International Series in
Operations Research & Management Science

Joe Zhu *Editor*

# Data Envelopment Analysis

## A Handbook of Models and Methods


Springer

# International Series in Operations Research & Management Science

Volume 221

**Series Editors**
Camille C Price
Stephen F. Austin State University, TX, USA

**Associate Series Editor**
Joe Zhu
Worcester Polytechnic Institute, MA, USA

**Founding Series Editor**
Frederick S. Hillier
Stanford University, CA, USA

More information about this series at http://www.springer.com/series/6161

Joe Zhu
Editor

# Data Envelopment Analysis

## A Handbook of Models and Methods

Springer

*Editor*
Joe Zhu
International Center for Auditing and Evaluation
Nanjing Audit University
Nanjing, P.R. China

School of Business
Worcester Polytechnic Institute
Worcester, MA 01545
USA

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-1-4899-7552-2 ISBN 978-1-4899-7553-9 (eBook)
DOI 10.1007/978-1-4899-7553-9

Library of Congress Control Number: 2015931384

Springer Boston New York Dordrecht London
© Springer Science+Business Media New York 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

This handbook complements the second edition of the *Handbook on Data Envelopment Analysis* (Cooper et al. 2011, Springer). Data envelopment analysis (DEA) is a "data-oriented" approach for evaluating the performance of a set of entities called Decision Making Units (DMUs) whose performance is categorized by multiple metrics. These performance metrics are indicated as inputs and outputs under DEA. Although DEA has a strong link to production theory in economics, the tool is also used for benchmarking in operations management where a set of measures is selected to benchmark the performance of manufacturing and service operations. In the circumstance of benchmarking, the efficient DMUs, as defined by DEA, may not necessarily form a "production frontier," but rather lead to a "best-practice frontier" (Cook et al. 2014).

Since the publication of the second edition of *Handbook on Data Envelopment Analysis*, there has been a significant amount of research on DEA methodology. As pointed out in a citation-based DEA survey by Liu et al. (2013), it is expected that the literature will grow to at least double its current size. With the recent publication of *Data Envelopment Analysis: A Handbook of Modeling Internal Structures and Networks* (Cook and Zhu 2014) written by experts on models and applications of DEA dealing with network and internal DMU structures, the current handbook is intended to represent another milestone in the progression of DEA. Written by experts, who are often major contributors to the DEA theory, it includes a collection of 16 chapters that represent the current state-of-the-art DEA research.

Chapter 1, by Färe, Grosskopf and Margaritis, provides an overview of the dual measurement of efficiency by means of distance functions and their value duals, the profit, revenue and cost functions.

Chapter 2, by Cook and Zhu, discusses cross-efficiency measures in DEA. While DEA has been proven an effective approach in identifying best practice frontiers, its flexibility in weighting multiple performance measures (inputs and outputs) and its nature of self-evaluation have been criticized. The cross efficiency method is developed as a DEA extension to rank DMUs and as a peer-evaluation approach. To complement Chap. 2, Lim and Zhu discuss how to use Variable Returns to Scale (VRS) models to develop cross efficiency in Chap. 3.

While the standard DEA assumes that data for inputs and outputs are continuous, there are situations where data are discrete and take form of integers. Chapter 4, by Kuosmanen, Keshvari and Kazemi Matin, examines the axiomatic foundations of integer DEA. The authors examine alternative efficiency metrics available for integer DEA, and consider estimation of the integer DEA technology under stochastic noise, modeling inefficiency and noise as Poisson distributed random variables.

There is a large literature on the use of weight restrictions in multiplier DEA models. In Chap. 5, Podinovski provides an alternative view of this subject from the perspective of dual envelopment DEA models in which weight restrictions can be interpreted as production trade-offs.

Chapter 6, by Olesen and Petersen, is concerned with development of indicators to determine whether or not the specification of the input and output space is supported by data in the sense that the variation in data is sufficient for estimation of a frontier of the same dimension as the input output space.

Kuosmanen, Johnson and Saastamoinen present a unified framework of productivity analysis, referred to as Stochastic Nonparametric Envelopment of Data (StoNED) in Chap. 7.

Chapter 8, by Pastor and Aparicio, presents an overview of the different approaches that have considered translation invariant DEA models. Translation invariance is a relevant property for dealing with non-positive input and/or non-positive output values in DEA.

Chapter 9, by Sahoo and Tone, provides a critical review of various possible estimation methods of scale economies in a non-parametric data envelopment analysis approach.

Chapter 10, by Zhu, describes several DEA models that can be used as a benchmarking tool where a set of DMUs is compared to existing standards. These existing standards can be in a form of DEA best-practice frontier or a set of pre-selected DMUs.

The conventional DEA assumes a set of homogeneous DMUs in the sense that each uses the same input and output measures (in varying amounts from one DMU to another). In some situations however, the assumption of homogeneity among DMUs may not apply. Chapter 11, by Cook, Harrison, Imanirad, Rouse and Zhu, presents DEA approaches for performance evaluation in the absence of homogeneity.

Chapter 12, by Kao, introduces a set of fuzzy DEA models in which data are missing and have to be estimated or predicted.

While it is generally assumed that all DEA outputs are impacted by all DEA inputs, there are many situations where this may not be the case. Chapter 13, by Cook and Zhu, extends the conventional DEA methodology to allow for the measurement of technical efficiency in situations where only partial input-to-output impacts exist. The new methodology involves viewing the DMU as a business unit, consisting of a set of mutually exclusive subunits, each of which can be treated in the conventional DEA sense.

In an effort to discriminate the performance of DMUs, the concept of super-efficiency is proposed in DEA. When applied to the variable returns to scale (VRS) situation, the resulting super-efficiency model may become infeasible for certain

DMUs. In Chap. 14, Chen and Du present a comprehensive overview of the infeasibility issues in the super efficiency DEA.

Chapter 15, by Zhou and Liu, discusses the treatment of undesirable measures in DEA. Chapter 16, by Asmild, reviews different ways of comparing the efficiency frontiers for subgroups within a data set, specifically program efficiency, the meta-technology (or technology gap) ratio and the global frontier difference index.

The current handbook focuses only on (new) models/approaches of DEA. Empirical studies using DEA will appear in *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications*, which I am currently editing. Along with other DEA handbooks, I hope that this handbook can serve as a reference for researcher and practitioners using DEA and as a guide for further development of DEA. I am indebted to the many DEA researchers worldwide for their continued effort in pushing the DEA research frontier. Without their work, many of the DEA models and approaches would not exist and be applied. I would like to thank the support from the Priority Academic Program Development of Jiangsu Higher Education Institutions in China.

September 2014 Joe Zhu

## References

Cook WD, Tone K, Zhu J (2014) Data envelopment analysis: prior to choosing a model. Omega 44:1–4

Cook WD, Zhu J (2014) Data envelopment analysis: a handbook of modeling internal structures and networks, Springer

Cooper WW, Seiford LM, Zhu J (2011) Handbook on data envelopment analysis, 2nd Edn., Springer

Liu JS, Lu LY, Lu WM, Lin BJ (2013) Data envelopment analysis 1978–2010: a citation based literature survey. Omega 41(1):3–15

# Contents

# Contributors

**Juan Aparicio** Center of Operations Research (CIO), University Miguel Hernandez of Elche, Elche, Spain

**Mette Asmild** IFRO, University of Copenhagen, Frederiksberg C, Denmark

**Yao Chen** International Center for Auditing and Evaluation, Nanjing Audit University, Nanjing, P.R. China

Manning School of Business, University of Massachusetts at Lowell, Lowell, MA, USA

**Wade D. Cook** Schulich School of Business, York University, Toronto, ON, Canada

**Juan Du** School of Economics and Management, Tongji University, Shanghai, Siping Road, P.R. China

**Rolf Färe** Department of Agricultural and Resource Economics, Oregon State University, Corvallis, OR, USA

Department of Economics and CERE, Umeå, Sweden

**Shawna Grosskopf** Department of Economics and CERE, Umeå, Sweden

**Julie Harrison** Department of Accounting & Finance, University of Auckland, Auckland, New Zealand

**Raha Imanirad** Schulich School of Business, York University, Toronto, ON, Canada

**Andrew Johnson** School of Business, Aalto University, Helsinki, Finland

Department of Industrial and Systems Engineering, Texas A&M University, Texas, TX, USA

**Chiang Kao** Department of Industrial and Information Management, National Cheng Kung University, Tainan, Taiwan

**Abolfazl Keshvari** Aalto University School of Business, Helsinki, Finland

**Timo Kuosmanen** Aalto University School of Business, Helsinki, Finland

**Sungmook Lim** Dongguk Business School, Dongguk University-Seoul, Seoul, South Korea

**Wenbin Liu** School of Business Administration, Hunan University, Changsha, China

Kent Business School, University of Kent, Canterbury, UK

**Dimitri Margaritis** Department of Accounting and Finance, University of Auckland, Auckland, New Zealand

**Reza Kazemi Matin** Department of Mathematics, College of Basic Science, Karaj Branch, Islamic Azad University, Alborz, Iran

**Ole B. Olesen** Department of Business and Economics, The University of Southern Denmark, Odense, Denmark

**Jesus T. Pastor** Center of Operations Research (CIO), University Miguel Hernandez of Elche, Elche, Spain

**Niels Chr. Petersen** Department of Business and Economics, The University of Southern Denmark, Odense, Denmark

**Victor V. Podinovski** Warwick Business School, University of Warwick, Coventry, UK

**Paul Rouse** Department of Accounting & Finance, University of Auckland, Auckland, New Zealand

**Antti Saastamoinen** School of Business, Aalto University, Helsinki, Finland

**Biresh K. Sahoo** Xavier Institute of Management, Bhubaneswar, India

**Kaoru Tone** National Graduate Institute for Policy Studies, Tokyo, Japan

**Zhongbao Zhou** School of Business Administration, Hunan University, Changsha, China

**Joe Zhu** School of Business, Worcester Polytechnic Institute, Worcester, MA, USA

# Chapter 1
# Distance Functions in Primal and Dual Spaces

**Rolf Färe, Shawna Grosskopf and Dimitri Margaritis**

**Abstract** This chapter provides an overview of the dual measurement of efficiency by means of distance functions and their value duals, the profit, revenue and cost functions. We start by showing how the Shephard (input) distance function in quantity space is a cost function in price space and how the cost function in quantity space is a distance function in price space. We then proceed to formulate a more unifying structure that allows for the simultaneous adjustment of inputs and outputs via establishing duality between the profit function and the directional (technology) distance function which also enables us to derive duality results for the revenue and cost functions as special cases. We complete our exposition by explaining how we can implement empirically dual forms of these efficiency measures either via activity analysis accounting for environmental technologies, slack-based measures and endogenous directional vectors or via parametric methods.

**Keywords** DEA · Directional distance functions · Shephard distance functions · Duality theory · Profit efficiency · Cost efficiency · Revenue efficiency · Slack-based measures · Endogenous directional vectors Generalized quadratic forms

## 1.1 Introduction

Introduced into economics by Ronald W. Shephard (1953), distance functions have proved to be useful tools for both theory and applied work. In this chapter we start with an in depth examination of their role in both primal (quantity) and dual (price) spaces. This forms the basis for the further study, including their key role in efficiency measurement and duality.

R. Färe (✉)
Department of Agricultural and Resource Economics,
Oregon State University Corvallis, OR, 97331, USA
e-mail: rolf.fare@oregonstate.edu

S. Grosskopf
Department of Economics and CERE, Umeå, Sweden

D. Margaritis
Department of Accounting and Finance, University of Auckland, Auckland, New Zealand

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations
Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_1

We extend our earlier work, Färe et al. (2007a) by introducing primal and dual optimization, slack-based directional distance functions and endogenous directional vectors.

Estimation issues complete the chapter including the nonparametric DEA/Activity Analysis estimators as well as an appendix addressing parametric estimation of distance functions.

## 1.2 Cost and Distance Functions in the Primal and Dual Spaces: An Introductory Example

In this section we introduce the idea of duality between distance functions and value functions.The specific example we use is the duality between the cost function and the input distance function. Both of these functions can be defined as an optimization of distance or value, where the latter is an inner product.

In order to make this concrete we begin with some notation. Let $x = (x_1, \ldots, x_N) \in \Re^N_+$ be an $N-$vector of non-negative inputs employed in producing a vector $y = (y_1, \ldots, y_M) \in \Re^M_+$ of non-negative outputs. We represent the technology which maps outputs into inputs by

$$L : \Re^N_+ \rightarrow L(y) \subseteq \Re^M_+, \tag{1.1}$$

where

$$L(y) = \{x : x \text{ can produce } y\},\ y \in \Re^M_+, \tag{1.2}$$

denotes the input requirement sets, one for each $y \in \Re^M_+$. These sets are assumed to satisfy standard set of conditions (axioms) including[1] **strong disposability of inputs**:

$$x' \geqq x^o \in L(y) \Rightarrow x' \in L(y)$$

and **convexity**

$$x', x^o \in L(y), 0 \leqq \lambda \leqq 1 \Rightarrow \lambda x' + (1 - \lambda)x^o \in L(y).$$

These conditions are required for duality theory, when we also require non-negative input prices $w = (w_1, \ldots, w_N) \in \Re^N_+$.

The cost and distance functions can be derived from the technology by optimization, specifically by minimization. Let

$$wx = \sum_{n=1}^{N} w_n x_n$$

[1] See Färe and Primont (1995) for a discussion of these axioms.

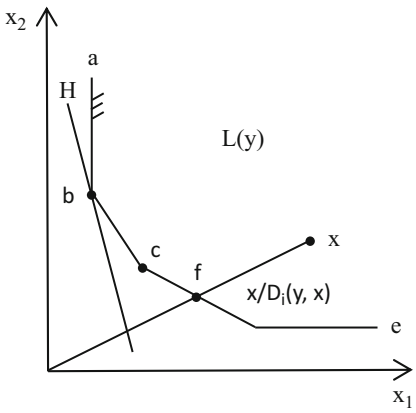



**Fig. 1.1** Cost and distance function minimizations

be the inner product of the price vector $w$ and the input quantity vector $x$.[2] The cost function is derived by the minimization of this inner product (value) as[3]

$$C(y, w) = \min_x \{wx : x \in L(y)\}. \tag{1.3}$$

The input distance function is derived by 'minimizing' the distance to the boundary of $L(y)$, where $L(y)$ is assumed to be closed, i.e.,

$$D_i(y, x) = \sup_\lambda \{\lambda : \frac{x}{\lambda} \in L(y)\}. \tag{1.4}$$

These two functions are illustrated in Fig. 1.1. The boundary of $L(y)$ is defined to be the isoquant for the output vector $y$, i.e.,

$$Isoq\ L(y) = \{x : x \in L(y),\ \lambda < 1 \Rightarrow \lambda x \notin L(y)\}, y \in \Re_+^M. \tag{1.5}$$

The isoquant for output vector $y$ is given by the line segments a-b-c-d-e and the input requirement set $L(y)$ is the area northeast of the isoquant.

Cost is minimized at $b$ for the inner product $wx$ expressed in terms of the hyperplane H. In general a hyperplane is described in terms of the inner product of prices and quantities, $wx$ as

$$H(w, c) = \{x : wx = c\}, \tag{1.6}$$

where $c$ is a constant and we are in the primal. To distinguish the dual hyperplane from the primal hyperplane $H(w, c)$ we refer to the dual or price space hyperplane for $wx$ as

$$\mathbf{H}(x, c) = \{w : wx = c\}, \tag{1.7}$$

[2] We think of $x \in \Re_+^N$ as belonging to the primal (quantity) space and $w \in (\Re_+^N)^*$ as belonging to its dual (price) space. Since $x$ is a vector of real numbers its dual space $(\Re^N)^*$ equals $\Re^N$. Hence in this case we need not distinguish between the primal and dual spaces. See Luenberger (2001).

[3] For the existence of the minimum see Färe and Primont (1995).

where the set is defined in prices rather than quantities.

To continue with our figure, the distance between $x$ and the isoquant along the ray $(x/||x||)$ is minimized at $f$, where

$$\hat{x} = x/D_i(y,x). \tag{1.8}$$

Note that the cost and distance functions are homogeneous of degree +1 in $w$ and $x$ respectively, i.e.,

$$C(y,\lambda w) = \lambda C(y,w), \lambda > 0, \tag{1.9}$$

and

$$D_i(y,\lambda x) = \lambda D_i(y,x), \lambda > 0. \tag{1.10}$$

Also note that since inputs are weakly disposable $(x \in L(y), \lambda > 1 \Rightarrow \lambda x \in L(y))$, the input distance function is a representation of the input requirement set

$$L(y) = \{x\colon D_i(y,x) \geqq 1\}, y \in \Re_+^M.$$

We also have that

$$Isoq\ L(y) = \{x\colon D_i(y,x) = 1\}, y \in \Re_+^M,$$

i.e., the distance function takes a value of one if and only if the input vector $x \in L(y)$ belongs to the isoquant, $x \in Isoq\ L(y)$. We next establish the dual relationship between the input distance function and the cost function. Since the cost function $C(y,w)$ minimizes the inner product $wx$ over all feasible input vectors $x \in L(y)$, we have

$$C(y,w) \leqq wx \text{ for all } x \in L(y). \tag{1.11}$$

We also know that

$$(x/D_i(y,x)) \in L(y), \tag{1.12}$$

thus

$$C(y,w) \leqq w(x/D_i(y,x)) = \frac{wx}{D_i(y,x)} \tag{1.13}$$

or

$$C(y,w)/wx \leqq 1/D_i(y,x), \tag{1.14}$$

where the normalized minimum cost is not larger than the reciprocal of the distance function. This inequality known as the Mahler inequality (Mahler 1939) is the basis for the following duality:

$$C(y,w) = \min_x \frac{wx}{D_i(y,x)} \tag{1.15}$$

$$D_i(y,x) = \min_w \frac{wx}{C(y,w)}$$

The first expression shows how the cost function may be recovered from the input distance function, and the second expression retrieves the distance function from the cost function.[4]

To find a similar primal and dual interpretation of the cost function we need to establish it as a distance function in price space. We begin by defining the 'price' technology as

$$\pounds(y,c) = \{w\colon C(y,w) \leqq c\}. \tag{1.16}$$

The distance function defined on $\pounds(y,c)$ is

$$\begin{aligned} min\{\lambda : w/\lambda \in \pounds(y,c)\} & \\ : C(y,w/\lambda \leqq c\} & \\ : C(y,w)/c) \leqq \lambda\} & \\ = C(y,w)/c, & \end{aligned} \tag{1.17}$$

showing that the (normalized) cost function is a distance function in price space.

Summing up: we have established that the cost function in quantity space is a distance function in price space and that the distance function in quantity space is a cost function in price space.[5]

## 1.3 Distance Functions and Their Duals

From the introductory example showing the relation between the input distance function (Shephard 1953) and the cost function we now turn to the directional distance functions and their duals.

Recall that $x \in \Re_+^N$ denotes inputs and $y \in \Re_+^M$ outputs. Another way of representing the technology follows:

$$T = \{(x,y)\colon x \text{ can produce } y\}$$

which is related to the input sets introduced previously as

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

A third model of the technology is the output set defined as

$$P(x) = \{y\colon x \text{ can produce y}, y \in \Re_+^M\},$$

[4] This requires convexity.

[5] This was first noted by Shephard (1953), in less detail.

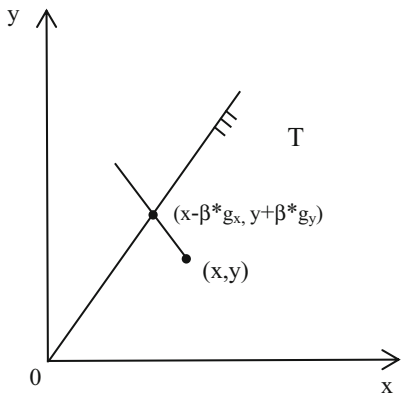



**Fig. 1.2** The directional technology distance function

and it follows that all three representations are related by

$$y \in P(x) \Leftrightarrow (x, y) \in T \Leftrightarrow x \in L(y). \tag{1.18}$$

In order to define the directional distance function we need to introduce a direction vector. Let $g = (g_x, g_y) \in \Re_+^N \times \Re_+^M, g \neq 0$, be a directional vector, a vector in which $(x, y)$ is projected to the boundary of the relevant technology set, in this case $T$.[6] The particular value of the direction vector is chosen by the researcher; there is no general rule for this choice. In Sect. 1.6 we show how one may endogenize the directional vector through optimization.

We are now ready to define the directional technology distance function,[7] following Chambers et al. (1988):

$$\vec{D}_T(x, y; g_x, g_y) = \max\{\beta : (x - \beta g_x, y + \beta g_y) \in T\} \tag{1.19}$$

is well-defined if there exists a $\beta \in \Re$ such that $(x - \beta g_x, y + \beta g_y) \in T$ and $+\infty$ otherwise. Figure 1.2 illustrates.

The input-output vector $(x, y)$ is expanded along the directional vector $(g_x, g_y)$ where $g_x$ is subtracted from $x$ and $g_y$ is added to $y$. The optimal $\beta^* = \vec{D}_T(x, y; g_x, g_y)$ is achieved at the boundary of $T$.

This distance function inherits its properties from those of the technology $T$, see Chambers et al. (1998) for detail. Given its properties, the distance function satisfies the **representation property**, i.e.,

$$\vec{D}_T(x, y; g_x, g_y) \geqq 0 \Leftrightarrow (x, y) \in T. \tag{1.20}$$

In words, the directional distance function fully describes technology and is consistent with the properties in $T$, i.e., it is a function representation of the technology set, $T$.

---

[6] Note that the input directional vector $g_x$ is taken here to be non-negative, however we 'deduct' it from the input vector $x$, analogous to the subtraction of cost from revenue to obtain profit.

[7] This was first introduced by Luenberger (1992) in the form of a shortage function.

From its definition—which has an additive structure—the distance function satisfies the translation property,

$$\vec{D}_T(x - \alpha g_x, y + \alpha g_y; g_x, g_y) = \vec{D}_T(x, y; g_x, g_y) - \alpha, \alpha \in \Re. \tag{1.21}$$

This condition corresponds to homogeneity of the input distance function. It will be shown to be important empirically when the distance function is given a functional form representation.

Next we introduce output price $p = (p_1, \ldots, p_M) \in \Re_+^M$, which together with input prices $w \in \Re_+^N$ are required to define the profit function as

$$\Pi(p, w) = \max_{x,y} \{py - wx : (x, y) \in T\} \tag{1.22}$$

and when it exists

$$\Pi(p, w) \geqq py - wx, \text{ for all } (x, y) \in T. \tag{1.23}$$

Since

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

from the definition of the distance function, it follows that

$$\frac{\Pi(p, w) - (py - wx)}{pg_y + wg_x} \geqq \vec{D}_T(x, y; g_x, g_y) \tag{1.24}$$

which establishes the relationship between the directional technology distance function and its dual profit function.

This inequality between the profit and distance function is the basis for

i) the Nerlovian profit indicator and
ii) the duality between the profit function and the directional technology distance function.

(The reader should recall the associated duality between the input distance function and the cost function.)

Chambers et al. (1998) used the above to define and name the Nerlovian Profit Indicator as

$$NPI = \frac{\Pi(p, w) - (py - wx)}{pg_y + wg_x}. \tag{1.25}$$

It subtracts observed from maximal profit and normalizes that by the value of the directional vectors. This indicator can be decomposed into technical efficiency

$$\vec{D}_T(x, y; g_x, g_y) = TE \tag{1.26}$$

which is measured by the distance function, and a residual term, $\vec{AE}_T$, called allocative efficiency. Together they form the additive decomposition

$$NPI = TE + \vec{AE}_T. \tag{1.27}$$

The duality between the profit function and the directional technology distance function is derived from the above inequality in (1.24) by optimizing over quantities $(x, y)$ and prices $(w, p)$, respectively.

$$\Pi(p, w) = \max_{x,y} p(y + \vec{D}_T(x, y; g_x, g_y)) - w(x - \vec{D}_T(x, y; g_x, g_y)) \tag{1.28}$$

and

$$\vec{D}_T(x, y; g_x, g_y) = \min_{w,p} \frac{\Pi(p, w) - (py - wx)}{pg_y + wg_x}. \tag{1.29}$$

Convexity of the technology $T$ is required for the distance function to be retrieved from $\Pi(p, w)$, i.e., for the last equality to hold.

The directional distance function includes other distance functions as special cases by restricting the directional vectors $g_x$ and $g_y$. If we first take $g = (g_x, 0)$, we have the directional input distance function, i.e.,

$$\vec{D}_i(x, y; g_x) = \vec{D}_i(x, y; g_x, 0). \tag{1.30}$$

By also choosing $py = py^*$, where $y^*$ is the profit optimizing output bundle, we can solve for the directional input distance function's dual cost function:

$$\vec{D}_i(x, y; g_x) \leqq \frac{wx - C(y, w)}{wg_x}. \tag{1.31}$$

If we further restrict the direction vector $g_x$ to be the observed input vector $x$, we find that

$$\begin{aligned} \vec{D}_i(x, y; x) &= \vec{D}_T(x, y; x, 0) \\ &= 1 - 1/D_i(y, x), \end{aligned} \tag{1.32}$$

which is the Shephard input distance function, and it follows that

$$\vec{D}_i(x, y : x) = 1 - \frac{1}{D_i(y, x)} \leqq \frac{wx - C(y, w)}{wx}. \tag{1.33}$$

Rearranging yields the inequality derived in our introductory example

$$\frac{C(y, w)}{wx} \leqq \frac{1}{D_i(y, x)}. \tag{1.34}$$

As shown in Sect. 1.2 this inequality is the basis for Shephard's cost/distance function duality. In addition if we add a residual (multiplicative) 'allocative efficiency' term, the Farrell (1957) decomposition of cost efficiency follows

$$\begin{array}{ccccc} \frac{C(y,w)}{wx} & = & \frac{1}{D_i(y,x)} & \times & AE_i \\ \textit{cost} & & \textit{technical} & & \textit{allocative} \\ \textit{efficiency} & & \textit{efficiency} & & \textit{efficiency} \end{array}$$

Returning to the directional input distance function, we note that it may be defined directly on the technology:

$$\begin{aligned} \vec{D}_i(x, y; g_x) &= \max\{\beta \colon (x - \beta g_x, y) \in T\} \\ &= \max\{\beta \colon (x - \beta g_x, y) \in L(y)\} \end{aligned} \tag{1.35}$$

If instead of setting $g_y = 0$, we restrict $g_x = 0$, we can derive the parallel on the output side, namely, the directional output distance function as follows

$$\vec{D}_o(x, y; g_y) = \vec{D}_T(x, y; 0, g_y), \tag{1.36}$$

and by choosing $wx = wx^*$, $x^*$ being the profit maximizing input vector we have

$$\vec{D}_o(x, y; g_y) \leqq \frac{R(x, p) - py}{pg_y}, \tag{1.37}$$

where

$$R(x, p) = \max_y\{py \colon y \in P(x)\} \tag{1.38}$$

is the revenue function. The inequality above is a Mahler inequality which may be used to formulate the duality between revenue and the directional output distance function as

$$R(x, p) = \max_y\{p(y + \vec{D}_o(x, y; g_y)g_y)\}. \tag{1.39}$$

Rearranging and accounting for the definition of the revenue function we have

$$\vec{D}_o(x, y; g_y) \leqq \max_p \frac{R(x, p) - py}{pg_y}. \tag{1.40}$$

Again by adding a residual allocative component we have a revenue indicator efficiency decomposition which is analogous to the Nerlovian profit efficiency decomposition:

$$\begin{array}{ccccc} \frac{R(x,p)-py}{pg_y} & = & \vec{D}_o(x, y; g_y) & + & \vec{AE}_o \\ \textit{revenue} & & \textit{technical} & & \textit{allocative} \\ \textit{indicator} & & \textit{indicator} & & \textit{indicator} \end{array}$$

When we further restrict the directional output vector to be equal to observed output, $g_y = y$, we find that

$$\vec{D}_o(x, y; g_y) = \frac{1}{D_o(x, y)} - 1, \tag{1.41}$$

which demonstrates that the directional output distance function and Shephard's output distance function

$$D_o(x, y) = \min_{\lambda}\{\lambda : y/\lambda \in P(x)\}, x \in \Re^N_+, \tag{1.42}$$

are closely related. Using this relationship we can derive a Farrell type output oriented (revenue) efficiency decomposition, namely

$$\begin{array}{ccccc} \frac{R(x,p)}{py} & = & \frac{1}{D_o(x,y)} & \times & AE_o \\ \textit{revenue} & & \textit{technical} & & \textit{allocative} \\ \textit{efficiency} & & \textit{efficiency} & & \textit{efficiency} \end{array}$$

We refer to Färe et al. (2008) for a summary of the relationships concerning duality and the efficiency inequalities discussed in this section, including an approximate reproduction of the associated diagram below.

$$\begin{array}{ccc} \frac{\Pi(p,w)(p,w)-(py-wx)}{pg_y+wg_x} & \geqq & \vec{D}_T(x,y;g_x,g_y) \\ \downarrow & & \downarrow \\ g_x = 0, wx = wx^* & & g_y = 0, py = py^* \\ \swarrow & & \searrow \\ \frac{R(x,p)-py}{pg_y} \geqq \vec{D}_o(x,y,g_y) & & \frac{wx-C(y,w)}{pg_x} \geqq \vec{D}_i(x,y;g_x) \\ \downarrow & & \downarrow \\ g_y = y & & g_x = x \\ \downarrow & & \downarrow \\ \frac{R(x,p)}{py} \geqq \frac{1}{D_o(x,y)} & & \frac{C(y,w)}{px} \leqq \frac{1}{D_i(y,x)} \end{array}$$

**Efficiency Inequalities**

Beginning at the top we have the most general case, the relationship between profit and the directional distance function. We continue by quoting from Färe et al. (2008)

> If we restrict our focus to scaling on either outputs or inputs (setting the appropriate directional vector equal to zero) and set observed cost (revenue) equal to its optimal value, we arrive at the middle row relationships between an additive revenue efficiency indicator (additive cost efficiency indicator) and the associated directional distance function. If we set the restricted direction vector equal to the value of the observation under evaluation, our cost and revenue inequalities reduce to the familiar Farrell relationships using Shephard distance functions. Summing up, we have given a brief overview of the relationships between the various distance functions and their value duals, including
> - directional technology distance function and the profit function,
> - Shephard output distance function and the revenue function, and
> - Shephard input distance function and the cost function.

## 1.4 Distance Functions and Efficiency Measurement

Economists distinguish between isoquants and efficient (sub)sets, which has implications for efficiency measurement. Following Russell and Schworm (2009), different efficiency measures have what they call different representation properties. For example, the Farrell input-oriented measure of technical efficiency indicates efficiency as membership in the isoquant, but not necessarily the efficient subset since these do not generally coincide in an activity analysis framework.

Färe (1975)[8] first addressed this issue by introducing an efficiency measure which takes value of one if and only if the input vector belongs to an efficient subset. He writes:

> ...the isoquant and the efficient subset need not coincide; this fact makes the Farrell measure of (in)efficiency inappropriate for technologies ....where these sets differ.

In this section we take up the topic of efficiency measurement with these two different indication properties, namely the isoquant and the efficient subset. We choose the input correspondence and input-oriented measures of technical efficiency as our focus.[9]

Denote the input correspondence as

$$L: \Re_+^M \rightarrow L(y) \in 2^{\Re_+^N}, \tag{1.43}$$

where the input sets are

$$L(y) = \{x: x \text{ can produce } y\}, y \in \Re_+^M. \tag{1.44}$$

The isoquants are defined as

$$Isoq\ L(y) = \{x: x \in L(y), \lambda x \notin L(y), \lambda < 1\}, y \in \Re_+^M, \tag{1.45}$$

and the efficient subsets are

$$Eff\ L(y) = \{x: x \in L(y), x^o \leqq x, x^o \neq x \Rightarrow x^o \notin L(y)\}, y \in \Re_+^M. \tag{1.46}$$

As mentioned above, these two subsets of $L(y)$ are in general not equal, as the Leontief production function below illustrates (Fig. 1.3):

$$y = min\{x_1, x_2\}. \tag{1.47}$$

This figure depicts the input requirement set for our simple Leontief technology. Its isoquant is bounded by the line segments ab-bc and their extensions, whereas the efficient subset is the set $\{b\}$.

---

[8] This paper is the consequence of many conversations with Professor R.W. Shephard.

[9] A 'must read' on this topic is W. Schworm and R. R. Russell: Axiomatic Foundations of Technical Efficiency Measurement (in progress).

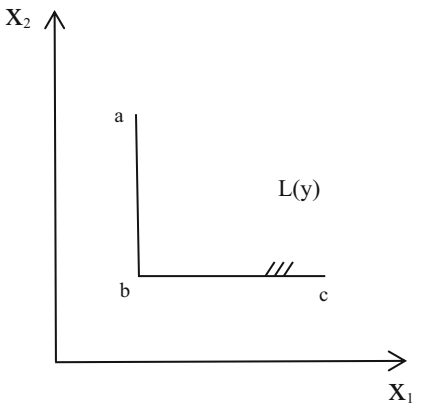



**Fig. 1.3** Leontief input requirement set

In general it is true that

$$Eff\ L(y) \subseteq Isoq\ L(y), \tag{1.48}$$

but as our example demonstrates—these sets need not coincide.[10]

Recall that the Farrell input measure of technical efficiency is the reciprocal of the input distance function, i.e.,

$$F_i(y,x) = (D_i(y,x))^{-1} = \inf_{\lambda}\{\lambda : \lambda x \in L(y)\}, \tag{1.49}$$

and since

$$D_i(y,x) = 1 \Leftrightarrow x \in Isoq\ L(y), \tag{1.50}$$

we have that the Farrell input measure of technical efficiency indicates the isoquant as consisting of efficient points, i.e.,

$$F_i(y,x) = 1 \Leftrightarrow x \in Isoq\ L(y), \tag{1.51}$$

which makes it inappropriate as a measure of 'efficiency' whenever

$$Eff\ L(y) \neq Isoq\ L(y). \tag{1.52}$$

Following Färe (1975) a variety of technical efficiency measures which 'indicate' efficiency relative to the efficient subset have been developed. Some are multiplicative such as the Russell Measure, proposed by Färe and Lovell (1978) and some are additive such as the Slack-Based Measure due to Tone (2001) or the directional distance

[10] This is often the case in the DEA/Activity Analysis case, and always when inputs are strictly positive as in Charnes et al. (1978).

function based measure in Färe and Grosskopf (2010). We focus our discussion on the input oriented version from Färe and Grosskopf (2010).[11] It is defined as

$$\begin{aligned} \vec{SD}_i(x^o, y^o; g_x) &= \max \beta_1 + \ldots \beta_N \\ s.t. \qquad & (x_1^o - \beta_1 g_{x_1}, x_2^o - \beta_2 g_{x_2}, \ldots, x_N^o - \beta_N g_{x_N}) \in L(y). \end{aligned} \tag{1.53}$$

Note that if we take $g_{x_n} = x_n^o, n = 1, \ldots, N$ we have

$$\begin{aligned} \vec{SD}_i(x^o, y^o; g_x) &= N \cdot 1 + (1 - \beta_1) - \ldots - (1 - \beta_N) \\ s.t. \qquad & ((x_1^o(1 - \beta_1), x_2^o(1 - \beta_2), \ldots, x_N^o(1 - \beta_N) \in L(y), \end{aligned} \tag{1.54}$$

which is an additive version of the Russell measure, see Färe et al. (2007b). One can prove that $\vec{SD}_i(x^o, y^o; g_x)$ indicates efficiency relative to the efficient subset, see Färe and Grosskopf (2010). As a final note, Färe, Grosskopf and Zelenyuk point out that the Russell-type efficiency measures are difficult to associate with a dual formulation.

## 1.5 DEA Estimators

Given a set of observations $(x^k, y^k), k = 1, \ldots, K$ of inputs and outputs we may construct the DEA/Activity Analysis technology as

$$\begin{aligned} T = \{(x, y) : & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n && n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m, && m = 1, \ldots, M \\ & z_k \geqq 0, && k = 1, \ldots, K\}. \end{aligned} \tag{1.55}$$

We assume that the data satisfy the Kemeny, Morgenstern and Thompson et al. (1956) conditions:

$$\begin{aligned} &(i) \ \sum_{m=1}^{M} y_{km} > 0, k = 1, \ldots, K, \qquad (ii) \ \sum_{k=1}^{K} y_{km} > 0, m = 1, \ldots, M, \\ &(iii) \ \sum_{n=1}^{N} x_{kn} > 0, k = 1, \ldots, K, \qquad (iv) \ \sum_{k=1}^{K} x_{kn} > 0, n = 1, \ldots, N. \end{aligned} \tag{1.56}$$

The condition (i) states that each Decision Making Unit (DMU) produces at least one type of output, (ii) states that each output is produced by at least one DMU.

[11] Färe and Grosskopf (2010) develop their model based on technology $T$ rather than as we do here with the input set $L(y)$.

Similarly the last two conditions require that each DMU uses some input and each input is used by some DMU. These conditions are weaker than originally imposed by Charnes et al. (1978), which required strictly positive inputs and outputs.

Given that (i–iv) hold, one can prove that $T$ is a closed convex set with bounded output sets $P(x)$. In addition the technology set $T$ defined above satisfies free disposability of inputs and outputs together with constant returns to scale (CRS). CRS is imposed through the non-negativity constraint on the intensity variables $z_k, k = 1, \ldots, K$.

Returning to our cost function, with input prices $w \in \Re^N_+, w \neq 0$, then the DEA version of the cost minimization problem for observation $k'$ is

$$C(y^{k'}, w) = \min_{z,x} wx \tag{1.57}$$

$$s.t. \quad \sum_{k=1}^{K} z_k x_{kn} \leqq x_n \; n = 1, \ldots, N$$

$$\sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm}, \; m = 1, \ldots, M$$

$$z_k \geqq 0, \; k = 1, \ldots, K.$$

Since the efficient subset $Eff \; L(y)$ is bounded and $L(y)$ is closed, one may have some prices equal to zero.

The dual problem is the input distance function which is solved for observation $k'$ as

$$(D_i(y^{k'}, x^{k'})^{-1}) = \min_{z,\lambda} \lambda \tag{1.58}$$

$$s.t. \quad \sum_{k=1}^{K} z_k x_{kn} \leqq \lambda x_{k'n} \; n = 1, \ldots, N$$

$$\sum_{k=1}^{K} z_k y_{km} \geqq y_{k'm}, \; m = 1, \ldots, M$$

$$z_k \geqq 0, \; k = 1, \ldots, K.$$

Thus the Farrell input-oriented measures of efficiency are

$$\underbrace{\frac{C(y^{k'}, w)}{wx^{k'}}}_{\substack{cost \\ efficiency}} = \underbrace{\frac{1}{D_i(y^{k'}, x^{k'})}}_{\substack{technical \\ efficiency}} \times \underbrace{AE_i}_{\substack{allocative \\ efficiency}}$$

Other Farrell type efficiency measures have similar DEA estimators and are left to the reader.

Next we introduce undesirable outputs, $b_1, \ldots, b_J$, such as pollution, and show how to adapt the DEA estimator to account for it. Let

$$T = \{(x, y, b)\colon x \text{ can produce } (y, b)\}. \tag{1.59}$$

We say that $y$ and $b$ are null joint if

$$(x, y, b) \in T \text{ and } y = 0 \Rightarrow b = 0. \tag{1.60}$$

In words, $y$ and $b$ are jointly produced (or $b$ is a byproduct of $y$), e.g., there is no fire $y$ without smoke $b$, or $y > 0 \Rightarrow b > 0$.

When we wish to model production with undesirable outputs where those outputs are regulated, we no longer wish to impose free disposability of those outputs, which would suggest that they may be costlessly disposed. Rather we assume that good and bad outputs are together weakly disposable:

$$(x, y, b) \in T, 0 \leqq \theta \leqq 1 \Rightarrow (x, \theta y, \theta b) \in T. \tag{1.61}$$

This states that proportional reductions in good and bad output together are feasible, which captures the idea that bads may be reduced by redirecting given inputs to abatement (and thereby reducing feasible good output). The DEA technology which includes these features of production with null jointness and weak disposability is written as

$$\begin{aligned} \{(x, y, b) : & \\ & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n,\ n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m,\ m = 1, \ldots, M \\ & \sum_{k=1}^{K} z_k b_{kj} = b_j,\ j = 1, \ldots, J \\ & z_k \geqq 0, k = 1, \ldots, K\}. \end{aligned} \tag{1.62}$$

Note that the $j = 1, \ldots, J$ constraints are strict equalities, which imposes joint weak disposability of good and bad outputs. Null jointness may be imposed by requiring:

v) $\sum_{j=1}^{J} b_{kj} > 0, k = 1, \ldots, K$
vi) $\sum_{k=1}^{K} b_{kj} > 0, j = 1, \ldots, J,$

i.e., each DMU produces some bad output and each bad output is produced by some DMU.

## 1.6 Endogenous Directional Vectors

In this section we show how the directional vectors for the directional distance functions may be solved for endogenously. We limit our exposition to the input-oriented case.[12]

In Sect. 1.4 we introduced a slack-based directional input distance function which we can be exploited to endogenize the choice of the directional input vector, $g_x$. First we specify the slack-based measure in a DEA framework, i.e.,

$$\begin{aligned} \vec{SD}_i(x^o, y^o; g_x) \; &= \max \; \beta_1 + \ldots \beta_N \qquad (1.63) \\ s.t. \quad & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n^o - \beta_n g_{x_n}, \; n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m^o, \; m = 1, \ldots, M \\ & z_k \geqq 0, \; k = 1, \ldots, K \end{aligned}$$

Next let us endogenize the directional vector $g_x = (g_{x_1}, \ldots, g_{x_N})$, so our revised problem becomes (with $\beta$ a scalar)

$$\begin{aligned} & \max_{z, \beta, g_x} \qquad (1.64) \\ s.t. \quad & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n^o - \beta g_{x_n}, \; n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m^o, \; m = 1, \ldots, M \\ & z_k \geqq 0, \; k = 1, \ldots, K \\ & \sum_{n=1}^{N} g_{x_n} = 1. \end{aligned}$$

Here we restrict $g_x$ to the unit simplex in order for our maximization problem to have a solution. Note that this is a nonlinear optimization problem.

To transform this into a linear form, we show that this problem is equivalent to our slack-based DEA model. To see this, consider our slack based measure with

[12] See Färe et al. (2013b) for the output orientation.

$g_x = (1, \ldots, 1)$:

$$\begin{aligned} \max \quad & \beta_1 + \ldots + \beta_N \\ s.t. \quad & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n^o - \beta_n \cdot 1, \ n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m^o, \ m = 1, \ldots, M \\ & z_k \geqq 0, \ k = 1, \ldots, K \\ & \sum_{n=1}^{N} g_{x_n} = 1. \end{aligned} \tag{1.65}$$

Let $\beta_n^*$ be an optimal solution and take $g_x$ such that

$$\begin{aligned} \beta_n^* &= \beta g_{x_n}, n = 1, \ldots, N \\ \sum_{n=1}^{N} \beta g_{x_n} &= 1, \end{aligned} \tag{1.66}$$

and write the slack-based problem as

$$\begin{aligned} \max \quad & \beta g_{x_1} + \ldots + \beta g_{x_N} = \beta(\textstyle\sum_{n=1}^{N} g_{x_n}) = \beta \\ s.t. \quad & \sum_{k=1}^{K} z_k x_{kn} \leqq x_n^o - \beta g_{x_n}, \ n = 1, \ldots, N \\ & \sum_{k=1}^{K} z_k y_{km} \geqq y_m^o, \ m = 1, \ldots, M \\ & z_k \geqq 0, \ k = 1, \ldots, K \\ & \sum_{n=1}^{N} g_{x_n} = 1, \end{aligned} \tag{1.67}$$

which is equivalent to our endogenous $g_x$ problem.

As an illustration of this problem, consider the following data set:

| $DMU$ | 1 | 2 |
|---|---|---|
| $y$ | 1 | 1 |
| $x_1$ | 1 | 2 |
| $x_2$ | 1 | 2 |

The slack-based solution to this problem for DMU 1 is

$$\beta_1^* = \beta_2^* = 0 \tag{1.68}$$

and we may take

$$g_1 = g_2 = 1/2. \tag{1.69}$$

For DMU 2 we have

$$\beta_1^* = \beta_2^* = 1 \tag{1.70}$$

and the directional vector for this DMU is

$$g_{x_n} = \frac{1}{1+1} = 1/2, n = 1, 2. \tag{1.71}$$

## 1.7 Appendix: Parametric Distance Functions

Data Envelopment Analysis or Activity Analysis is often referred to as a nonparametric representation of production technology since they do not require specification of a (parametric) functional form such as Cobb-Douglas, translog or quadratic functions for estimation. Although the focus in this chapter is nonparametric models[13] the chapter would be incomplete without a discussion of parametric estimation of distance functions.[14]

Recall that by its definition the directional distance function satisfies the translation property while the Shephard distance functions satisfy homogeneity. As we will demonstrate, these two properties generate different parametric forms.

Let $F : \Re^I \to \Re, h : \Re \to \Re$ and $\xi : \Re \to \Re$ be real-valued functions with $\xi^{-1}$ well-defined. If $a_i, a_{ij}$ are real constants and $q_i \in \Re_+$, we say that

$$\xi^{-1}(F(q)) = a_o + \sum_{i=1}^{I} a_i h(q_i) + \sum_{i=1}^{I}\sum_{j=1}^{I} a_{ij} h(q_i)h(q_j) \tag{1.72}$$

is a generalized quadratic function (Chambers 1988), a transformed quadratic function (Diewert 2002) or a function having a second order Taylor's series approximation (Färe and Sung 1986).

$F$ is homogeneous of degree +1 if

$$F(\lambda q) = \lambda F(q), \lambda > 0 \tag{1.73}$$

and satisfies the translation property if

$$F(q + \alpha g) = F(q) + \alpha, \alpha \in \Re, \tag{1.74}$$

where $g = (g_1, \ldots, g_I) \in \Re^I$ is a directional vector.

Note that $F : \Re^I \to \Re$ is linear in the parameters $a_o, a_i, a_{ij}$, and hence can be estimated using linear programming or linear regression.

---

[13] An alternative form of nonparametric estimators which is considered to be econometric is discussed in Martins-Filho in Färe et al. (2013a).

[14] This appendix builds on Chambers et al. (2013).

The generalized quadratic form together with either homogeneity or translation produce two sets of (two) functional equations:

$$\xi(F(q)) = a_o + \sum_{i=1}^{I} a_i h(q_i) + \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} h(q_i)h(q_j) \tag{1.75}$$

$$F(\lambda q) = \lambda F(q) \text{ homogeneity property}$$

and

$$\xi(F(q)) = a_o + \sum_{i=1}^{I} a_i h(q_i) + \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} h(q_i)h(q_j) \tag{1.76}$$

$$F(q + \alpha g) = F(q) + \alpha \text{ translation property}$$

The first set of equations with the homogeneity property has two solutions (see Färe and Sung (1986):

$$F(q) = a_o + \sum_{i=1}^{I} a_i \ln(q_i) + \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} \ln(q_i) \ln(q_j) \tag{1.77}$$

and

$$F(q) = \left( a_o + \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} q_i^{r/2} q_j^{r/2} \right)^{1/r}. \tag{1.78}$$

The first solution is known as the translog function (Christensen et al. 1971) and the second is known as the quadratic mean of order $r$ (Denny 1974; Diewert 1976). Note that the translog has both first order parameters $a_i$ as well as second order parameters $a_{ij}$, while the quadratic function has only second order parameters, since homogeneity requires that $a_i = 0$.

The second set of equations with the translation property also has two solutions (Färe and Lundberg 2006). Assuming that $g = (1, \ldots, 1)$ we have

$$F(q) = a_o + \sum_{i=1}^{I} a_i q_i + \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} q_i q_j \tag{1.79}$$

the quadratic function and

$$F(q) = \frac{1}{2\lambda} \ln \left( \sum_{i=1}^{I}\sum_{j=1}^{J} a_{ij} \exp(\lambda q_i) \exp(\lambda q_j) \right), \lambda \neq 0, \tag{1.80}$$

called 'quadratic exponential mean of order s', here $s = \lambda$ (Kolm 1976) and exponential mean of order s due to Diewert and Wales (1988).

Again the first solution has both first and second order terms while the second has only second order terms.

We conclude that if a generalized quadratic function can be justified, Shephard's distance functions should be parameterized as translog, while the directional distance function should be specified as quadratic functions.

## References

Chambers RG (1988) Applied production analysis: a dual approach. Cambridge University Press, Cambridge

Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions, and Nerlovian efficiency. J Optim Theory Appl 98(2):351–364

Chambers RG, Färe R, Grosskopf S, Vardanyan M (2013) Generalized quadratic revenue functions. J Econom 173:11–21

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Christensen LR, Jorgenson DW, Lau L (1971) Conjugate duality and the transcendental logarithmic production function. Econometrica 39:255–276

Denny MC (1974) The relationship between functional forms for the production system. Can J Econ 7:21–31

Diewert WE (1976) Exact and superlative index numbers. J Econom 4:115–145

Diewert WE (2002) The quadratic approximation lemma and decomposition of superlative indexes. J Econ Soc Meas 28:63–88

Diewert WE, Wales TJ (1988) Normalized quadratic systems of consumer demand functions. Can J Econ 26:77–106

Farrell MJ (1957) The measurement of productive efficiency. J Royal Stat Soc Ser A, General 120:3, 253–281

Färe R (1975) Efficiency and the production function. Z Nationalökonomie 35:317–324

Färe R, Grosskopf S (2010) Directional distance functions and slack-based measures of efficiency. Eur J Oper Res 200:320–322

Färe R, Lovell CAK (1978) Measuring the technical efficiency of production. J Econ Theory 19:150–162

Färe R, Lundberg A (2006) Parameterizing the shortage function, Mimeo

Färe R, Primont D (1995) Multi-output production and duality: theory and applications. Kluwer Academic Publishers, Boston

Färe R, Sung KJ (1986) On second order Taylor's series approximations and linear homogeneity. Aequationes Mathematicae 30:180–186

Färe R, Grosskopf S, Whittaker G (2007a) Distance functions: with applications to DEA. In: Zhu J, Cook WD (eds) Modeling data structures, irregularities and structural complexities in DEA. Springer, New York

Färe R, Grosskopf S, Zelenyuk V (2007b) Finding common ground: efficiency indices. In: Färe R, Grosskopf S, Primont D (eds) Aggregation, efficiency and measurement. Springer, New York

Färe R, Grosskopf S, Margaritis D (2008) Efficiency and productivity: malmquist and more. In: Fried H, Lovell CK, Schmidt S (eds) The measurement of productive efficiency and productivity. Oxford University Press, New York

Färe R, Grosskopf S, Pasurka C (2013a) On nonparametric estimation: with focus on agriculture. Ann Rev Resour Econ 5:93–110

Färe R, Grosskopf S, Whittaker G (2013b) Directional distance functions: endogenous directions based on endogenous normalization constraints. J Product Anal 40:267–269

Kemeny JG, Morgenstern O, Thompson GL (1956) A generalization of the von Neumann model of expanding economy. Econometrica 24:115–135
Kolm SC (1976) Unequal inequalities II. J Econ Theory 13:82–111
Luenberger DG (1969) Optimization by vector space methods. Wiley, New York
Luenberger DG (1992) Benefit functions and duality. J Math Econ 21:461–481
Mahler K (1939) Ein Übertragungsprinzip für konvexe körper. Casopis pro Pestovani Matematiky a Fysiky 64:93–102
Russell RR, Schworm W (2009) Axiomatic foundations of efficiency measurement on data-generated technologies. J Product Anal 31:77–86
Shephard RW (1953) Cost and production functions. Princeton University Press, Princeton
Tone K (2001) A slack-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509

# Chapter 2
# DEA Cross Efficiency

**Wade D. Cook and Joe Zhu**

**Abstract** Data envelopment analysis (DEA) provides a relative efficiency measure for peer decision making units (DMUs) with multiple inputs and outputs. While DEA has been proven an effective approach in identifying the best practice frontiers, its flexibility in weighting multiple inputs and outputs and its nature of self-evaluation have been criticized. The cross efficiency method was developed as a DEA extension to rank DMUs with the main idea being to use DEA to do peer evaluation, rather than in pure self-evaluation mode. However, cross efficiency scores obtained from the original DEA model are generally not unique, and depend on which of the alternate optimal solutions to the DEA linear programs is used. The current chapter discusses various cross efficiency approaches in dealing with non-unique solutions from DEA

**Keywords** Data Envelopment Analysis (DEA) · Cross efficiency · Multiplicative · Cobb-Douglas

## 2.1 Introduction

While DEA has been proven an effective approach in identifying best practice frontiers, its flexibility in weighting multiple inputs and outputs and its nature of self-evaluation have been criticized. The cross efficiency method was developed as a DEA extension to rank DMUs (Sexton et al. 1986), with the main idea being to use DEA to do peerevaluation, rather than to have it operate in a pure self-evaluation mode. Cross efficiency has been further investigated by Doyle and Green (1994). There are mainly two advantages of the cross-evaluation method. It provides an ordering among DMUs, and it eliminates unrealistic weight schemes without requiring the elicitation of weight restrictions from application area experts (e.g., Anderson et al. 2002).

W. D. Cook (✉)
Schulich School of Business, York University, M3J 1P3 Toronto, ON, Canada
e-mail: wcook@schulich.yorku.ca

J. Zhu
School of business, Worcester Polytechnic Institute, 01609 Worcester, MA, USA

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_2

Cross efficiency evaluation has been used in various applications, e.g., efficiency evaluations of nursing homes (Sexton et al. 1986), R&D project selection (Oral et al. 1991), preference voting (Green et al. 1996), and others. However, as noted in Doyle and Green (1994), the non-uniqueness of the DEA optimal weights/multipliers possibly reduces the usefulness of cross efficiency. Specifically, cross efficiency scores obtained from the original DEA methodology are generally not unique. Thus, depending on which of the alternate optimal solutions to the DEA linear programs is used, it may be possible to improve a DMU's (cross efficiency) performance rating, but generally only by worsening the ratings of others. With that in mind, Sexton et al. (1986) and Doyle and Green (1994) propose the use of a secondary goal to deal with the non-unique DEA solutions. They developed aggressive (benevolent) model formulations to identify optimal weights that not only maximize the efficiency of a particular DMU under evaluation, but also minimize (maximize) the average efficiency of other DMUs

In the current chapter, we present the standard DEA cross efficiency method, and discuss several approaches that have been developed to address the non-uniqueness issue dicussed above. These approaches include the game cross efficiency methodology of Liang et al. (2008a) and the maximum cross efficiency concept (Cook and Zhu 2014) based upon a set of log-linear DEA models.

## 2.2 Cross Efficiency

Suppose we have a set of $n$ DMUs and each $DMU_j$ have $s$ different outputs and $m$ different inputs. We denote the $i$th input and $r$th output of $DMU_j (j = 1,2,\ldots,n)$ as $x_{ij} (i = 1,\ldots,m)$ and $y_{rj} (r = 1,\ldots,s)$, respectively. Cross efficiency is generally presented as a two-phase process. Specifically, phase 1 is the self-evaluation phase where DEA scores are calculated using the constant returns-to-scale (CRS) DEA model of Charnes et al. (1978). In the second phase, the multipliers arising from phase 1 are applied to all peer DMUs to arrive at the so-called cross evaluation score for each of those DMUs.

Phase 1: Suppose $DMU_d$ is under evaluation by the CRS model (Charnes et al. 1978). Then that DMU's (self-evaluation) efficiency score is determined by the following DEA model

$$\begin{aligned} \text{Max} \quad & E_{dd} = \frac{\sum_{r=1}^{s} u_{rd} y_{rd}}{\sum_{i=1}^{m} v_{id} x_{id}} \\ s.t. \quad & E_{dj} = \frac{\sum_{r=1}^{s} u_{rd} y_{rj}}{\sum_{i=1}^{m} v_{id} x_{ij}} \leq 1, j = 1,2,\ldots,n. \\ & u_{rd} \geq 0, r = 1,\ldots,s. \\ & v_{id} \geq 0, i = 1,\ldots,m. \end{aligned} \tag{2.1}$$

where $v_{id}$ and $u_{rd}$ represent $i$th input and $r$th output weights for $DMU_d$.

Phase 2: The cross efficiency of $DMU_j$, using the weights that $DMU_d$ has chosen in model (2.1), is given by

$$E_{dj} = \frac{\sum_{r=1}^{s} u_{rd}^* y_{rj}}{\sum_{i=1}^{m} v_{id}^* x_{ij}}, d, j = 1,2,\ldots,n \tag{2.2}$$

where (*) denotes optimal values in model (2.1). For $DMU_j (j = 1, 2, \cdots, n)$, an average of all $E_{dj} (d = 1,2,\ldots,n)$,

$$\overline{E}_j = \frac{1}{n} \sum_{d=1}^{n} E_{dj}, \tag{2.3}$$

is referred to as the *cross efficiency score for* $DMU_j$.

We should point out that each individual $E_{dj}$ is called cross efficiency and the average defined in (2.3) is also called cross efficiency in the DEA literature. In general, "cross efficiency" refers to the average defined in (2.3), not the individual scores defined in (2.2).

While the DEA model (2.1) is a non-linear model, model (2.1) is usually solved in its equivalent multiplier model,

$$\max \quad E_{dd} = \sum_{r=1}^{s} u_{rd} y_{rd}$$

Subject to

$$\begin{aligned} &\sum_{i=1}^{m} v_{id} x_{id} = 1 \\ &\sum_{r=1}^{s} u_{rd} y_{rj} - \sum_{i=1}^{m} v_{id} x_{ij} \le 0 \quad j = 1,\ldots.,n \\ &u_{rd}, v_{id} \ge 0 \end{aligned} \tag{2.4}$$

Due to the fact that the above cross efficiency is based upon input-oriented models, cross efficiency scores are not greater than one.

We here briefly illustrate the concept of cross efficiency by adopting the cross efficiency matrix from Doyle and Green (1994). In Fig. 2.1, we have six DMUs. $E_{dj}$ is the (cross) efficiency of $DMU_j$ based upon a set of DEA weights calculated for $DMU_d$. This set of DMU weights gives the best efficiency score for $DMU_d$ under evaluation by a DEA model, and $E_{dd}$ (in the leading diagonal) is the DEA efficiency for $DMU_k$. The cross efficiency for a given $DMU_j$ is defined as the arithmetic average down column j, given by $\bar{E}_j$. (We point out that in Doyle and Green (1994), the efficiency score for DMU $k$ is not included as part of the average.)

Obviously, $E_{dj}$ $(d \neq j)$ and $\bar{E}_j$ are not unique due to the often-present multiple optimal DEA weights in model (2.4), for example. As a result of this non uniqueness, the cross efficiency concept has been criticized as unreliable.

| Rating DMU \ Rated DMU | 1 | 2 | 3 | 4 | 5 | 6 | Averaged appraisal of peers |
|---|---|---|---|---|---|---|---|
| 1 | $\mathbf{E_{11}}$ | $E_{12}$ | $E_{13}$ | $E_{14}$ | $E_{15}$ | $E_{16}$ | $A_1$ |
| 2 | $E_{21}$ | $\mathbf{E_{22}}$ | $E_{23}$ | $E_{24}$ | $E_{25}$ | $E_{26}$ | $A_2$ |
| 3 | $E_{31}$ | $E_{32}$ | $\mathbf{E_{33}}$ | $E_{34}$ | $E_{35}$ | $E_{36}$ | $A_3$ |
| 4 | $E_{41}$ | $E_{42}$ | $E_{43}$ | $\mathbf{E_{44}}$ | $E_{45}$ | $E_{46}$ | $A_4$ |
| 5 | $E_{51}$ | $E_{52}$ | $E_{53}$ | $E_{54}$ | $\mathbf{E_{55}}$ | $E_{56}$ | $A_5$ |
| 6 | $E_{61}$ | $E_{62}$ | $E_{63}$ | $E_{64}$ | $E_{65}$ | $\mathbf{E_{66}}$ | $A_6$ |
| | $\overline{E}_1$ | $\overline{E}_2$ | $\overline{E}_3$ | $\overline{E}_4$ | $\overline{E}_5$ | $\overline{E}_6$ | |

Averaged appraisal by peers (peer appraisal)

**Fig. 2.1** Cross efficiency matrix. (Doyle and Green 1994)

Note that the above discussion is based upon input-orientation. Similarly, we can use output-oriented models to calculate cross efficiency. In this case, $E_{dj}$ in (2.2) becomes

$$E_{dj} = \frac{\sum_{i=1}^{m} v_{id}^* x_{ij}}{\sum_{r=1}^{s} u_{rd}^* y_{rj}} \tag{2.5}$$

where $v_{id}^*$ and $u_{rd}^*$ are optimal values in the following output-oriented model when $DMU_d$ is under evaluation

$$\begin{aligned} \min \quad & E_{dd} = \frac{\sum_{i=1}^{m} v_{id} x_{id}}{\sum_{r=1}^{s} u_{rd} y_{rd}} \\ s.t. \quad & E_{dj} = \frac{\sum_{i=1}^{m} v_{id} x_{ij}}{\sum_{r=1}^{s} u_{rd} y_{rj}} \geq 1, j = 1,2,\ldots,n \\ & v_{id} \geq 0, i = 1,\ldots,m \\ & u_{rd} \geq 0, r = 1,\ldots,s. \end{aligned} \tag{2.6}$$

The above model (2.6) is equivalent to the following output-oriented CRS multiplier model:

$$\text{Min} \quad \sum_{r=1}^{s} v_{id} x_{id}$$

subject to

$$\begin{aligned} & \sum_{i=1}^{m} v_{id} x_{ij} - \sum_{r=1}^{s} u_{rd} y_{rj} \geq 0, j = 1,2,\ldots,n \\ & \sum_{i=1}^{m} u_{rd} y_{rd} = 1 \\ & v_{id} \geq 0, i = 1,\ldots,m \\ & u_{rd} \geq 0, r = 1,\ldots,s \end{aligned} \tag{2.7}$$

**Table 2.1** Numerical example

| DMU | Input 1 | Input 2 | Input 3 | Output 1 | Output 2 |
|---|---|---|---|---|---|
| 1 | 7 | 7 | 7 | 4 | 4 |
| 2 | 5 | 9 | 7 | 7 | 7 |
| 3 | 4 | 6 | 5 | 5 | 7 |
| 4 | 5 | 9 | 8 | 6 | 2 |
| 5 | 6 | 8 | 5 | 3 | 6 |

**Table 2.2** Input-oriented CRS efficiency and optimal multipliers

| DMU | CRS efficiency | x1 | x2 | x3 | y1 | y2 |
|---|---|---|---|---|---|---|
| 1 | 0.68571 | 0.00000 | 0.14286 | 0.00000 | 0.17143 | 0.00000 |
| 2 | 1.00000 | 0.07143 | 0.07143 | 0.00000 | 0.14286 | 0.00000 |
| 3 | 1.00000 | 0.00000 | 0.16667 | 0.00000 | 0.00000 | 0.14286 |
| 4 | 0.85714 | 0.07143 | 0.07143 | 0.00000 | 0.14286 | 0.00000 |
| 5 | 0.85714 | 0.00000 | 0.00000 | 0.20000 | 0.00000 | 0.14286 |

Note that under the output-oriented case, all cross efficiency scores are not less than one, as the output-oriented CRS efficiency score is not less than one. Then, the output-oriented DEA cross efficiency score can be defined in a similar manner as in (4.3).

Finally, the above discussion is based upon CRS. Similar developments can be obtained under non-CRS situations. We, however, point out that negative cross efficiency scores can be obtained under non-CRS conditions, for example, variable returns-to-scale (VRS) (see Lim and Zhu (in press) or Chap. 3.

## 2.3 Numerical Example

Throughout the chapter, we use the numerical example shown in Table 2.1 to illustrate various cross efficiency approaches. This example is from Liang et al. (2008a) and has five DMUs with three inputs and two outputs.

Table 2.2 reports the CRS efficiency scores obtained from model (2.4) along with a set of optimal multipliers. Based upon this set of multipliers, an input-oriented cross efficiency matrix is provided in Table 2.3. Tables 2.4 and 2.5 report cross efficiency results based upon the output-oriented model (2.7). As we can see, unlike the standard CRS efficiency scores, the input-oriented cross efficiency score is not (always) the reciprocal of the associated output-oriented cross efficiency score.

**Table 2.3** Input-oriented standard cross efficiency matrix

| | Cross efficiency matrix | | | | |
|---|---|---|---|---|---|
| Rating DMU | DMU1 | DMU2 | DMU3 | DMU4 | DMU5 |
| 1 | 0.68571 | 0.93333 | 1.00000 | 0.80000 | 0.45000 |
| 2 | 0.57143 | 1.00000 | 1.00000 | 0.85714 | 0.42857 |
| 3 | 0.48980 | 0.66667 | 1.00000 | 0.19048 | 0.64286 |
| 4 | 0.57143 | 1.00000 | 1.00000 | 0.85714 | 0.42857 |
| 5 | 0.40816 | 0.71429 | 1.00000 | 0.17857 | 0.85714 |
| Cross efficiency | 0.54531 | 0.86286 | 1.00000 | 0.57667 | 0.56143 |

**Table 2.4** Output-oriented CRS efficiency and optimal multipliers

| DMU | CRS efficiency | x1 | x2 | x3 | y1 | y2 |
|---|---|---|---|---|---|---|
| 1 | 1.45833 | 0.00000 | 0.20833 | 0.00000 | 0.25000 | 0.00000 |
| 2 | 1.00000 | 0.07143 | 0.07143 | 0.00000 | 0.14286 | 0.00000 |
| 3 | 1.00000 | 0.00000 | 0.16667 | 0.00000 | 0.00000 | 0.14286 |
| 4 | 1.16667 | 0.08333 | 0.08333 | 0.00000 | 0.16667 | 0.00000 |
| 5 | 1.16667 | 0.00000 | 0.00000 | 0.23333 | 0.00000 | 0.16667 |

**Table 2.5** Output-oriented standard cross efficiency matrix

| CRS | Cross efficiency matrix | | | | |
|---|---|---|---|---|---|
| Rating DMU | DMU1 | DMU2 | DMU3 | DMU4 | DMU5 |
| 1 | 1.45833 | 1.07143 | 1.00000 | 1.25000 | 2.22222 |
| 2 | 1.75000 | 1.00000 | 1.00000 | 1.16667 | 2.33333 |
| 3 | 2.04167 | 1.50000 | 1.00000 | 5.25000 | 1.55556 |
| 4 | 1.75000 | 1.00000 | 1.00000 | 1.16667 | 2.33333 |
| 5 | 2.45000 | 1.40000 | 1.00000 | 5.60000 | 1.16667 |
| Cross efficiency | 1.89000 | 1.19429 | 1.00000 | 2.88667 | 1.92222 |

## 2.4 Maximum Log Cross Efficiency

To address the non-uniqueness in cross efficiency, the idea of secondary goals was introduced, with the original proposal being to maximize or minimize the average appraisal of peers as indicated by $A_k$ in Fig. 2.1. Specifically, $A_k$ is the arithmetic average across the row $k$. However, due to the DEA model (CCR multiplier model) used, $E_{kj} = \frac{\sum_r \mu_{rk} y_{rj}}{\sum_i v_{ik} x_{ij}}$, where $y_{rj}$, $(r = 1, 2, \ldots, s)$ are outputs and $x_{ij}, (i = 1, 2, \ldots, m)$ are inputs for $DMU_j$, and $\mu_{rk}$, $v_{ik}$ are corresponding output and input weights chosen by $DMU_k$.

Thus, $A_k = \frac{1}{n}\sum_j E_{kj}$ appears in the form of a non-linear fractional problem that cannot be converted into linear format. To remedy this, Sexton et al. (1986), and Doyle and Green (1994) suggested the use of linear surrogates for the secondary goal in form of the numerators in $E_{kj}$ minus the sum of the denominators, and modified ratios that can be converted into linear relations. However, due to the fact that these surrogates are not equivalent to the optimal values of $A_k$, the resulting cross efficiency scores are, at best, approximations of these optimal values.

While such approaches as those of Doyle and Green (1994), and those suggested by others (e.g., Liang et al. 2008b), help to reduce the variability of DEA optimal weights, these approaches all produce cross efficiency scores that differ from one another. Cook and Zhu (2014), on the other hand, propose to use multiplicative DEA models developed in Charnes et al. (1982) and Charnes et al. (1983) to obtain maximum (*and unique*) cross efficiency scores under the condition that each DMU's DEA efficiency score remains unchanged. To introduce the Cook and Zhu (2014) approach, we need first to present the multiplicative DEA models.

## 2.5 Multiplicative DEA Model

Charnes et al. (1982) introduce the following multiplicative DEA model when $DMU_o$ is under evaluation

$$\begin{aligned} \max \quad & \frac{\prod_{r=1}^{s} y_{ro}^{\mu_r}}{\prod_{i=1}^{m} x_{io}^{v_i}} \\ s.t. \quad & \frac{\prod_{r=1}^{s} y_{rj}^{\mu_r}}{\prod_{i=1}^{m} x_{ij}^{v_i}} \leq 1, \quad j = 1, \ldots, n \\ & \mu_r, v_i \geq 1 \end{aligned} \tag{2.8}$$

Taking logarithms (to any base), model (2.8) becomes

$$\max \sum_{r=1}^{s} \mu_r \hat{y}_{ro} - \sum_{i=1}^{m} v_i \hat{x}_{io}$$

subject to

$$\begin{aligned} & \sum_{r=1}^{s} \mu_r \hat{y}_{rj} - \sum_{i=1}^{m} v_i \hat{x}_{ij} \leq 0 \\ & \mu_r, v_i \geq 1 \end{aligned} \tag{2.9}$$

where ($\hat{}$) denotes logarithms.

The dual to model (2.9) can be written as

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \hat{x}_{ij} + s_i^- = \hat{x}_{io} \quad i = 1,2,\ldots,m; \\ &\sum_{j=1}^{n} \lambda_j \hat{y}_{rj} - s_r^+ = \hat{y}_{ro} \quad r = 1,2,\ldots,s; \\ &\lambda_j, s_i^-, s_r^+ \geq 0 \end{aligned} \tag{2.10}$$

It can be seen that model (2.10) is actually the CRS additive model. We therefore call model (2.8) (and its equivalents) the CRS multiplicative DEA model.

Charnes et al. (1983) introduce the following multiplicative DEA model when $DMU_o$ is under evaluation

$$\begin{aligned} \max \quad & \frac{e^{\eta} \prod_{r=1}^{s} y_{rk}^{\mu_r}}{e^{\xi} \prod_{i=1}^{m} x_{ik}^{\nu_i}} \\ s.t. \quad & \frac{e^{\eta} \prod_{r=1}^{s} y_{rj}^{\mu_r}}{e^{\xi} \prod_{i=1}^{m} x_{ij}^{\nu_i}} \leq 1, \quad j = 1,\ldots,n \\ & \eta, \xi \geq 0,\ \mu_r, \nu_i \geq 1 \end{aligned} \tag{2.11}$$

Taking logarithms (to any base), model (2.11) becomes

$$\max \quad \eta - \xi + \sum_{r=1}^{s} \mu_r \hat{y}_{ro} - \sum_{i=1}^{m} \nu_i \hat{x}_{io}$$

subject to

$$\begin{aligned} &\eta - \xi + \sum_{r=1}^{s} \mu_r \hat{y}_{rj} - \sum_{i=1}^{m} \nu_i \hat{x}_{ij} \leq 0 \\ &\eta, \xi \geq 0 \\ &\mu_r, \nu_i \geq 1 \end{aligned} \tag{2.12}$$

where ($\hat{}$) denotes logarithms.

The dual to model (2.12) is

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\begin{aligned}
&\sum_{j=1}^{n} \lambda_j \hat{x}_{ij} + s_i^- = \hat{x}_{io}\, i = 1,2,\ldots,m; \qquad (2.13)\\
&\sum_{j=1}^{n} \lambda_j \hat{y}_{rj} - s_r^+ = \hat{y}_{ro}\; r = 1,2,\ldots,s;\\
&\sum_{j=1}^{n} \lambda_j = 1\\
&\lambda_j, s_i^-, s_r^+ \geq 0
\end{aligned}$$

Obviously, model (2.13) is the VRS version of the additive model. We therefore call model (2.11) (and its equivalent) the VRS multiplicative DEA model.

The above two multiplicative DEA models identify Cobb-Douglas production functions directly from observations (see Charnes et al. (1982, 1983) for more discussions.)

Model (2.8) or (2.9) yields the best efficiency score for $DMU_o$ with a set of "weights" chosen by $DMU_o$. Denote an optimal set of weights by $\mu_{ro}^*, \nu_{io}^*$, and the efficiency score from (2.8) as $\theta_o^*$. Then cross efficiency of $DMU_j$ using the weights that $DMU_o$ has chosen, is given by

$$E_{oj} = \frac{\prod_{r=1}^{s} y_{rj}^{\mu_{ro}^*}}{\prod_{i=1}^{m} x_{ij}^{\nu_{io}^*}} \qquad (2.14)$$

The efficiency score for $DMU_o$ obtained from model (2.8) is $E_{oo} = \theta_o^*$

Then we define the following *geometric average* peer appraisal cross efficiency score as the CRS multiplicative cross efficiency score

$$\bar{E}_j = \left(\prod_{k=1}^{n} E_{kj}\right)^{1/n} = \left(\prod_{k=1}^{n} \frac{\prod_{r=1}^{s} y_{rj}^{\mu_{rk}^*}}{\prod_{i=1}^{m} x_{ij}^{\nu_{ik}^*}}\right)^{1/n} \qquad (2.15)$$

The VRS multiplicative cross efficiency score can be defined in a similar manner. Specifically, for a $DMU_k$ under evaluation of model (2.11), we have

$$E_{kj} = \frac{e^{\eta_k^*} \prod_{r=1}^{s} y_{rj}^{\mu_{rk}^*}}{e^{\xi_k^*} \prod_{i=1}^{m} x_{ij}^{\nu_{ik}^*}} \qquad (2.16)$$

as the cross efficiency of $DMU_j$ using the weights that $DMU_k$ has chosen, where $\eta_k^*, \xi_k^*, \mu_{rk}^*, \nu_{ik}^*$ are optimal solutions from model (2.11) or (2.12).

Then we define the following *geometric average* peer appraisal VRS multiplicative cross efficiency score as

$$\bar{E}_j = \left(\prod_{k=1}^{n} E_{kj}\right)^{1/n} = \left(\prod_{k=1}^{n} \frac{e^{\eta_k^*} \prod_{r=1}^{s} y_{rj}^{\mu_{rk}^*}}{e^{\xi_k^*} \prod_{i=1}^{m} x_{ij}^{\nu_{ik}^*}}\right)^{1/n} \tag{2.17}$$

where $E_{kk}$ is the optimal value to model (2.11) or (2.12).

## 2.6 Maximum Log Cross Efficiency

We now present the approach developed in Cook and Zhu (2014). These authors point out that one can maximize the average cross efficiency score $\bar{E}_j$(defined in (2.15)) subject to the condition that $E_{kk} = \theta_k^*$ for all $k = 1, \ldots, n$. Specifically, for $DMU_{j_o}$ we have

$$\begin{aligned}
\max \quad & \left(\prod_{k=1}^{n} \frac{\prod_{r=1}^{s} y_{rj_o}^{\mu_{rk}}}{\prod_{i=1}^{m} x_{ij_o}^{\nu_{ik}}}\right)^{1/n} \\
\text{s.t.} \quad & \frac{\prod_{r=1}^{s} y_{rj}^{\mu_{rk}}}{\prod_{i=1}^{m} x_{ij}^{\nu_{ik}}} \le 1, \quad j = 1, \ldots, n, k = 1, \ldots n \\
& E_{kk} = \frac{\prod_{r=1}^{s} y_{rk}^{\mu_{rk}}}{\prod_{i=1}^{m} x_{ik}^{\nu_{ik}}} = \theta_k^*, k = 1, \ldots, n \\
& \mu_{\mathrm{rk}}, \nu_{ik} \ge 1, \quad k = 1, \ldots n; i - 1, \ldots, m; r = 1, \ldots, s
\end{aligned} \tag{2.18}$$

Making logarithmic transformations in (2.18), we arrive at the following linear program

$$\begin{aligned}
\max \quad & \frac{1}{\mathrm{n}} \left(\sum_{k} \sum_{r} \mu_{rk} \hat{y}_{r j_o} - \sum_{k} \sum_{i} \nu_{ik} \hat{x}_{i j_0}\right) \\
\text{s.t.} \quad & \sum_{r=1}^{s} \mu_{rk} \hat{y}_{rj} - \sum_{i=1}^{m} \nu_{ik} \hat{x}_{ij} \le 0, \quad j, k = 1, \ldots, n \\
& \sum_{r=1}^{s} \mu_{rk} \hat{y}_{rk} - \sum_{i=1}^{m} \nu_{ik} \hat{x}_{ik} = \ln(\theta_k^*), \quad k = 1, \ldots, n \\
& \mu_{\mathrm{rk}}, \nu_{ik} \ge 1, \quad k = 1, \ldots n; i - 1, \ldots, m; r = 1, \ldots, s
\end{aligned} \tag{2.19}$$

where "$\wedge$" denotes data in logarithmic form. Since logarithms are used in the process, we call this type of cross efficiency (the optimal value to model (2.19)) "Maximum Log Cross Efficiency".

Cook and Zhu (2014) point out that the attractive feature of the proposed multiplicative approach is that the resulting cross efficiency score (the objective function in model (2.19)) is uniquely determined; this is not the case for any of the other approaches taken up to now. Specifically, in the standard cross efficiency, the fact that alternate optimal solutions can occur, gives rise to non-unique peer ratings for a DMU, and hence the average of these (the cross efficiency score for that DMU) is not uniquely determined. It is noted that while there may as well be alternate optimal solutions $\mu_{rk}^*, v_{ik}^*$ yielding this unique optimal value in (2.19), this fact in and of itself is immaterial. It is the uniqueness of the cross efficiency score, not the multipliers that lead to it, that matters.

In addition to the above development based upon CRS, we have the following VRS maximum log cross efficiency model

$$\begin{aligned}
\max \quad & \left(\prod_{k=1}^{n} \frac{e^{\eta_k} \prod_{r=1}^{s} y_{rj_o}^{\mu_{rk}}}{e^{\xi_k} \prod_{i=1}^{m} x_{ij_o}^{v_{ik}}}\right)^{1/n} \\
\text{s.t.} \quad & \frac{e^{\eta_k} \prod_{r=1}^{s} y_{rj}^{\mu_{rk}}}{e^{\xi_k} \prod_{i=1}^{m} x_{ij}^{v_{ik}}} \le 1, \quad j = 1, \ldots, n, k = 1, \ldots n \\
& E_{kk} = \frac{e^{\eta_k} \prod_{r=1}^{s} y_{rk}^{\mu_{rk}}}{e^{\xi_k} \prod_{i=1}^{m} x_{ik}^{v_{ik}}} = \theta_k^*, k = 1, \ldots, n \\
& \eta_k, \xi_k \ge 0, \mu_{\mathrm{rk}}, v_{ik} \ge 1, \quad k = 1, \ldots n; i = 1, \ldots, m; r = 1, \ldots s
\end{aligned} \tag{2.20}$$

Making logarithmic transformations in (2.20), we arrive at the following linear program

$$\begin{aligned}
\max \quad & \frac{1}{\mathrm{n}} \left(\sum_k \eta_k + \sum_k \sum_r \mu_{rk} \hat{y}_{r j_o} - \sum_k \xi_k - \sum_k \sum_i v_{ik} \hat{x}_{i j_0}\right) \\
\text{s.t.} \quad & \eta_k + \sum_{r=1}^{s} \mu_{rk} \hat{y}_{rj} - \xi_k - \sum_{i=1}^{m} v_{ik} \hat{x}_{ij} \le 0, \quad j, k = 1, \ldots, n \\
& \eta_k + \sum_{r=1}^{s} \mu_{rk} \hat{y}_{rk} - \xi_k - \sum_{i=1}^{m} v_{ik} \hat{x}_{ik} = \ln(\theta_k^*), \quad k = 1, \ldots, n \\
& \eta_k, \xi_k \ge 0, \mu_{\mathrm{rk}}, v_{ik} \ge 1, \quad k = 1, \ldots n; i = 1, \ldots, m; r = 1, \ldots s
\end{aligned} \tag{2.21}$$

where "$\wedge$" denotes data in logarithmic form.

To demonstrate the above approach, we apply it to the numerical example in Table 2.1. Table 2.6 reports the results from models (2.12) and (2.18) under CRS. Table 2.7 reports standard (multiplicative) cross efficiency matrix based upon (2.15). Table 2.8 reports the maximum Log cross efficiency matrix under CRS.

We next apply the numerical example in Table 2.1 to the VRS model. The efficiency scores from model (2.12) are reported in the first column of Table 2.9.

**Table 2.6** CRS multiplicative results

| DMU | Efficiency from model (2.12) | Standard cross efficiency (2.15) | Maximum log cross efficiency (2.18) |
|---|---|---|---|
| 1 | 0.1348 | 0.0543 | 0.0563 |
| 2 | 1 | 0.6704 | 0.7603 |
| 3 | 1 | 1 | 1 |
| 4 | 0.1314 | 0.0841 | 0.1013 |
| 5 | 0.2332 | 0.0763 | 0.0763 |

**Table 2.7** Standard CRS multiplicative log cross efficiency matrix

| | Rated DMU | | | | |
|---|---|---|---|---|---|
| Rating DMU | 1 | 2 | 3 | 4 | 5 |
| 1 | **0.1348** | 0.6900 | 1 | 0.1314 | 0.1739 |
| 2 | 0.0021 | **1** | 1 | 0.0910 | 0.0016 |
| 3 | 0.1122 | 0.5333 | **1** | 0.0517 | 0.2332 |
| 4 | 0.1348 | 0.6900 | 1 | **0.1314** | 0.1739 |
| 5 | 0.1122 | 0.5333 | 1 | 0.0517 | **0.2332** |
| $\overline{E}_j$ defined in (15) | 0.0543 | 0.6704 | 1 | 0.0841 | 0.0763 |

**Table 2.8** Maximum log cross efficiency matrix under CRS

| | Rated DMU | | | | |
|---|---|---|---|---|---|
| Rating DMU | 1 | 2 | 3 | 4 | 5 |
| 1 | **0.1348** | 0.6900 | 1 | 0.1314 | 0.1740 |
| 2 | 0.0021 | **1** | 1 | 0.0910 | 0.0016 |
| 3 | 0.1348 | 1 | **1** | 0.1314 | 0.23323 |
| 4 | 0.1348 | 0.6901 | 1 | **0.1314** | 0.1739 |
| 5 | 0.1122 | 0.5334 | 1 | 0.0517 | **0.2332** |
| $\overline{E}_j$ (maximum) | 0.0563 | 0.7603 | 1 | 0.1013 | 0.0763 |

Column 2 reports the cross efficiency using (2.17), and the he last column reports the maximum cross efficiency scores based upon model (2.21). The related standard cross efficiency matrix is displayed in Table 2.10. It can be seen that the standard cross efficiency scores of DMUs 1 and 5 are at their maximum. Table 2.11 presents the cross efficiency scores as calculated by model (2.21).

We finally apply our approach to a data set of 37 project proposals relating to the Turkish iron and steel industry (see Oral et al. 1991). Each project is characterized by

**Table 2.9** VRS multiplicative results

| DMU | Efficiency model (2.1) | Standard cross efficiency | Maximum cross efficiency |
|---|---|---|---|
| 1 | 0.1599 | 0.1354 | 0.1354 |
| 2 | 1 | 0.6858 | 0.7777 |
| 3 | 1 | 1 | 1 |
| 4 | 0.1607 | 0.1310 | 0.1402 |
| 5 | 0.2571 | 0.1755 | 0.1755 |

**Table 2.10** Standard VRS multiplicative log cross efficiency matrix

| | Rated DMU | | | | |
|---|---|---|---|---|---|
| Rating DMU | 1 | 2 | 3 | 4 | 5 |
| 1 | **0.1599** | 0.5333 | 1 | 0.1143 | 0.2571 |
| 2 | 0.1054 | **1** | 1 | 0.1607 | 0.0990 |
| 3 | 0.1599 | 0.5333 | **1** | 0.1143 | 0.2571 |
| 4 | 0.1054 | 1 | 1 | **0.1607** | 0.0990 |
| 5 | 0.1599 | 0.5333 | 1 | 0.1143 | **0.2571** |
| $\overline{E}_j$ defined in (17) | 0.1354 | 0.6858 | 1 | 0.1310 | 0.1755 |

**Table 2.11** Maximum log cross efficiency matrix under VRS

| | Rated DMU | | | | |
|---|---|---|---|---|---|
| Rating DMU | 1 | 2 | 3 | 4 | 5 |
| 1 | **0.1599** | 0.5333 | 1 | 0.1143 | 0.2571 |
| 2 | 0.1054 | **1** | 1 | 0.1607 | 0.0990 |
| 3 | 0.1599 | 1 | **1** | 0.1607 | 0.2571 |
| 4 | 0.1054 | 1 | 1 | **0.1607** | 0.0990 |
| 5 | 0.1599 | 0.5333 | 1 | 0.1143 | **0.2571** |
| $\overline{E}_j$ defined in (3) | 0.1354 | 0.7777 | 1 | 0.1402 | 0.1755 |

five output measures: direct economic contribution, indirect economic contribution, technological contribution, scientific contribution and social contribution. The single input is the cost. The results are reported in Table 2.12 where columns 2 and 3 report the multiplicative efficiency and its related standard Log cross efficiency based upon model (2.12), and column 4 reports the maximum cross efficiency based upon model (2.21).

**Table 2.12** Log cross efficiency for project selection

| Project | Multiplicative efficiency | Standard log cross efficiency | Maximum log cross efficiency |
|---|---|---|---|
| 1 | 1 | 0.2654 | 0.3160 |
| 2 | 0.5579 | 0.0606 | 0.0764 |
| 3 | 0.0003 | 0.0000 | 0.0000 |
| 4 | 0.0398 | 0.0001 | 0.0002 |
| 5 | 0.1210 | 0.0016 | 0.0025 |
| 6 | 0.1598 | 0.0076 | 0.0115 |
| 7 | 0.0883 | 0.0011 | 0.0020 |
| 8 | 0.1312 | 0.0008 | 0.0013 |
| 9 | 1 | 0.0117 | 0.0202 |
| 10 | 0.2293 | 0.0031 | 0.0039 |
| 11 | 0.3043 | 0.0071 | 0.0089 |
| 12 | 0.0387 | 0.0000 | 0.0000 |
| 13 | 0.0503 | 0.0002 | 0.0003 |
| 14 | 1 | 0.3983 | 0.4464 |
| 15 | 0.1956 | 0.0103 | 0.0160 |
| 16 | 0.1908 | 0.0010 | 0.0013 |
| 17 | 1 | 0.2103 | 0.2198 |
| 18 | 0.1584 | 0.0005 | 0.0007 |
| 19 | 0.1493 | 0.0010 | 0.0013 |
| 20 | 0.0063 | 0.0000 | 0.0000 |
| 21 | 1 | 0.0218 | 0.0282 |
| 22 | 0.3407 | 0.0091 | 0.0115 |
| 23 | 1 | 0.2253 | 0.2606 |
| 24 | 1 | 0.0272 | 0.0361 |
| 25 | 0.0295 | 0.0000 | 0.0000 |
| 26 | 0.5612 | 0.0459 | 0.0542 |
| 27 | 0.5744 | 0.0372 | 0.0417 |
| 28 | 0.0505 | 0.0001 | 0.0001 |
| 29 | 0.2665 | 0.0066 | 0.0081 |
| 30 | 0.2165 | 0.0027 | 0.0034 |
| 31 | 1 | 0.0009 | 0.0014 |
| 32 | 0.0384 | 0.0000 | 0.0000 |

**Table 2.12** (continued)

| Project | Multiplicative efficiency | Standard log cross efficiency | Maximum log cross efficiency |
|---|---|---|---|
| 33 | 0.0122 | 0.0000 | 0.0000 |
| 34 | 1 | 0.0000 | 0.0000 |
| 35 | 1 | 0.0039 | 0.0053 |
| 36 | 1 | 0.0479 | 0.0556 |
| 37 | 1 | 0.0056 | 0.0083 |

**Table 2.13** Selected projects

| Project | Budget | Game cross efficiency |
|---|---|---|
| 14 | 95 | Yes |
| 1 | 84.2 | Yes |
| 23 | 75.6 | Yes |
| 17 | 32.1 | Yes |
| 2 | 90 | No |
| 36 | 64.1 | Yes |
| 26 | 69.3 | Yes |
| 27 | 57.1 | Yes |
| 24 | 92.3 | No |
| 21 | 74.4 | Yes |
| 9 | 95.9 | No |
| 15 | 83.8 | Yes |
| 22 | 90 | No |

Based upon a project selection rule, which chooses projects by decreasing values of DEA cross efficiency scores, until the budget for the program (e.g., 1000) is exhausted, 13 projects are selected as shown in Table 2.13. The last column shows whether a selected project is also selected by the game cross efficiency approach of Liang et al. (2008a), which we will introduce in the next section. The difference can be due to the fact that the game cross efficiency approach is based upon the standard DEA whereas our approach is based upon the log linear frontier.

## 2.7 Game Cross Efficiency

Liang et al. (2008a) develop an approach called Game Cross Efficiency, which is based upon the concept of DEA cross efficiency. As pointed out by Liang et al. (2008a), in many DEA applications, some form of direct or indirect competition may exist among the DMUs under evaluation. Certainly any setting where DMUs

compete for scarce funds, competition is present by definition. R&D project proposals submitted by different departments in an organization can be viewed as DMUs, and subjected to a DEA analysis. These proposals are clearly competing for available funds. Candidates in a preferential election setting can be looked upon as DMUs, and competition is obviously present. An academic applying for research grants is in competition with other academics. Participants in organized sporting events such as the Olympic games, constitute competitive DMUs. When DMUs are viewed as players in a game, cross efficiency scores may be viewed as payoffs, and each DMU may choose to take a non-cooperative game stance to the extent that it will attempt to maximize its (worst possible) payoff.

The idea of game cross efficiency can be presented as follows. For *each* competing $DMU_j$, a multiplier bundle is determined that optimizes the efficiency score for $j$, with the additional constraint that the resulting score for DMU $d$ should be at or above DMU $d$'s estimated best performance, *in a cross-efficiency sense*. In game cross efficiency case, rather than using the ideal score for $DMU_d$, we strive to use a score which will actually be representative of its final measure of performance. The problem, of course, arises that we will not know this best performance score for $d$ until the best performances of all other DMUs are known as well. To combat this "chicken and egg" phenomenon, Liang et al. (2008a) adopt an iterative approach that leads to an equilibrium.

To make these ideas more concrete, suppose that in a game sense, one player $DMU_d$ is given an efficiency score $\alpha_d$, and that another player $DMU_j$ then tries to maximize its own efficiency, subject to the condition that $\alpha_d$ cannot be decreased. The *game cross efficiency* for $DMU_j$ relative to $DMU_d$ is defined as

$$\alpha_{dj} = \frac{\sum_{r=1}^{s} u_{rj}^{d} y_{rj}}{\sum_{i=1}^{m} v_{ij}^{d} x_{ij}}, \quad d = 1, 2, \ldots, n \tag{2.22}$$

where $u_{rj}^{d}$ and $v_{ij}^{d}$ are optimal weights in model (2.23) below. The subscript $dj$ is intended to indicate that $DMU_j$ is permitted only to choose weights that will not deteriorate the current efficiency of $DMU_d$.

The difference between the standard cross efficiency (2.2) and the game cross efficiency (2.22) is that weights in (2.22) are not necessarily an optimal solution in the DEA model (2.4), but rather are a feasible solution to the CRS multiplier model (2.4). Such a definition allows DMUs to choose (negotiate) a set of weights, (hence a form of cross efficiency scores), that are best for all of the DMUs. So, in this sense, Liang et al. (2008a) adopt a non-cooperative game approach.

We use the following model to calculate the game cross efficiency defined in (2.22) for each $DMU_j$ (given that cross efficiency score of $DMU_d$ cannot be less than $\alpha_d$)

$$\text{Max} \quad \sum_{r=1}^{s} u_{rj}^{d} y_{rj}$$

subject to

$$\begin{aligned}
&\sum_{i=1}^{m} v_{ij}^{d} x_{il} - \sum_{r=1}^{s} u_{rj}^{d} y_{rl} \geq 0, l = 1,2,\ldots,n \\
&\sum_{i=1}^{m} v_{ij}^{d} x_{ij} = 1 \\
&\alpha_d \times \sum_{i=1}^{m} v_{ij}^{d} x_{id} - \sum_{r=1}^{s} u_{rj}^{d} y_{rd} \leq 0 \\
&v_{ij}^{d} \geq 0, i = 1,\ldots,m \\
&u_{rj}^{d} \geq 0, r = 1,\ldots,s
\end{aligned} \tag{2.23}$$

where $\alpha_d \leq 1$ is a parameter. This model (2.23) is very similar to the CRS multiplier model, except for the additional constraint of $\alpha_d \times \sum_{i=1}^{m} v_{ij}^{d} x_{id} - \sum_{r=1}^{s} u_{rj}^{d} y_{rd} \leq 0$ which ensures that the (cross) efficiency score of $DMU_d$ cannot be less than $\alpha_d$. Model (2.23) is referred to as the *DEA game cross efficiency model*, based upon $DMU_d$. Note that model (2.23) maximizes the efficiency of $DMU_j$, under the condition that the efficiency of a given $DMU_d$, is not less than a given value ( $\alpha_d$). Thus, the efficiency of $DMU_j$ is further constrained by the requirement that the ratio efficiency of $DMU_d$ is not less than its original average cross efficiency.

The $\alpha_d$ in model (2.23) initially takes the value given by the average original cross efficiency of $DMU_d$ defined in (2.3). When the algorithm converges, this $\alpha_d$ becomes the best (average) game-cross efficiency score.

For each $DMU_j$, model (2.23) is solved $n$ times, once for each $d = 1, \ldots, n$. Note that for each $d$, at optimality, $\sum_{i=1}^{m} v_{ij}^{d} x_{ij} = 1$ holds for $DMU_j$ ($j = 1, 2, \ldots, n$). Therefore, for each $DMU_j$, the optimal value to model (2.23) actually represents a game cross efficiency with respect to $DMU_d$, as defined in (2.22). In other words, for each $DMU_j$, $\alpha_j = \frac{1}{n} \sum_{d=1}^{n} \sum_{r=1}^{s} u_{rj}^{d*}(\alpha_d) y_{rj}$ is called the input-oriented (average) *game cross efficiency* for $DMU_j$, where $u_{rj}^{d*}(\alpha_d)$ represent an optimal solution to model (2.23).

Note that the average game cross efficiency no longer represents a regular DEA cross efficiency value. Liang et al. (2008a) show that optimal game cross efficiency scores constitute a Nash Equilibrium point.

We now present the procedure for determining the *best* average input-oriented game-cross efficiency for $DMU_j$, as described in Liang et al. (2008a).

**Algorithm Step1**: Solve model (2.4) and obtain a set of original DEA cross efficiency scores $\overline{E_d}$ defined in (2.3). Let $t = 1$ and $\alpha_d = \alpha_d^1 = \overline{E_d}$.

**Step2**: Solve model (2.23). Let $\alpha_j^2 = \frac{1}{n} \sum_{d=1}^{n} \sum_{r=1}^{s} u_{rj}^{d*}(\alpha_d^1) y_{rj}$ or in a general format,

$$\alpha_j^{t+1} = \frac{1}{n} \sum_{d=1}^{n} \sum_{r=1}^{s} u_{rj}^{d*}(\alpha_d^t) y_{rj},$$

where $u_{rj}^{d*}(\alpha_d^t)$ represents optimal value of $u_{rj}^d$ in model (2.23) when $\alpha_d = \alpha_d^t$.

**Step3**: If $|\alpha_j^{t+1} - \alpha_j^t| \geq \delta$ for some $j$, where $\delta$ is a user-specified small positive value, then let $\alpha_d = \alpha_d^{t+1}$ and go to Step 2. If $|\alpha_j^{t+1} - \alpha_j^t| < \delta$ for all j, then stop. $\alpha_j^{t+1}$ is the best average game-cross efficiency given to $DMU_j$. (In calculation, we can set $\delta = 0.001$ , for example.)

In Step1, the $\overline{E_d}$ represent traditional (average) cross efficiency scores for $DMU_d, d = 1,2,\ldots,n$, and are the initial values for $\alpha_d$ (denoted as $\alpha_d^1$) in model (2.23). Although the cross efficiency scores may not be unique, Liang et al. (2008a) show that any initial values for $\alpha_d$ (or any traditional cross efficiency scores), will lead to unique game-cross efficiency scores. When the algorithm stops, since $\sum_{r=1}^{s} u_{rj}^{d*}(\alpha_d^t) y_{rj}$ is the optimal value to model (2.23), $\alpha_j^{t+1} = \frac{1}{n}\sum_{d=1}^{n}\sum_{r=1}^{s} u_{rj}^{d*}(\alpha_d^t) y_{rj}, \quad t \geq 1$ is unique. Also, the notation $\alpha_d = \alpha_d^t, \quad t \geq 1$, given in Step 2, means that in model (2.23) $\alpha_d$ is replaced with $\alpha_d^t$. Step 3 is used to indicate when to terminate the process of executing model (2.23).

In a similar manner, we can develop an output-oriented game cross efficiency approach. In this case, we rely on the output-oriented CRS model. First, $\alpha_{dj}$, *game cross efficiency* for $DMU_j$ relative to $DMU_d$, is defined as

$$\alpha_{dj} = \frac{\sum_{i=1}^{m} v_{ij}^d x_{ij}}{\sum_{r=1}^{s} u_{rj}^d y_{rj}}, \; d = 1,2,\ldots,n \tag{2.24}$$

Similar to model (2.23), we have the following output-oriented model when $DMU_j$ is under evaluation

$$\text{Min} \quad \sum_{i=1}^{m} v_{ij}^d x_{ij}$$

subject to

$$\begin{aligned}
&\sum_{i=1}^{m} v_{ij}^d x_{il} - \sum_{r=1}^{s} u_{rj}^d y_{rl} \geq 0, l = 1,2,\ldots,n \\
&\sum_{r=1}^{s} u_{rj}^d y_{rj} = 1 \\
&\sum_{i=1}^{m} v_{ij}^d x_{id} - \alpha_d \times \sum_{r=1}^{s} u_{rj}^d y_{rd} \leq 0 \\
&v_{ij}^d \geq 0, i = 1,\ldots,m \\
&u_{rj}^d \geq 0, r = 1,\ldots,s
\end{aligned} \tag{2.25}$$

where $\alpha_d \geq 1$ is a parameter. This model (2.25) is very similar to the output-oriented CRS multiplier model, except for the additional constraint of $\sum_{i=1}^{m} v_{ij}^d x_{id} - \alpha_d \times \sum_{r=1}^{s} u_{rj}^d y_{rd} \leq 0$ which ensures that the (cross) efficiency score of $DMU_d$ cannot

**Table 2.14** Game cross efficiency

| DMU | Input-oriented game cross efficiency[a] | Output-oriented game cross efficiency | Output-oriented VRS game cross efficiency[a] |
|---|---|---|---|
| 1 | 0.63813 | 1.58809 | 1.45102 |
| 2 | 0.97638 | 1.01304 | 1 |
| 3 | 1 | 1 | 1 |
| 4 | 0.79833 | 1.22833 | 1.24265 |
| 5 | 0.66659 | 1.60533 | 1.46813 |

[a]See the Conclusions section for discussions on the VRS game cross efficiency

be greater than $\alpha_d$. Note that under output-oriented model, a larger score indicates worse performance.

This $\alpha_d$ initially takes the value given by the (average) original output-oriented cross efficiency of $DMU_d$. When the algorithm converges, this $\alpha_d$ becomes the best (average) game-cross efficiency score. Model (2.25) is referred to as the output-oriented *DEA game cross efficiency model*.

For each $DMU_j$, model (2.25) is solved $n$ times, once for each $d = 1, \ldots, n$. Note that for each $d$, at optimality, $\sum_{r=1}^{s} u_{rj}^{d} y_{rj} = 1$ holds for $DMU_j$ $(j = 1, 2, \ldots, n)$. Therefore, for each $DMU_j$, the optimal value to model (2.25) actually represents a game cross efficiency with respect to $DMU_d$, as defined in (2.24). Namely, for each $DMU_j, \alpha_j = \frac{1}{n} \sum_{d=1}^{n} \sum_{i=1}^{m} v_{ij}^{d*}(\alpha_d) x_{ij}$ is called the output-oriented (average) *game cross efficiency* for $DMU_j$, where $v_{ij}^{d*}(\alpha_d)$ represent an optimal solution to model (2.25).

Liang et al. (2008a) provide detailed discussion on the numerical example in Table 2.1 based upon the input-oriented game cross efficiency model (2.23). We here only provide the results for both models (2.23) and (2.25) in Table 2.14. In both cases, the standard DEA cross efficiency scores in Tables 2.3 and 2.5 are used as the initial $\alpha_d$, and we set $\delta = 0.001$. The input-oriented game cross efficiency scores are reached after 10 iterations, and the output-oriented game cross efficiency scores are reached after 27 iterations.

## 2.8 Conclusions

The above discussion is based upon CRS. We can also develop game cross efficiency under the condition of VRS. However, due to the fact that the input-oriented VRS model can yield negative cross efficiency, we here only present the game cross efficiency based upon the output-oriented VRS model where cross efficiency scores are always positive.

The output-oriented VRS game cross efficiency $DMU_j$ relative to $DMU_d$ is given by

$$\alpha_{dj} = \frac{\sum_{i=1}^{m} \omega_{ij}^{d} x_{ij} + v^{d}}{\sum_{r=1}^{s} \mu_{rj}^{d} y_{rj}}, \quad d = 1, 2, \ldots, n \tag{2.26a}$$

**Table 2.15** Output-oriented VRS cross efficiency matrix

| | VRS | Cross efficiency matrix | | | | |
|---|---|---|---|---|---|---|
| DMU | Cross efficiency | DMU1 | DMU2 | DMU3 | DMU4 | DMU5 |
| DMU1 | 1.63167 | 1.41667 | 1.00000 | 1.00000 | 1.16667 | 2.11111 |
| DMU2 | 1.09429 | 1.41667 | 1.00000 | 1.00000 | 1.16667 | 2.11111 |
| DMU3 | 1.00000 | 1.45833 | 1.07143 | 1.00000 | 1.25000 | 2.22222 |
| DMU4 | 2.07000 | 1.41667 | 1.00000 | 1.00000 | 1.16667 | 2.11111 |
| DMU5 | 1.94444 | 2.45000 | 1.40000 | 1.00000 | 5.60000 | 1.16667 |

Based upon the output-oriented VRS multiplier model and model (2.25), we have the following VRS model for obtaining the output-oriented VRS game cross efficiency score.

$$\text{Min} \quad \sum_{i=1}^{m} v_{ij}^{d} x_{ij} + v^{d}$$

subject to

$$\begin{aligned}
&\sum_{i=1}^{m} v_{ij}^{d} x_{il} - \sum_{r=1}^{s} u_{rj}^{d} y_{rl} + v^{d} \geq 0, l = 1,2,\ldots,n \\
&\sum_{r=1}^{s} u_{rj}^{d} y_{rj} = 1 \qquad (2.26b) \\
&\sum_{i=1}^{m} v_{ij}^{d} x_{id} - \alpha_{d} \times \sum_{r=1}^{s} u_{rj}^{d} y_{rd} + v^{d} \leq 0 \\
&v_{ij}^{d} \geq 0, i = 1,\ldots,m \\
&u_{rj}^{d} \geq 0, r = 1,\ldots,s \\
&v^{d} \, free
\end{aligned}$$

where $\sum_{i=1}^{m} v_{ij}^{d} x_{id} - \alpha_{d} \times \sum_{r=1}^{s} u_{rj}^{d} y_{rd} + v^{d} \leq 0$ ensures that the game cross efficiency score of $DMU_d$ cannot be greater than $\alpha_d$. Note that under output-oriented model, a larger score indicates worse performance.

We use the regular output-oriented VRS cross efficiency as the starting point for our game cross efficiency scores. Table 2.15 shows a VRS cross efficiency matrix along with cross efficiency scores shown in column 2.

## References

Anderson TR, Hollingsworth KB, Inman LB (2002) The fixed weighting nature of a cross-evaluation model. J Product Anal 18(1):249–255

Charnes A, Cooper WW, Rhodes E (1978) Measuring efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Seiford L, Stutz J (1982) A multiplicative model for efficiency analysis. Socio-Econ Plan Sci 16(5):223–224

Charnes A, Cooper WW, Seiford LM, Stutz J (1983) Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments. Oper Res Lett 2:101–103

Cook WD, Zhu J (2014) DEA Cobb-Douglas frontier and cross efficiency. J Oper Res Soc 65(2):265–268

Doyle J, Green R (1994) Efficiency and cross efficiency in DEA: derivations, meanings and the uses. J Oper Res Soc 45(5):567–578

Green R, Doyle J, Cook W (1996) Preference voting and project ranking using DEA and cross-evaluation. Eur J Oper Res 90:461–472

Liang L, Wu J, Cook WD, Zhu J (2008a) The DEA game cross efficiency model and its Nash equilibrium. Oper Res 56:278–1288

Liang L, Wu J, Cook WD, Zhu J (2008b) Alternative secondary goals in DEA cross efficiency evaluation. Int J Prod Econ 113:1025–1030

Lim S, Zhu J (in press) DEA Cross-efficiency evaluation under variable returns to scale. J Oper Res Soc

Oral M, Kettani O, Lang P (1991) A methodology for collective evaluation and selection of industrial R & D projects. Manage Sci 37(7):871–883

Sexton TR, Silkman RH, Hogan AJ (1986) Data envelopment analysis: critique and extensions. In: Silkman RH (ed) Measuring efficiency: an assessment of data envelopment analysis, vol 32. Jossey-Bass, San Francisco, pp 73–105

# Chapter 3
# DEA Cross Efficiency Under Variable Returns to Scale

**Sungmook Lim and Joe Zhu**

**Abstract** While cross-efficiency evaluation has been used in a wide variety of DEA applications due to its attractive mechanism, so has it been for DEA models mostly with constant returns to scale (CRS) assumption. This is due to the fact that negative VRS cross-efficiency arises for some DMUs. Since there exist many instances that require the use of the VRS DEA model, it is imperative to develop cross-efficiency measures under VRS. This chapter introduces a recent development of DEA cross-efficiency evaluation approach under VRS, which is motivated by the observation that cross-efficiency evaluation is closely related to the issue of incorporation of weight restrictions and that negative VRS cross-efficiency is related to free production of outputs. The new approach is based upon a geometric interpretation of the relationship between the CRS and VRS DEA models that the VRS model can be cast as a series of CRS models under translated Cartesian coordinate systems. We illustrate this approach using a simple example and show how the VRS negative cross-efficiency problem is addressed under the new framework.

**Keywords** Data envelopment analysis · Cross-efficiency · Variable returns to scale (VRS) · Free production of outputs

## 3.1 Introduction

As discussed in Chap. 2, cross-efficiency evaluation incorporates peer-appraisal in addition to self-appraisal into DEA models to address the criticism of too much flexibility in weighting multiple inputs and outputs. While cross-efficiency evaluation has been used in a wide variety of DEA applications due to its attractive mechanism, so

S. Lim (✉)
Dongguk Business School, Dongguk University—Seoul,
100-715 Seoul, South Korea
e-mail: sungmook@dongguk.edu

J. Zhu
School of Business, Worcester Polytechnic Institute,
01609 Worcester, MA, USA
e-mail: jzhu@wpi.edu

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_3

has it been for DEA models mostly with constant returns to scale (CRS) assumption, such as the CRS model or CCR model of Charnes et al. (1978). The literature, to the extent of the authors' knowledge, has been almost silent on (and has not properly addressed) the issue that cross-efficiency evaluation for input-oriented DEA models with variable returns to scale (VRS) assumption, such as the input-oriented VRS model or the BCC model of Banker et al. (1984), has the problem of negative cross-efficiency for some units. There are many instances where a change in inputs does not result in the same change in outputs that require the use of the VRS DEA model. If DMUs (e.g., bank branches) are of various sizes, then the VRS model is more appropriate to use so that a small sized DMU is not benchmarked against large sized DMUs. The VRS DEA model is one of the basic DEA models that is widely used in various DEA applications. Therefore, it is imperative to develop cross-efficiency measures under the condition of VRS. Wu et al. (2009) propose to add an additional constraint in the VRS model so that cross-efficiencies are non-negative. However, such a modification does not properly address the root cause of the negative VRS cross-efficiency problem.

In the current chapter, we present an approach of DEA cross-efficiency evaluation under VRS developed by Lim and Zhu (2014). Their approach is based upon the observation that cross-efficiency evaluation is closely related to the issue of incorporation of weight restrictions in the sense that each DMU is evaluated by weights chosen by other DMUs in addition to its own. It is well known that the incorporation of weight restrictions in DEA models may result in their infeasibility or non-positive efficiency scores of some units. Podinovski and Bouzdine-Chameeva (2013) finds that these problems arise when weight restrictions induce free production of outputs (i.e., positive outputs with zero inputs) in the underlying technology, which is unacceptable from the production theory point of view. Applying the same concept, Lim and Zhu (2014) find that the problem of negative cross-efficiency in the input-oriented VRS DEA model arises when a DMU is cross-evaluated by a weight vector associated with an efficient frontier which extends to induce free production of outputs in the underlying technology. They claim that such problematic weights are invalid (or unacceptable) to be used for cross-efficiency evaluation and need to be adjusted. To develop a way of resolving the problem of negative cross-efficiency in the input-oriented VRS DEA model, they develop a geometric interpretation of the relationship between the VRS and CRS models. They show that every DMU, via solving the VRS model, seeks for a translation of the Cartesian coordinate system and an optimal bundle of weights such that its CRS-efficiency score, measured under the chosen coordinate system, is maximized. Therefore, VRS cross-efficiency is related to the CRS cross-efficiency measures. Using the fact that any efficient frontier does not extend to induce free production of outputs in the CRS model, they propose that cross-efficiency evaluation for the VRS model should be done via a series of CRS models under translated Cartesian coordinate systems.

## 3.2 Negative Cross Efficiency and Free Production

In this section, we illustrate the problem of negative cross-efficiency and show how it is related to free production of outputs, as discussed in Lim and Zhu (2014). Assume that there are $n$ DMUs which consume $m$ inputs to produce $s$ outputs. DMU $k$ ($k = 1,2,\ldots,n$) uses a vector of inputs $x_k = (x_{1k},\ldots,x_{mk})^T \in R_+^m$ to produce a vector of outputs $y_k = (y_{1k},\ldots,y_{sk})^T \in R_+^s$. The input-oriented VRS model in multiplier form is as follows:

$$\begin{aligned} \max \quad & \sum_{r=1}^{s} u_r y_{r0} - \xi \\ \text{s.t.} \quad & \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} + \xi \geq 0, \quad j = 1,\ldots,n \\ & \sum_{i=1}^{m} v_i x_{i0} = 1 \\ & v_i, \quad u_r \geq \varepsilon \quad \forall i, r, \quad \xi \text{ free in sign.} \end{aligned} \tag{3.1}$$

where $\varepsilon$ is a positive non-Archimedean infinitesimal. When the above model is solved, an input-oriented efficiency score of $\text{DMU}_0$ and input-oriented cross-efficiencies of the other DMUs (evaluated by $\text{DMU}_0$) are obtained together. Specifically, a (conventional input-oriented) *cross-efficiency* of $\text{DMU}_j$ is given by

$$e_{0j}^I = \frac{\sum_{r=1}^{s} u_r^* y_{rj} - \xi^*}{\sum_{i=1}^{m} v_i^* x_{ij}} \tag{3.2}$$

where * denotes an optimal solution to the model. Due to the free variable $\xi$ cross-efficiency calculated by (3.2) may be negative when $\xi > 0$, which results in problematic situation. Averaging $e_{ij}^I$ over $i$, we get a (conventional input-oriented) *cross-efficiency score* of $\text{DMU}_j$. Note that $e_{00}^I$ is an input-oriented efficiency score of $\text{DMU}_0$.

Following is the output-oriented VRS model in multiplier form:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} v_i x_{i0} + \xi \\ \text{s.t.} \quad & \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} + \xi \geq 0, \quad j = 1,\ldots,n \\ & \sum_{r=1}^{s} u_r y_{r0} = 1 \\ & v_i, \quad u_r \geq \varepsilon \quad \forall i, r, \quad \xi \text{ free in sign.} \end{aligned} \tag{3.3}$$

When the above model is solved, an output-oriented efficiency score of $\text{DMU}_0$ and output-oriented cross-efficiencies of the other DMUs (evaluated by $\text{DMU}_0$) are obtained together. Specifically, a (conventional output-oriented) *cross-efficiency* of $\text{DMU}_j$ is given by

$$e_{0j}^{O} = \frac{\sum_{i=1}^{m} v_i^* x_{ij} + \xi^*}{\sum_{r=1}^{s} u_r^* y_{rj}} \tag{3.4}$$

where $v_i^*$, $u_r^*$ and $\xi^*$ are an optimal bundle of weights to model (3.3). Differently from the input-oriented case, the problem of negative cross-efficiency does not occur due to the first set of constraints in model (3.3). Averaging $e_{ij}^{O}$ over $i$, we get a (conventional output-oriented) *cross-efficiency score* of $\text{DMU}_j$. Note that $e_{00}^{O}$ is an output-oriented efficiency score of $\text{DMU}_0$.

The free variable $\xi$ provides an indication of the type of returns to scale (RTS) that prevails at a particular DMU under evaluation. Specifically, increasing returns to scale (IRS) (decreasing returns to scale (DRS)) prevails at $(x_0, y_0)$ if and only if $\xi^* < 0$ ($\xi^* > 0$) for all optimal solutions to the VRS model, whereas CRS prevails at $(x_0, y_0)$ if and only if $\xi^* = 0$ in any optimal solution (see Banker et al. (2011)).

It is worthwhile to note that the VRS model itself involves cross-efficiency evaluation in its constraints (and the same is true with the CRS model). Model (3.1) dictates that each DMU seeks for an optimal bundle of weights while making cross-efficiencies of the other DMUs not exceed unity. Also note that these cross-efficiencies are measured (in a linearized form within the constraints of model (3.1)) no matter which type of RTS prevails at cross-evaluated DMUs. In other words, optimal weights chosen by a DMU exhibiting one type of RTS (say IRS) are used to cross-evaluate the other DMUs exhibiting different types of RTS (say CRS or DRS) within model (3.1). A similar observation can be made in model (3.3). This interpretation provides a justification of the use of cross-efficiency evaluation in DEA as a peer-appraisal approach, particularly under the VRS assumption.

Now let us illustrate how the problem of negative cross-efficiency arises in the input-oriented VRS model using a simple one-input and one-output example. Suppose the data set in Table 3.1 is given, which consists of seven DMUs with a single input and a single output. An input-oriented VRS efficiency score of each DMU along with its optimal weights are provided in Table 3.1. It also reports optimal solutions to the output-oriented model (3.3) as well as RTS classifications.

Figure 3.1 plots the data set and the supporting hyperplane associated with an optimal bundle of weights chosen by DMU F in model (3.1). Hyperplane $H_F$ represents an optimal bundle of weights $(v^*,\ u^*,\ \xi^*) = \left(\frac{1}{4}, \frac{3}{8}, \frac{5}{4}\right)$ for DMU F, with which DMU F attains an efficiency score of unity. Using an input-oriented radial distance measure, (conventional input-oriented) cross-efficiencies of DMUs D, E, and G evaluated by DMU F can be determined with reference to the hyperplane $H_F$. For instance, a cross-efficiency of DMU D is $\overline{D^0D^1}/\overline{D^0D}$, which is $\frac{1}{6}$, a cross-efficiency of DMU E is $\overline{E^0E^1}/\overline{E^0E}$, which is $\frac{5/2}{3} = \frac{5}{6}$, and a cross-efficiency of DMU G is $\overline{\text{G}^0\text{G}^1}/\overline{\text{G}^0\text{G}}$, which is $\frac{11/2}{7} = \frac{11}{14}$. The same results can also be obtained when we use (3.2) for calculating cross-efficiency: $e_{\text{FD}}^{I} = \frac{3/2-5/4}{3/2} = \frac{1}{6}, e_{\text{FE}}^{I} = \frac{15/8-5/4}{3/4} = \frac{5}{6}$,

**Table 3.1** Data and VRS model solutions

| DMU | Input | Output | Input-oriented VRS model (1) | | | | | Output-oriented VRS model (3) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $v^*$ | $u^*$ | $\xi^*$ | Efficiency score | RTS | $v^*$ | $u^*$ | $\xi^*$ | Efficiency score | RTS |
| A | 1 | 1 | 1 | $\frac{1}{4}$ | $-\frac{3}{4}$ | 1 | IRS | 4 | 1 | $-3$ | 1 | IRS |
| B | 4 | 2 | $\frac{1}{4}$ | $\frac{1}{16}$ | $-\frac{3}{16}$ | $\frac{5}{16}$ | IRS | $\frac{1}{2}$ | $\frac{1}{2}$ | 1 | 3 | DRS |
| C | $\frac{3}{2}$ | 3 | $\frac{2}{3}$ | $\frac{1}{3}$ | 0 | 1 | CRS | $\frac{2}{3}$ | $\frac{1}{3}$ | 0 | 1 | CRS |
| D | 6 | 4 | $\frac{1}{6}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{3}{8}$ | DRS | $\frac{1}{12}$ | $\frac{1}{4}$ | $\frac{7}{6}$ | $\frac{5}{3}$ | DRS |
| E | 3 | 5 | $\frac{1}{3}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | 1 | DRS | $\frac{4}{15}$ | $\frac{1}{5}$ | $\frac{1}{5}$ | 1 | DRS |
| F | 4 | 6 | $\frac{1}{4}$ | $\frac{3}{8}$ | $\frac{5}{4}$ | 1 | DRS | $\frac{1}{9}$ | $\frac{1}{6}$ | $\frac{5}{9}$ | 1 | DRS |
| G | 7 | 7 | $\frac{1}{7}$ | $\frac{3}{7}$ | 2 | 1 | DRS | $\frac{1}{21}$ | $\frac{1}{7}$ | $\frac{2}{3}$ | 1 | DRS |

and $e^I_{\mathrm{FG}} = \frac{21/8-5/4}{7/4} = \frac{14}{11}$. While any problem, at least seemingly, doesn't occur in calculating cross-efficiencies of these DMUs, a difficulty will be encountered in determining cross-efficiencies of the other DMUs positioned below the horizontal line (labelled $x$'-axis) that intersects the $y$-axis at $\mathrm{O}' = \left(0, \ \frac{\xi^*}{u^*}\right) = \left(0, \ \frac{10}{3}\right)$. In fact, model (3.1) forces cross-efficiencies of DMUs A, B, and C to be determined with reference to the 'negative-input' segment of hyperplane $H_F$. For instance, a cross-efficiency of DMU A is $\overline{\mathrm{A}^0\mathrm{A}^1}/\overline{\mathrm{A}^0\mathrm{A}}$, which is $\frac{-7/2}{1} = -\frac{7}{2}$, and a cross-efficiency of DMU B is $\overline{\mathrm{B}^0\mathrm{B}^1}/\overline{\mathrm{B}^0\mathrm{B}}$, which is $\frac{-2}{4} = -\frac{1}{2}$. The negative sign is due to the position of $\mathrm{A}^1$ and $\mathrm{B}^1$ (left to the $y$-axis). Note that the same results can be obtained when we use (3.2): $e^I_{\mathrm{FA}} = \frac{3/8-5/4}{1/4} = -\frac{7}{2}$, and $e^I_{\mathrm{FB}} = \frac{3/4-5/4}{1} = -\frac{1}{2}$. The same problem of negative cross-efficiency occurs for DMU C as well.

This problem is caused by situations where weights chosen by some DMUs are invalid for cross-evaluating other DMUs; an optimal bundle of weights chosen by DMU F is not valid for determining cross-efficiencies of DMUs A, B, and C. To justify this, it is worthwhile to note that the efficient frontier associated with the optimal weights chosen by DMU F extends to induce the point $\mathrm{O}'$ which represents a free production of outputs in the underlying technology. Furthermore, model (3.1) forces DMUs A, B, and C to be cross-evaluated with reference to the invalid part of the extended efficient frontier that emanates from the unacceptable free production point $\mathrm{O}'$ and points southwest. This implies that some kind of adjustment is required for those invalid weights to be properly used for cross-efficiency evaluation. We proceed to examine the case of negative values of $\xi^*$.

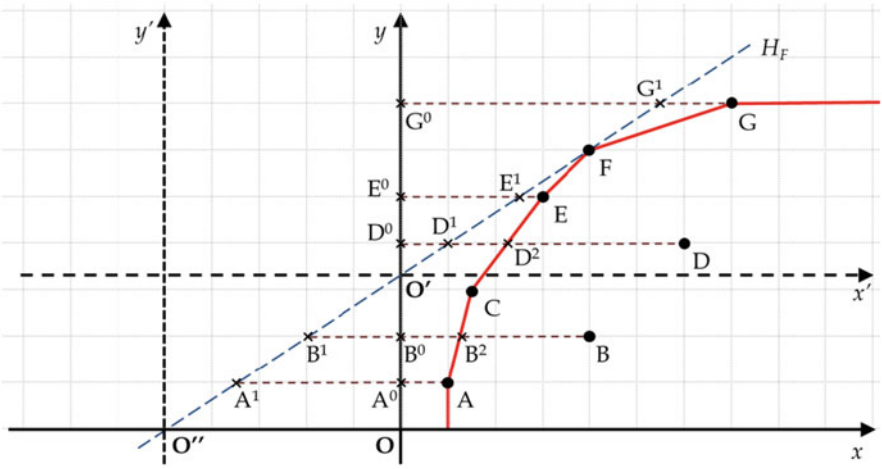



**Fig. 3.1** Input-oriented cross-efficiency evaluation by DMU F

Figure 3.2 shows the supporting hyperplane associated with an optimal bundle of weights chosen by DMU A. Hyperplane $H_A$ represents an optimal bundle of weights $(v^*, u^*, \xi^*) = (1, \frac{1}{4}, -\frac{3}{4})$ for DMU A, with which DMU A attains an efficiency score of unity. Using an input-oriented radial distance measure, (conventional input-oriented) cross-efficiencies of the other DMUs evaluated by DMU A can be determined with reference to hyperplane $H_A$. For instance, a cross-efficiency of DMU D is $\overline{D^0D^1}/\overline{D^0D}$, which is $\frac{7/4}{6} = \frac{7}{24}$, and a cross-efficiency of DMU G is $\overline{G^0G^1}/\overline{G^0G}$, which is $\frac{5/2}{7} = \frac{5}{14}$. Note that the same results can be obtained when we use (3.2) for calculating cross-efficiency; $e^I_{AD} = \frac{1-(-3/4)}{6} = \frac{7}{24}$ and $e^I_{AG} = \frac{7/4-(-3/4)}{7} = \frac{5}{14}$. When the optimal value of the free variable, $\xi^*$, is non-positive, the problem of negative cross-efficiency does not seem to occur. In other words, when a DMU is cross-evaluated by another DMU exhibiting IRS, it always attains a positive cross-efficiency.

Although seemingly unproblematic for this case, the claim made in the above (using the case of DMU F) still applies. According to Podinovski and Bouzdine-Chameeva (2013), a technology is said to allow free production of outputs when it is possible to produce positive outputs with zero inputs. This definition can be extended to include the case of negative outputs with zero inputs such as point $O' = (0, -3)$ in Fig. 3.2. This case may be interpreted as consumption (opposite to production) of outputs without any inputs, and it can be considered as extended free disposability of outputs which may not be unacceptable. However, here it is assumed to be unacceptable, which provides a more general framework. 'Positive outputs with zero inputs' and 'negative outputs with zero inputs' are referred to as 'type I' and 'type II' of free production of outputs, respectively. Note that the efficient frontier associated with the optimal weights chosen by DMU A extends to induce the point O′ (in Fig. 3.2), which represents a type II free production of outputs in the underlying

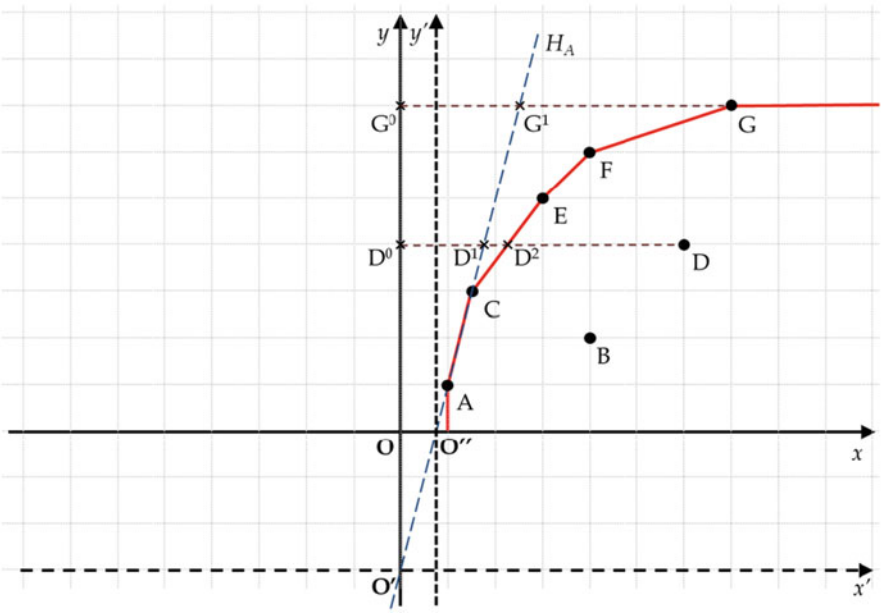



**Fig. 3.2** Input-oriented cross-efficiency evaluation by DMU A

technology. Model (3.1) forces the other DMUs to be cross-evaluated with reference to the invalid efficient frontier (which extends to induce the unacceptable type II free production point O′) for them.

The problem of negative cross-efficiency cannot be actually observed in the output-oriented VRS model because its constraints prevent it. Figure 3.3 plots the same data set and the supporting hyperplane associated with an optimal bundle of weights chosen by DMU F in the output-oriented model (3.3). Hyperplane $H_F$ represents an optimal bundle of weights $(v^*, u^*, \xi^*) = \left(\frac{1}{9}, \frac{1}{6}, \frac{5}{9}\right)$ for DMU F, with which DMU F attains an efficiency score of unity. Using an output-oriented radial distance measure, (conventional output-oriented) cross-efficiencies of the other DMUs evaluated by DMU F can be determined with reference to the hyperplane $H_F$. For instance, a cross-efficiency of DMU A is $\overline{A^0A^1}/\overline{A^0A}$, which is $\frac{4}{1} = 4$, a cross-efficiency of DMU B is $\overline{B^0B^1}/\overline{B^0B}$, which is $\frac{6}{2} = 3$, and a cross-efficiency of DMU G is $\overline{G^0G^1}/\overline{G^0G}$, which is $\frac{8}{7}$. The same results can also be obtained by using the conventional output-oriented cross-efficiency formula (3.4): $e^O_{FA} = \frac{1/9+5/9}{1/6} = 4, e^O_{FB} = \frac{4/9+5/9}{2/6} = 3$, and $e^O_{FG} = \frac{7/9+5/9}{7/6} = \frac{8}{7}$. It can be easily shown that cross-efficiencies of the remaining DMUs are also positive. This is quite obvious due to the geometric structure of the output-oriented radial distance measure used in model (3.3). In other words, the problem of negative cross-efficiency never occurs, as in the case of the optimal value of the free variable $\xi^*$ being non-positive. However, the same weight invalidity problem related to free production

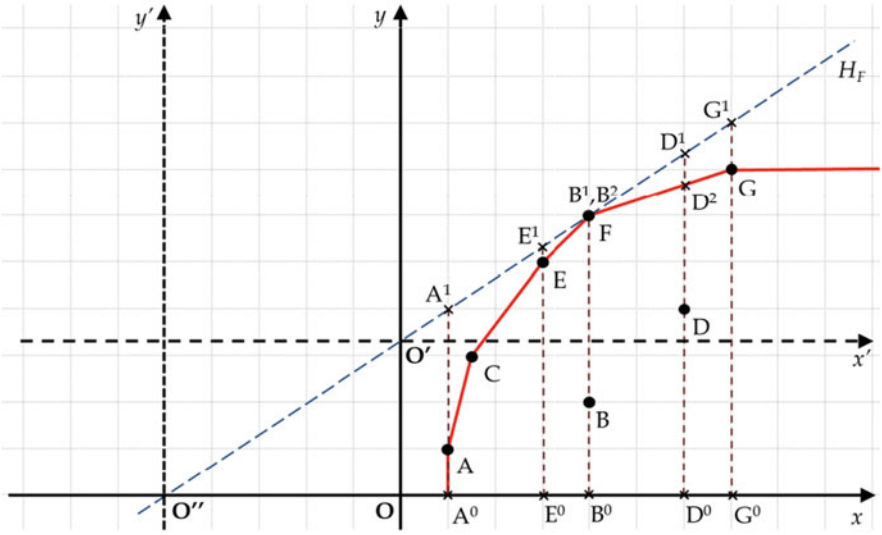



**Fig. 3.3** Output-oriented cross-efficiency evaluation by DMU F

of outputs occurs in cross-efficiency evaluation for the output-oriented VRS DEA model as well, since the input-oriented and output-oriented VRS models have the same efficient frontier structure. Observe that the efficient frontier associated with the optimal weights chosen by DMU F in Fig. 3.3 extends to induce the point $O'$ which represents a (type I) free production of outputs in the underlying technology, and thus the optimal bundle of weights chosen by DMU F is not valid for determining cross-efficiencies of the other DMUs. A similar argument can be made with DMU A in which the associated efficient frontier extends to induce a type II free production of outputs.

The above observations lead to the conclusion that cross-efficiency evaluation via the conventional VRS model is not proper no matter whether the problem of negative cross-efficiency actually arises or not. Therefore, for the VRS DEA model, some significant change of the framework of cross-efficiency evaluation should be developed. Lim and Zhu (2014) accomplish this based on a geometric view of the relationship between the VRS and CRS models, which will be presented in the next section.

## 3.3 DEA and Coordinate Systems: A Geometric Link Between the VRS and CRS Models

To lay a foundation for valid cross-efficiency evaluation in the VRS DEA model, Lim and Zhu (2014) provide a geometric interpretation of the VRS model as a series of CRS models under translated Cartesian coordinate systems, which is formalized by the following theorem.

**Theorem 1** Given any optimal solution $(v^*,\ u^*,\ \xi^*)$ to model (3.1) chosen by a VRS-efficient DMU (denoted $\mathrm{DMU}_0$), a CRS-efficiency score of $\mathrm{DMU}_0$, measured under the translated Cartesian coordinate system defined by an adjusted origin $\mathrm{O}^* = (-\frac{\beta_1\xi^*}{v_1^*}, \ldots, -\frac{\beta_m\xi^*}{v_m^*}, \frac{\beta_{m+1}\xi^*}{u_1^*}, \ldots, \frac{\beta_{m+s}\xi^*}{u_s^*})$, is unity, for any $\beta_k \in R^+$ $(k = 1, \ldots, m+s)$ such that $\sum_{k=1}^{m+s} \beta_k = 1$.

**Proof** Under the translated Cartesian coordinate system with origin $\mathrm{O}^*$, the input-oriented VRS model (3.5) with the same set of DMUs is presented as follows:

$$
\begin{aligned}
\max \quad & \sum_{r=1}^{s} \mu_r \left( y_{r0} - \frac{\beta_{m+r}\xi^*}{u_r^*} \right) - \kappa \\
\text{s.t.} \quad & \sum_{i=1}^{m} \nu_i \left( x_{ij} + \frac{\beta_i\xi^*}{v_i^*} \right) - \sum_{r=1}^{s} \mu_r \left( y_{rj} - \frac{\beta_{m+r}\xi^*}{u_r^*} \right) + \kappa \geq 0, \quad j = 1, \ldots, n \\
& \sum_{i=1}^{m} \nu_i \left( x_{i0} + \frac{\beta_i\xi^*}{v_i^*} \right) = 1 \\
& \nu_i, \quad \mu_r \geq \varepsilon \ \ \forall i, r, \ \kappa \text{ free}
\end{aligned} \tag{3.5}
$$

Consider a solution $(\nu,\ \mu,\ \kappa) = (\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0)$ with $\Gamma = 1 + \xi^* \sum_{k=1}^{m} \beta_k$. Note that that $\Gamma > 0$ since $\xi^* \geq -1 + \sum_{r=1}^{s} u_r^* y_{r0} > -1$ and $\sum_{k=1}^{m} \beta_k \leq 1$. Plugging $(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0)$ in the first set of constraints of model (3.5), we obtain

$$
\begin{aligned}
& \sum_{i=1}^{m} \frac{v_i^*}{\Gamma} \left( x_{ij} + \frac{\beta_i\xi^*}{v_i^*} \right) - \sum_{r=1}^{s} \frac{u_r^*}{\Gamma} \left( y_{rj} - \frac{\beta_{m+r}\xi^*}{u_r^*} \right) \\
& \quad = \frac{1}{\Gamma} \left( \sum_{i=1}^{m} v_i^* x_{ij} - \sum_{r=1}^{s} u_r^* y_{rj} + \xi^* \sum_{k=1}^{m+s} \beta_k \right) \\
& \quad = \frac{1}{\Gamma} \left( \sum_{i=1}^{m} v_i^* x_{ij} - \sum_{r=1}^{s} u_r^* y_{rj} + \xi^* \right) \geq 0, \quad \forall j
\end{aligned}
$$

where the non-negativity is ensured due to the fact that $(v^*,\ u^*,\ \xi^*)$ is a feasible solution to model (3.1) and $\Gamma > 0$. If $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ is plugged in the second set of constraints of model (3.5), we obtain

$$
\sum_{i=1}^{m} \frac{v_i^*}{\Gamma} \left( x_{i0} + \frac{\beta_i\xi^*}{v_i^*} \right) = \frac{1}{\Gamma} \left( \sum_{i=1}^{m} v_i^* x_{i0} + \xi^* \sum_{k=1}^{m} \beta_k \right) = \frac{1}{\Gamma} \left( 1 + \xi^* \sum_{k=1}^{m} \beta_k \right) = 1
$$

using the fact that $\sum_{i=1}^{m} v_i^* x_{i0} = 1$. Therefore, $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ is a feasible solution to model (3.5). Now examine its objective value, which is

$$
\sum_{r=1}^{s} \frac{u_r^*}{\Gamma} \left( y_{r0} - \frac{\beta_{m+r}\xi^*}{u_r^*} \right) = \frac{1}{\Gamma} \left( \sum_{r=1}^{s} u_r^* y_{r0} - \xi^* \sum_{k=m+1}^{m+s} \beta_k \right)
$$

$$= \frac{1}{\Gamma}\left(\sum_{r=1}^{s} u_r^* y_{r0} - \xi^* + \xi^*\left(1 - \sum_{k=m+1}^{m+s} \beta_k\right)\right)$$
$$= \frac{1}{\Gamma}\left(1 + \xi^* \sum_{k=1}^{m} \beta_k\right) = 1.$$

Hence, $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ is an optimal solution to model (3.5) with which $DMU_0$ attains an efficiency score of unity, implying that $DMU_0$ is VRS-efficient at which CRS prevails since $\kappa = 0$ in the optimal solution. This leads to the conclusion that $DMU_0$ is CRS-efficient under the translated coordinate system. □

We also provide a parallel theorem for the output-oriented case as follows:

**Theorem 2** Given any optimal solution $(v^*, u^*, \xi^*)$ to model (3.3) chosen by a VRS-efficient DMU (denoted $DMU_0$), a CRS-efficiency score of $DMU_0$, measured under the translated Cartesian coordinate system defined by an adjusted origin $O^* = (-\frac{\beta_1 \xi^*}{v_1^*}, \ldots, -\frac{\beta_m \xi^*}{v_m^*}, \frac{\beta_{m+1}\xi^*}{u_1^*}, \ldots, \frac{\beta_{m+s}\xi^*}{u_s^*})$, is unity, for any $\beta_k \in R^+ (k = 1, \ldots, m + s)$ such that $\sum_{k=1}^{m+s} \beta_k = 1$.

**Proof** Under the translated Cartesian coordinate system with origin $O^*$, the output-oriented VRS model (3.6) with the same set of DMUs is presented as follows:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{m} v_i \left(x_{i0} + \frac{\beta_i \xi^*}{v_i^*}\right) + \kappa \\
\text{s.t.} \quad & \sum_{i=1}^{m} v_i \left(x_{ij} + \frac{\beta_i \xi^*}{v_i^*}\right) - \sum_{r=1}^{s} \mu_r \left(y_{rj} - \frac{\beta_{m+r}\xi^*}{u_r^*}\right) + \kappa \geq 0, \quad j = 1, \ldots, n \\
& \sum_{r=1}^{s} \mu_r y_{r0} = 1 \qquad (3.6) \\
& v_i, \quad \mu_r \geq \varepsilon \ \ \forall i, r, \quad \kappa \text{ free}
\end{aligned}$$

Consider a solution $(v, \mu, \kappa) = \left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ with $\Gamma = 1 - \xi^* \sum_{k=m+1}^{m+s} \beta_k$. Note that $\Gamma > 0$ since $\xi^* = 1 - \sum_{i=1}^{m} v_i^* x_{i0} < 1$ and $\sum_{k=m+1}^{m+s} \beta_k \leq 1$. Plugging $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ in the first set of constraints of model (3.6), we obtain

$$\sum_{i=1}^{m} \frac{v_i^*}{\Gamma}\left(x_{ij} + \frac{\beta_i \xi^*}{v_i^*}\right) - \sum_{r=1}^{s} \frac{u_r^*}{\Gamma}\left(y_{rj} - \frac{\beta_{m+r}\xi^*}{u_r^*}\right)$$
$$= \frac{1}{\Gamma}\left(\sum_{i=1}^{m} v_i^* x_{ij} - \sum_{r=1}^{s} u_r^* y_{rj} + \xi^* \sum_{k=1}^{m+s} \beta_k\right)$$
$$= \frac{1}{\Gamma}\left(\sum_{i=1}^{m} v_i^* x_{ij} - \sum_{r=1}^{s} u_r^* y_{rj} + \xi^*\right) \geq 0, \quad \forall j$$

where the non-negativity is ensured due to the fact that $(v^*,\ u^*,\ \xi^*)$ is a feasible solution to model (3.3) and. $\Gamma > 0$. If $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma},\ 0\right)$ is plugged in the second set of constraints of model (3.6), we obtain

$$\sum_{r=1}^{s} \frac{u_r^*}{\Gamma}\left(y_{i0} - \frac{\beta_{m+r}\xi^*}{u_r^*}\right) = \frac{1}{\Gamma}\left(\sum_{r=1}^{s} u_r^* y_{r0} - \xi^* \sum_{k=m+1}^{m+s} \beta_k\right)$$
$$= \frac{1}{\Gamma}\left(1 - \xi^* \sum_{k=m+1}^{m+s} \beta_k\right) = 1$$

using the fact that $\sum_{r=1}^{s} u_r^* y_{r0} = 1$. Therefore, $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}, 0\right)$ is a feasible solution to model (3.6). Now examine its objective value, which is

$$\sum_{i=1}^{m} \frac{v_i^*}{\Gamma}\left(x_{i0} + \frac{\beta_i \xi^*}{v_i^*}\right) = \frac{1}{\Gamma}\left(\sum_{i=1}^{m} v_i^* x_{i0} + \xi^* \sum_{k=1}^{m} \beta_k\right)$$
$$= \frac{1}{\Gamma}\left(\sum_{i=1}^{m} v_i^* x_{i0} + \xi^* - \xi^*\left(1 - \sum_{k=1}^{m} \beta_k\right)\right)$$
$$= \frac{1}{\Gamma}\left(1 - \xi^* \sum_{k=m+1}^{m+s} \beta_k\right) = 1.$$

Hence, $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma},\ 0\right)$ is an optimal solution to model (3.6) with which $\mathrm{DMU}_0$ attains an efficiency score of unity, implying that $\mathrm{DMU}_0$ is VRS-efficient at which CRS prevails since $\kappa = 0$ in the optimal solution. This leads to the conclusion that $\mathrm{DMU}_0$ is CRS-efficient under the translated coordinate system. □

Theorem 1 (Theorem 2) indicates that each DMU, via solving the VRS model, seeks for an optimal bundle of weights and free variable value with which its input-oriented (output-oriented) CRS-efficiency score, measured under a translated Cartesian coordinate system, is maximized (minimized). In addition, the theorems show that the location of the adjusted origin of the translated coordinate system is associated with the chosen optimal bundle of weights and free variable value. It should be noted here that the efficient frontier determined by the CRS model does not extend to induce free production of outputs, and thus the CRS model does not suffer from the problem of negative cross-efficiency. Therefore, we expect that the problem of negative cross-efficiency can be effectively resolved by transforming the VRS model into a series of CRS models.

Although Theorem 1 and Theorem 2 deal with only VRS-efficient DMUs, they can be applied implicitly to VRS-inefficient ones as well since inefficient DMUs can choose the same weights with their reference points (projections) on the efficient frontier to maximize (or minimize in the output-oriented case) their CRS efficiency scores. However, it should be pointed out that an input-oriented VRS efficiency score of an inefficient DMU under the original coordinate system may differ from its input-oriented CRS efficiency score under a translated coordinate system chosen by

Theorem 1. On the other hand, an output-oriented VRS efficiency score of an inefficient DMU under the original coordinate system coincides with its output-oriented CRS efficiency score under a translated coordinate system chosen by Theorem 2.

On the basis of Theorem 1 and Theorem 2, the following corollaries, which are slight modifications of the original versions in Lim and Zhu (2014), can be established to provide a link between VRS optimal weights and free production. Their proofs can be found in Lim and Zhu (2014).

**Corollary 1** The supporting hyperplane of the efficient frontier associated with an optimal bundle of weights in model (3.1) or model (3.3) chosen by a VRS-efficient DMU exhibiting DRS extends to induce type I free production of outputs in the underlying technology.

**Corollary 2** The supporting hyperplane of the efficient frontier associated with an optimal bundle of weights in model (3.1) or model (3.3) chosen by a VRS-efficient DMU exhibiting IRS extends to induce type II free production of outputs in the underlying technology.

If a VRS-efficient DMU exhibits CRS (i.e., it is CRS-efficient), the corresponding efficient frontier may or may not extend to induce free production of outputs, depending on the sign of the chosen $\xi^*$. However, an optimal solution to model (3.1) where $\xi^* = 0$ always exists for a CRS-efficient DMU, and it is assumed that such solution is always chosen. (In applications, once the CRS condition is identified with a DMU, we can use the CRS model to calculate cross efficiency scores.)

Corollaries 1 and 2 suggest that optimal weights chosen by DMUs at which either DRS or IRS prevails are not valid for cross-evaluating other DMUs, and this provides a good rationale for the development of the current cross-efficiency evaluation approach in the VRS DEA model, which will be presented in the next section. Before we proceed we illustrate Theorem 1, Theorem 2, and the related corollaries using the example introduced in the previous section.

For an illustration for the input-oriented case, we use DMU F which is VRS-efficient and exhibits DRS under the original coordinate system with origin O, depicted in Fig. 3.1. Its optimal bundle of weighs chosen to model (3.1) is $(v^*,\ u^*,\ \xi^*) = (\frac{1}{4}, \frac{3}{8}, \frac{5}{4})$. The supporting hyperplane $H_F$ intersects the $y$-axis at $\mathrm{O}' = (0, \frac{\xi^*}{u^*}) = (0, \frac{10}{3})$ and the $x$-axis at $\mathrm{O}'' = (-\frac{\xi^*}{v^*}, 0) = (-5,\ 0)$. Any point on the line $\overline{\mathrm{O}'\mathrm{O}''}$ can be represented by a convex combination of the two extreme points; $\mathrm{O}^* = (-\frac{\beta_1\xi^*}{v^*},\ \frac{\beta_2\xi^*}{u^*}) = (-5\beta_1,\ \ \frac{10}{3}\beta_2)$ where $\beta_1 + \beta_2 = 1$ and $\beta_1,\ \ \beta_2 \in R^+$. Observe that the supporting hyperplane $H_F$ of the efficient frontier associated with the optimal weights chosen by DMU F extends to induce the point $\mathrm{O}'$ (in Fig. 3.1) which represents a type I free production of outputs, as indicated by Corollary 1. Therefore, the optimal weights to model (3.1) chosen by DMU F at which DRS prevails are not valid for cross-evaluating the other DMUs; for some DMUs such as A, B, and C, this invalidity results in the actually observable realization of negative cross-efficiency.

On the other hand, an input-oriented VRS efficiency score of DMU F, measured under the translated Cartesian coordinate system with an adjusted origin $\mathrm{O}^*$, is unity

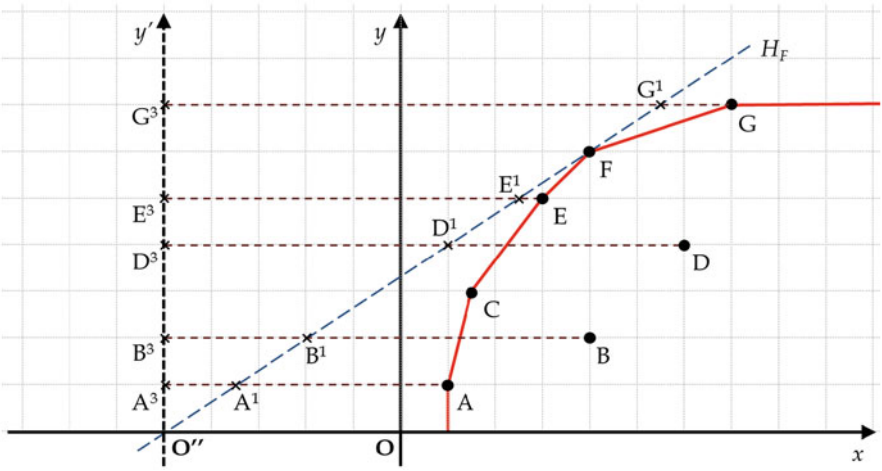



**Fig. 3.4** Input-oriented cross-efficiency evaluation by DMU F under a translated coordinate system

and CRS prevails at DMU F, meaning that DMU F is CRS-efficient under the translated coordinate system, as indicated by Theorem 1. Considering a special case where $\beta_1 = 1$ and $\beta_2 = 0$, the Cartesian coordinate system defined by the $x$-axis and the $y$'-axis with the adjusted origin $O^* = O'' = (-5, 0)$ renders DMU F CRS-efficient. Note that, under the translated coordinate system, the optimal weights $(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}) = (\frac{v^*}{1+\xi^*\beta_1}, \frac{u^*}{1+\xi^*\beta_1}) = (\frac{1/4}{9/4}, \frac{3/8}{9/4}) = (\frac{1}{9}, \frac{1}{6})$ chosen by DMU F, represented again by $H_F$, are valid for cross-evaluating the other DMUs since they do not induce free production of outputs. For instance, under the translated coordinate system, the coordinates of DMU A are (6,1), and its input-oriented cross-efficiency using the weights $(\frac{1}{9}, \frac{1}{6})$ is $e^I_{FA} = \frac{1/6}{6/9} = \frac{1}{4}$. This score can also be determined geometrically under the translated coordinate system in Fig. 3.4: $e^I_{FA} = \overline{A^3A^1}/\overline{A^3A} = \frac{3/2}{6} = \frac{1}{4}$. The coordinates of DMU B under the translated coordinate system are (9, 2), and its input-oriented cross-efficiency using the weights $(\frac{1}{9}, \frac{1}{6})$ is $e^I_{FB} = \frac{2/6}{9/9} = \frac{1}{3}$. This score can also be determined geometrically under the translated coordinate system in Fig. 3.4: $e^I_{FB} = \overline{B^3B^1}/\overline{B^3B} = \frac{3}{9} = \frac{1}{3}$.

For another illustration for the input-oriented case, we use DMU A which is VRS-efficient and exhibits IRS under the original coordinate system with origin O, depicted in Fig. 3.2. Its optimal bundle of weighs chosen to model (3.1) is $(v^*, u^*, \xi^*) = (1, \frac{1}{4}, -\frac{3}{4})$. The supporting hyperplane $H_A$ intersects the $y$-axis at $O' = (0, \frac{\xi^*}{u^*}) = (0, -3)$ and the $x$-axis at $O'' = (-\frac{\xi^*}{v^*}, 0) = (\frac{3}{4}, 0)$. Any point on the line $\overline{O'O''}$ can be represented by a convex combination of the two extreme points; $O^* = (-\frac{\beta_1\xi^*}{v^*}, \frac{\beta_2\xi^*}{u^*}) = (\frac{3}{4}\beta_1, -3\beta_2)$ where $\beta_1 + \beta_2 = 1$ and $\beta_1, \beta_2 \in R^+$. Observe that the supporting hyperplane $H_A$ of the efficient frontier associated with the optimal weights chosen by DMU A extends to induce the point $O'$ (in Fig. 3.2) which represents a type II free production of outputs, as indicated by Corollary 2. Therefore,

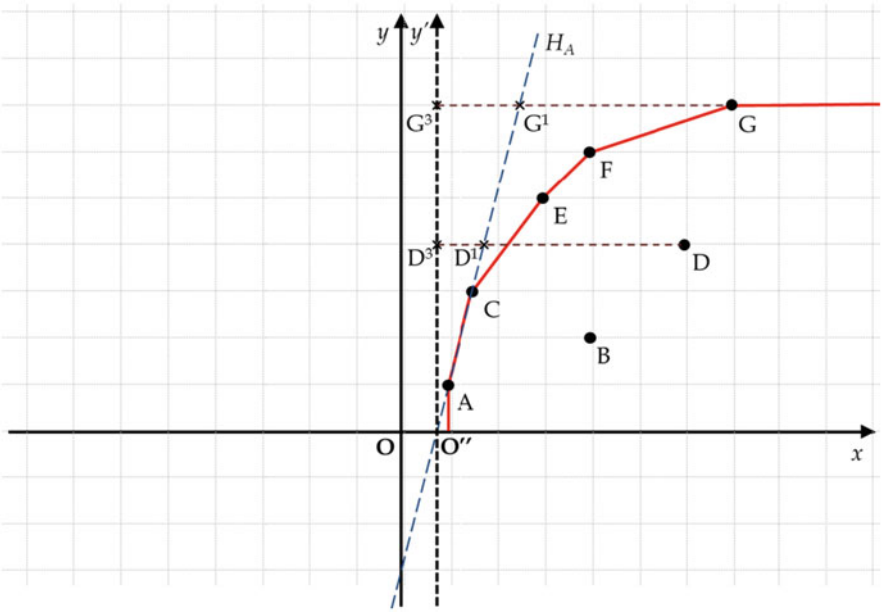



**Fig. 3.5** Input-oriented cross-efficiency evaluation by DMU A under a translated coordinate system

the optimal weights to model (3.1) chosen by DMU A at which IRS prevails are not valid for cross-evaluating the other DMUs, although this invalidity does not result in actually observable realization of negative cross-efficiency.

On the other hand, an input-oriented VRS efficiency score of DMU A, measured under the translated Cartesian coordinate system with an adjusted origin $\mathrm{O}^*$, is unity and CRS prevails at DMU A, meaning that DMU A is CRS-efficient under the translated coordinate system, as indicated by Theorem 1. Considering a special case where $\beta_1 = 1$ and $\beta_2 = 0$, the Cartesian coordinate system defined by the $x$-axis and the $y$'-axis with the adjusted origin $\mathrm{O}^* = \mathrm{O}'' = (\frac{3}{4}, 0)$ renders DMU A CRS-efficient. Note that, under the translated coordinate system, the optimal weights $(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}) = (\frac{v^*}{1+\xi^*\beta_1}, \frac{u^*}{1+\xi^*\beta_1}) = (\frac{1}{1/4}, \frac{1/4}{1/4}) = (4, 1)$ chosen by DMU A, represented again by $H_A$, are valid for cross-evaluating the other DMUs since they do not induce free production of outputs. For instance, under the translated coordinate system, the coordinates of DMU D are $(\frac{21}{4}, 4)$, and its input-oriented cross-efficiency using the weights (4,1) is $e^I_{\mathrm{AD}} = \frac{4}{21}$. This score can also be determined geometrically under the translated coordinate system in Fig. 3.5: $e^I_{\mathrm{AD}} = \overline{\mathrm{D}^3\mathrm{D}^1}/\overline{\mathrm{D}^3\mathrm{D}} = \frac{1}{21/4} = \frac{4}{21}$. The coordinates of DMU G under the translated coordinate system are $(\frac{25}{4}, 7)$,and its input-oriented cross-efficiency using the weights $(4, 1)$ is $e^I_{\mathrm{AG}} = \frac{7}{25}$. This score can also be determined geometrically under the translated coordinate system in Fig. 3.5: $e^I_{\mathrm{AG}} = \overline{\mathrm{G}^3\mathrm{G}^1}/\overline{\mathrm{G}^3\mathrm{G}} = \frac{7/4}{25/4} = \frac{7}{25}$.

Turning to the output-oriented case, we revisit Fig. 3.3 for an illustration. Observe that DMU F is VRS-efficient and exhibits DRS under the original coordinate system with origin O, as depicted in Fig. 3.3. Its optimal bundle of weighs chosen to model (3.3) is $(v^*,\ u^*,\ \xi^*) = (\frac{1}{9}, \frac{1}{6}, \frac{5}{9})$. The supporting hyperplane $H_F$ intersects the $y$-axis at $\mathrm{O}' = (0,\ \frac{\xi^*}{u^*}) = (0,\ \frac{10}{3})$ and the $x$-axis at $\mathrm{O}'' = (-\frac{\xi^*}{v^*},\ 0) = (-5, 0)$. Any point on the line $\overline{\mathrm{O}'\mathrm{O}''}$ can be represented by a convex combination of the two extreme points; $\mathrm{O}^* = (-\frac{\beta_1\xi^*}{v^*},\ \frac{\beta_2\xi^*}{u^*}) = (-5\beta_1,\ \frac{10}{3}\beta_2)$ where $\beta_1 + \beta_2 = 1$ and $\beta_1, \beta_2 \in R^+$. Observe that the supporting hyperplane $H_F$ of the efficient frontier associated with the optimal weights chosen by DMU F extends to induce the point $\mathrm{O}'$ (in Fig. 3.3) which represents a type I free production of outputs, as indicated by Corollary 1. Therefore, the optimal weights to model (3.3) chosen by DMU F at which DRS prevails are not valid for cross-evaluating the other DMUs, although this invalidity does not result in actually observable realization of negative cross-efficiency.

On the other hand, an output-oriented VRS efficiency score of DMU F, measured under the translated Cartesian coordinate system with an adjusted origin $\mathrm{O}^*$, is unity and CRS prevails at DMU F, meaning that DMU F is CRS-efficient under the translated coordinate system, as indicated by Theorem 2. Considering a special case where $\beta_1 = 1$ and $\beta_2 = 0$, the Cartesian coordinate system defined by the $x$-axis and the $y$'-axis with the adjusted origin $\mathrm{O}^* = \mathrm{O}'' = (-5, 0)$ renders DMU F CRS-efficient. Note that, under the translated coordinate system, the optimal weights $(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}) = (\frac{v^*}{1-\xi^*\beta_2}, \frac{u^*}{1-\xi^*\beta_2}) = (\frac{1/9}{1}, \frac{1/6}{1}) = (\frac{1}{9}, \frac{1}{6})$ chosen by DMU F, represented again by $H_F$, are valid for cross-evaluating the other DMUs since they do not induce free production of outputs. For instance, under the translated coordinate system, the coordinates of DMU A are (6,1), and its output-oriented cross-efficiency using the weights $(\frac{1}{9}, \frac{1}{6})$ is $e^O_{\mathrm{FA}} = \frac{6/9}{1/6} = 4$. This score can also be geometrically confirmed under the translated coordinate system (defined by the $x$-axis and the $y$'-axis) in Fig. 3.3. Actually, the output-oriented radial distance measurement is not altered by input-oriented translations of the coordinate system, as is well known. The coordinates of DMU G under the translated coordinate system are, (12, 7), and its output-oriented cross-efficiency using the weights $(\frac{1}{9}, \frac{1}{6})$ is $e^O_{\mathrm{FG}} = \frac{12/9}{7/6} = \frac{8}{7}$. This score can also be confirmed geometrically under the translated coordinate system in Fig. 3.3.

## 3.4 Cross Efficiency in the VRS Model

Recall that Theorem 1, Theorem 2, and the corollaries show that optimal weights chosen by a VRS-efficient DMU exhibiting IRS or DRS are not valid for cross-evaluating other DMUs because the corresponding supporting hyperplane of the efficient frontier extends to induce type I or type II free production of outputs. They also imply a geometric relationship between the VRS and CRS models which can be stated as "*the VRS model for any DMU can be casted as the CRS model for the same DMU under a translated Cartesian coordinate system*." Using the fact that any supporting hyperplane of the efficient frontier does not extend to induce free

production of outputs in the CRS model and thus is always valid for cross-efficiency evaluation, Lim and Zhu (2014) propose that cross-efficiency evaluation for the VRS model should be done via a series of CRS models under translated Cartesian coordinate systems.

Let us first examine the input-oriented case. For an intuitive exposition, DMU F in Fig. 3.1 is examined. Solving the input-oriented VRS model for DMU F is equivalent to solving the input-oriented CRS model for the same unit under a translated Cartesian coordinate system. While there exist numerous choices of an adjusted origin according to Theorem 1 (*i.e.*, any point along $\overline{O'O''}$ will do), a convenient choice will be either a point on the $x$-axis or one on the $y$-axis. However, if the point on the $y$-axis, O′, is chosen for an adjusted origin, some units (DMUs A, B, and C) will have negative outputs under the translated coordinate system, which is not acceptable. This is not the case with the point on the $x$-axis, O″, and therefore it is selected for an adjusted origin. It should be noted that O″ is not the only choice for $O^*$ along $\overline{O'O''}$ which does not give rise to the negative-output problem, and resulting cross-efficiencies depend on the choice of $O^*$. However, the choice of an adjusted origin on the $x$-axis makes it possible to derive a general formula (that does not depend on coefficients $\beta_k$) for VRS cross-efficiency as follows in the subsequent paragraphs. With such translation of the coordinate system applied, it becomes valid for DMU F to cross-evaluate the other DMUs with reference to its supporting hyperplane $H_F$.

A similar reasoning can be applied to DMU A in Fig. 3.2. Solving the input-oriented VRS model for DMU A is equivalent to solving the input-oriented CRS model for the same unit under a translated Cartesian coordinate system. Again there exist numerous choices of an adjusted origin along $\overline{O'O''}$ according to Theorem 1. While a convenient choice will be either a point on the $x$-axis or one on the $y$-axis, the point on the $x$-axis, O″, can be selected to ensure consistency. With such translation of the coordinate system applied, it becomes valid for DMU A to cross-evaluate the other DMUs with reference to its supporting hyperplane $H_A$.

Now a general formula for cross-efficiency evaluation in the input-oriented VRS model is developed. Suppose that $DMU_0$ cross-evaluates $DMU_j$ using its optimal solution $(v^*, u^*, \xi^*)$ to model (3.1). The translation of the coordinate system is considered defined by an adjusted origin $O^* = (-\frac{\beta_1\xi^*}{v_1^*}, \ldots, -\frac{\beta_m\xi^*}{v_m^*}, 0, \ldots, 0)$ where zero repeats $s$ times for the output-associated coordinates, $\sum_{k=1}^{m}\beta_k = 1$ and $\beta_k \in R^+$. Under this translated coordinate system, an input-oriented CRS cross-efficiency $e_{0j}^I$ of $DMU_j$ is determined using the optimal weights $(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma})$ chosen by $DMU_0$, where $\Gamma = 1 + \xi^*\sum_{k=1}^{m}\beta_k$ as defined in Theorem 1, as follows:

$$e_{0j}^{I*} = \frac{\sum_{r=1}^{s}\frac{u_r^*}{\Gamma}y_{rj}}{\sum_{i=1}^{m}\frac{v_i^*}{\Gamma}\left(x_{ij} + \frac{\beta_i\xi^*}{v_i^*}\right)} = \frac{\sum_{r=1}^{s}u_r^*y_{rj}}{\sum_{i=1}^{m}v_i^*\left(x_{ij} + \frac{\beta_i\xi^*}{v_i^*}\right)} = \frac{\sum_{r=1}^{s}u_r^*y_{rj}}{\sum_{i=1}^{m}v_i^*x_{ij} + \xi^*} \quad (3.7)$$

This formula can be used for an input-oriented VRS cross-efficiency of $DMU_j$ (evaluated by $DMU_0$) under the original coordinate system. Note that the final formula does not involve coefficients $\beta_r$ and it has the same form with the inverse of the conventional cross-efficiency formula used in the output-oriented VRS model. According to

the following proposition, input-oriented VRS cross-efficiencies calculated by (3.7) are positive and less than or equal to unity.

**Proposition 1**

$$0 < e_{0j}^{I*} \leq 1$$

**Proof** Optimal weights used in (3.7) are obtained by solving model (3.1). The first set of constraints in model (3.1) can be rewritten as follows:

$$\sum_{i=1}^{m} v_i x_{ij} + \xi \geq \sum_{r=1}^{s} u_r y_{rj} > 0, \quad j = 1, \ldots, n$$

Therefore, it follows that $0 < e_{0j}^{I*} \leq 1$ for all $j$. □

Let us now turn to the output-oriented case. As pointed out in Sect. 3.2, the same weight invalidity problem related to free production of outputs occurs in cross-efficiency evaluation for the output-oriented VRS model as well, although the problem of negative cross-efficiency cannot be actually observed in the output-oriented VRS model due to its constraints. Therefore, it is required to develop this case. Suppose that $\mathrm{DMU}_0$ cross-evaluates $\mathrm{DMU}_j$ using its optimal solution $(v^*, u^*, \xi^*)$ to model (3.3). The translation of the coordinate system is considered defined by $\mathrm{O}^*$ (the same choice with the input-oriented case). Under this translated coordinate system, an output-oriented CRS cross-efficiency $e_{0j}^{O}$ of $\mathrm{DMU}_j$ is determined using the optimal weights $\left(\frac{v^*}{\Gamma}, \frac{u^*}{\Gamma}\right)$ chosen by $\mathrm{DMU}_0$, where $\Gamma = 1 - \xi^* \sum_{k=m+1}^{m+s} \beta_k$ as defined in Theorem 2, as follows:

$$e_{0j}^{O*} = \frac{\sum_{i=1}^{m} \frac{v_i^*}{\Gamma} \left( x_{ij} + \frac{\beta_i \xi^*}{v_i^*} \right)}{\sum_{r=1}^{s} \frac{u_r^*}{\Gamma} y_{rj}} = \frac{\sum_{i=1}^{m} v_i^* x_{ij} + \xi^* \sum_{i=1}^{m} \beta_i}{\sum_{r=1}^{s} u_r^* y_{rj}} = \frac{\sum_{i=1}^{m} v_i^* x_{ij} + \xi^*}{\sum_{r=1}^{s} u_r^* y_{rj}} \tag{3.8}$$

This formula can be used for an output-oriented VRS cross-efficiency of $\mathrm{DMU}_j$ (evaluated by $\mathrm{DMU}_0$) under the original coordinate system. Note that the final formula does not involve coefficients $\beta_r$ and it coincides with the conventional output-oriented cross-efficiency formula. This coincidence is obvious due to the well-known fact that the output-oriented VRS model is invariant with respect to the input-oriented translation of the coordinate system. Output-oriented VRS cross-efficiencies calculated by (3.8) are greater than or equal to one, as shown in the following proposition.

**Proposition 2**

$$e_{0j}^{O*} \geq 1$$

**Proof** Optimal weights used in (3.8) are obtained by solving model (3.3). The first set of constraints in model (3.3) can be rewritten as follows:

$$\sum_{i=1}^{m} v_i x_{ij} + \xi \geq \sum_{r=1}^{s} u_r y_{rj} > 0, \quad j = 1, \ldots, n$$

Therefore, it follows that $e_{0j}^{O*} \geq 1$ for all $j$. □

As pointed out in Sect. 3.3, an input-oriented VRS cross-efficiency of a DMU evaluated by the DMU itself, calculated using (3.7), differs from its (simple) input-oriented VRS efficiency score calculated using (3.2). This means that one DMU will be given $n + 1$ scores; $n$ cross-efficiencies and one (simple) efficiency score. Lim and Zhu (2014) suggests to average these $n$ cross-efficiencies to calculate an input-oriented VRS cross-efficiency score of the DMU. An alternative is to average all $n + 1$ scores. In case of the output-oriented VRS model, one DMU is given $n$ scores consisting of $n - 1$ cross-efficiencies ($e_{0j}^{O*}$, $j \neq 0$) and one (simple) efficiency score ($e_{00}^{O*}$), as in the conventional approach. All $n$ scores can be averaged to calculate an output-oriented VRS cross-efficiency score of the DMU. An alternative is not to include the simple efficiency score as part of the average, which is suggested in Doyle and Green (1994).

Tables 3.2 and 3.3 show an input-oriented and an output-oriented VRS cross-efficiency matrix, respectively, for the example data set (seven DMUs) given in Table 3.1. In case of the input-oriented VRS model, each DMU is given eight scores; seven input-oriented VRS cross-efficiencies and one input-oriented VRS (simple) efficiency score. An input-oriented VRS cross-efficiency score of each DMU is calculated by averaging its eight scores, which is given in the last row of Table 3.2. In case of the output-oriented VRS model, each DMU is given seven scores; six output-oriented VRS cross-efficiencies (excluding cross-efficiency rated by itself) and one output-oriented VRS (simple) efficiency score. Note that output-oriented VRS (simple) efficiency scores can be found on the diagonal of the cross-efficiency matrix. An output-oriented VRS cross-efficiency score of each DMU is calculated by averaging its seven scores, which is given in the last row of Table 3.3.

Considering that cross-efficiency evaluation in DEA is a method of peer-evaluation, $e_{0j}^{I*}$ (or $e_{0j}^{O*}$) implements peer-evaluation for the VRS DEA model more fully. Note that the VRS DEA model allows a DMU to choose an optimal value $\xi^*$ in addition to ($v^*$, $u^*$), where $\xi^*$ indicates the RTS type of the DMU. By Theorem 1 (and Theorem 2), we see that an optimal choice ($v^*$, $u^*$, $\xi^*$) of a DMU determines the normal vector ($v^*$, $u^*$) of the supporting hyperplane associated with the efficient frontier (onto which the DMU is projected), as well as the origin of a new Cartesian coordinate system under which the DMU exhibits CRS (*i.e.*, the most productive scale size). The new origin can be determined based on the value of $\xi^*$. Note that the CRS DEA model determines only ($v^*, u^*$) fixing $\xi^* = 0$. Therefore, cross-efficiency evaluation in the VRS DEA model should properly take into

**Table 3.2** Input-oriented VRS cross-efficiency matrix

| | Rated DMU | | | | | | |
|---|---|---|---|---|---|---|---|
| Rating DMU | A | B | C | D | E | F | G |
| A | 1 | $\frac{2}{13}$ | 1 | $\frac{4}{21}$ | $\frac{5}{9}$ | $\frac{6}{13}$ | $\frac{7}{25}$ |
| B | 1 | $\frac{2}{13}$ | 1 | $\frac{4}{21}$ | $\frac{5}{9}$ | $\frac{6}{13}$ | $\frac{7}{25}$ |
| C | $\frac{1}{2}$ | $\frac{1}{4}$ | 1 | $\frac{1}{3}$ | $\frac{5}{6}$ | $\frac{3}{4}$ | $\frac{1}{2}$ |
| D | $\frac{3}{7}$ | $\frac{6}{19}$ | 1 | $\frac{4}{9}$ | 1 | $\frac{18}{19}$ | $\frac{21}{31}$ |
| E | $\frac{3}{7}$ | $\frac{6}{19}$ | 1 | $\frac{4}{9}$ | 1 | $\frac{18}{19}$ | $\frac{21}{31}$ |
| F | $\frac{1}{4}$ | $\frac{1}{3}$ | $\frac{9}{13}$ | $\frac{6}{11}$ | $\frac{15}{16}$ | 1 | $\frac{7}{8}$ |
| G | $\frac{1}{5}$ | $\frac{1}{3}$ | $\frac{18}{31}$ | $\frac{3}{5}$ | $\frac{15}{17}$ | 1 | 1 |
| Input-oriented VRS (simple) efficiency score | 1 | $\frac{5}{16}$ | 1 | $\frac{3}{8}$ | 1 | 1 | 1 |
| Input-oriented VRS cross-efficiency score | 0.6009 | 0.2711 | 0.9091 | 0.3905 | 0.8455 | 0.8210 | 0.6612 |

account the role of $\xi^*$. Recall that the role of $\xi^*$ is to determine a new Cartesian coordinate system under which the evaluated DMU attains the most productive scale size (i.e., CRS efficiency). Lim and Zhu (2014) define the general concept of peer-evaluation in DEA as follows: 'each DMU cross-evaluates other peer DMUs under its own best evaluation environment'. Here *the best evaluation environment* refers to the weights on the input-output factors as well as the new coordinate system that are most favourable to the DMU. Under this best evaluation environment, the DMU itself attains the highest efficiency score as well as the most productive scale size. This concept of peer-evaluation can be implemented by $e_{0j}^{I*}$ (or $e_{0j}^{O*}$) by allowing each DMU to cross-evaluate other peer DMUs under its own best evaluation environment properly represented by all components of the DMU's optimal weights ($v^*$, $u^*$, $\xi^*$).

## 3.5 Conclusions

While cross-efficiency evaluation has been regarded as a powerful extension of DEA, its use has been limited to the case of the CRS model up to recently. In this chapter, we have introduced the approach of Lim and Zhu (2014) for cross-efficiency evaluation

**Table 3.3** Output-oriented VRS cross-efficiency matrix

| | Rated DMU | | | | | | |
|---|---|---|---|---|---|---|---|
| Rating DMU | A | B | C | D | E | F | G |
| A | 1 | $\frac{13}{2}$ | 1 | $\frac{21}{4}$ | $\frac{9}{5}$ | $\frac{13}{6}$ | $\frac{25}{7}$ |
| B | 3 | 3 | $\frac{7}{6}$ | 2 | 1 | 1 | $\frac{9}{7}$ |
| C | 2 | 4 | 1 | 3 | $\frac{6}{5}$ | $\frac{4}{3}$ | 2 |
| D | 5 | 3 | $\frac{31}{18}$ | $\frac{5}{3}$ | $\frac{17}{15}$ | 1 | 1 |
| E | $\frac{7}{3}$ | $\frac{19}{6}$ | 1 | $\frac{9}{4}$ | 1 | $\frac{19}{18}$ | $\frac{31}{21}$ |
| F | 4 | 3 | $\frac{13}{9}$ | $\frac{11}{6}$ | $\frac{16}{15}$ | 1 | $\frac{8}{7}$ |
| G | 5 | 3 | $\frac{31}{18}$ | $\frac{5}{3}$ | $\frac{17}{15}$ | 1 | 1 |
| Output-oriented VRS (simple) efficiency score | 1 | 3 | 1 | $\frac{5}{3}$ | 1 | 1 | 1 |
| Output-oriented VRS cross-efficiency score | 3.1905 | 3.6667 | 1.2937 | 2.5238 | 1.1905 | 1.2222 | 1.6395 |

under VRS, which is based on a novel geometric view of the relationship between the VRS and CRS models.

As noted by Lim and Zhu (2014), two other alternative approaches can be applied. One is to incorporate a non-negativity constraint on $\sum_{r=1}^{s} u_r y_{rj} - \xi$, $\forall j$ in model (3.1), as is suggested by Wu et al. (2009). Although this seems a quick fix, it can be easily observed that this fix distorts the feasible region of multipliers. Specifically, for the example given in Table 3.1, the non-negativity constraints collectively become $\frac{\xi}{u} \leq \min y_j = 1$. Note that, for a feasible solution $(v, u, \xi)$, $\frac{\xi}{u}$ denotes the $y$-intercept of the associated supporting hyperplane. With this distorted feasible region of multipliers, DMUs F and G cannot attain an efficiency score of unity, even though they are VRS-efficient. Therefore, this alternative does not correctly address the issue of negative VRS cross efficiency score.

The other is just not to use problematic (or invalid) optimal weights, such as the one chosen by DMU G in the example, for cross-efficiency calculation. In other words, only optimal weights chosen by DMUs exhibiting CRS are used for cross-efficiency evaluation. While this approach is reasonable in that only valid optimal weights are used in the calculation of cross-efficiency, it may cause an unbalanced (or partial) cross-efficiency evaluation. In other words, it prevents a full range of peer evaluation.

It may appear that under VRS cross efficiency, a small sized DMU can be benchmarked against a large sized DMU. However, this particular issue should not be a concern under the concept of cross efficiency. Lim and Zhu (2014) point out that the

basic idea of cross efficiency is peer evaluation, namely applying one DMU's perspective (manifested in its optimal bundle of weights) to others. As recently pointed out by Cook et al. (2014), although DEA has strong link to production theory in economics, the tool is also used for benchmarking in operations management, where a set of measures is selected to benchmark the performance of manufacturing and service operations. In the circumstance of benchmarking, the efficient DMUs, as defined by DEA, may not necessarily form a "production frontier", but rather lead to a "best-practice frontier". Under this general concept, size or frontier type should not be an issue of concern and the proposed approach works. VRS and CRS are just two terms used to characterize the shapes of the DEA best practice frontier. Without the concept of RTS, these two different shapes of DEA frontiers still exist. VRS simply offers a tighter envelopment. Furthermore, under multiple inputs and multiple outputs, it is difficult to define what constitutes a small- or large-sized DMU. The idea of cross efficiency is to benchmark DMUs against each other, regardless of their size, whether under CRS or VRS. Under CRS cross efficiency, a large-sized DMU is also benchmarked against a small-sized DMU. If one uses the concept of RTS to characterize the shape of the DEA frontier and to classify DMUs, one can clearly see the difference between CRS and VRS cross efficiency. In general, a set of DMUs can be classified into three groups: IRS, CRS, and DRS. Under the CRS cross efficiency, all the CRS efficient facets are applied to IRS and DRS DMUs, while under the VRS cross efficiency IRS, CRS, and DRS efficient facets are applied to all DMUs. Since the general concept of cross efficiency is to look at the performance of a DMU by using other DMUs' weights or facets, it is reasonable to apply IRS, CRS, and DRS facets to all DMUs and to generate VRS cross efficiency

With the approach introduced in the current chapter, we can now use the cross-efficiency concept under the VRS assumption. We note that non-uniqueness of cross-efficiency resulting from multiple optimal multipliers is still an issue with the VRS cross-efficiency approach. As in Doyle and Green (1994), we can add a set of secondary goals in the proposed VRS cross efficiency approach. See also Liang et al. (2008a) and Lim (2012). We can also develop a game cross efficiency approach as in Liang et al. (2008b). These are possible future research topics.

## References

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30(9):1078–1092

Banker RD, Cooper WW, Seiford LM, Zhu J (2011) Returns to scale in data envelopment analysis.In: Cooper WW, Seiford LM, Zhu J (eds) Handbook on data envelopment analysis. Springer, New York

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Cook WD, Tone K, Zhu J (2014) Data envelopment analysis: prior to choosing a model. Omega 44:1–4

Doyle J, Green R (1994) Efficiency and cross-efficiency in DEA: derivations, meanings and uses. J Oper Res Soc 45(5):567–578

Liang L, Wu J, Cook WD, Zhu J (2008a) Alternative secondary goals in DEA cross-efficiency evlauation. Int J Product Econ 113(2):1025–1030

Liang L, Wu J, Cook WD, Zhu J (2008b) The DEA game cross-efficiency model and its nash equilibrium. Oper Res 56(5):1278–1288

Lim S (2012) Minimax and maximin formulations of cross-efficiency in DEA. Comput Ind Eng 62(3):726–731

Lim S, Zhu J (12 March 2014) DEA cross-efficiency evaluation under variable returns to scale. J Oper Res Soc (advance online publication). doi:10.1057/jors.2014.13

Podinovski VV, Bouzdine-Chameeva T (2013) Weight restrictions and free production in data envelopment analysis. Oper Res 61(2):426–437

Wu J, Liang L, Chen Y (2009) DEA game cross-efficiency approach to Olympic rankings. Omega 37(4):909–918

# Chapter 4
# Discrete and Integer Valued Inputs and Outputs in Data Envelopment Analysis

**Timo Kuosmanen, Abolfazl Keshvari and Reza Kazemi Matin**

**Abstract** Standard axioms of free disposability, convexity and constant returns to scale employed in *Data Envelopment Analysis* (DEA) implicitly assume continuous, real-valued inputs and outputs. However, the implicit assumption of continuous data will never hold with exact precision in real world data. To address the discrete nature of data explicitly, various formulations of *Integer DEA* (IDEA) have been suggested. Unfortunately, the axiomatic foundations and the correct mathematical formulation of IDEA technology has caused considerable confusion in the literature. This chapter has three objectives. First, we re-examine the axiomatic foundations of IDEA, demonstrating that some IDEA formulations proposed in the literature fail to satisfy the axioms of free disposability of continuous inputs and outputs, and natural disposability of discrete inputs and outputs. Second, we critically examine alternative efficiency metrics available for IDEA. We complement the IDEA formulations for the radial input measure with the radial output measure and the directional distance function. We then critically discuss the additive efficiency metrics, demonstrating that the optimal slacks are not necessarily unique. Third, we consider estimation of the IDEA technology under stochastic noise, modeling inefficiency and noise as Poisson distributed random variables.

*Abbreviations of key concepts referred to in this chapter*: DEA = Data Envelopment Analysis, DMU = Decision Making Unit, CNLS = Convex Nonparametric Least Squares, IDEA = Integer DEA, MILP = Mixed Integer Linear Programming, RTS = Returns To Scale, SFA = Stochastic Frontier Analysis, StoNED = Stochastic Nonparametric Envelopment of Data.
*Abbreviations of articles frequently cited in this chapter*: KJS = Kuosmanen, Johnson and Saastamoinen (in this volume), KKM = Kuosmanen and Kazemi Matin (2009), KMK = Kazemi Matin and Kuosmanen (2009), KSM = Khezrimotlagh, Salleh, and Mohsenpour (2012, 2013a, 2013b), LV = Lozano and Villa (2006, 2007).

T. Kuosmanen (✉) · A. Keshvari
Aalto University School of Business, Helsinki, Finland
e-mail: timo.kuosmanen@aalto.fi

R. K. Matin
Department of Mathematics, College of Basic Science, Karaj Branch,
Islamic Azad University, Alborz, Iran

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_4

**Keywords** Axiomatic production theory · Efficiency analysis · Mixed integer linear programming · Stochastic noise

## 4.1 Introduction

Data envelopment analysis (DEA, Charnes et al. 1978) is an axiomatic, mathematical programming approach to assessing efficiency of decision making units (DMUs).[1] DEA does not assume any particular functional form for the frontier, but relies on the axioms of production theory, most importantly, free disposability, convexity, and some specification of returns to scale (i.e., variable, non-increasing, non-decreasing, or constant). The standard axioms of free disposability, convexity and constant returns to scale employed in DEA implicitly assume continuous, real-valued inputs and outputs. In contrast, input-output data used in applications are always discrete because the precision of measurement is necessarily restricted to a limited number of decimal digits. Therefore, the implicit assumption of continuous data will never hold with exact precision in real world data.

From a practical point of view, this is not a problem if the observed discrete data can be meaningfully approximated by continuous variables. For example, if the labor input is measured by the number of hours worked, rounded to the nearest integer, and the measured input varies between 1000 and 100,000 h across evaluated DMUs, then the continuous approximation of the discrete data of labor input is perfectly valid as the possible rounding error is small (at most 0.1 %) relative to the measured input. In contrast, if the labor input is the number of workers performing certain function (e.g., firm managers, university professors, hospital physicians), and the DMUs under evaluation are small, the rounding error can become a significant issue. For example, Kuosmanen and Kazemi Matin (2009) consider efficiency analysis of university departments where the number of professors and the number of published articles are examples of integer valued input and output variables. Suppose a university department currently has three professors. Suppose further that the conventional DEA analysis suggests the efficient level of professors is 2.7. How should this result be interpreted? If we round up the efficient number of professors to 3, then the evaluated DMU will appear as efficient, even though the DEA analysis indicates input efficiency of 90 %. However, rounding the input target downwards to 2 may result as an infeasible solution. Since the conventional DEA implicitly assumes all inputs and outputs to be real-valued, the estimated DEA frontier does not necessarily provide meaningful reference points if one simply rounds the input or output targets to the nearest whole number.

[1] We will henceforth use the term "DMU" to refer to any entity that transforms inputs to output, including both non-profit firms and for-profit companies. DMU can refer to a production plant, facility, or sub-division of a company, or to an aggregate entity such as an industry, a region, or a country.

Lozano and Villa (2006, 2007) (henceforth LV) were the first to address this issue explicitly in DEA.[2] They proposed to estimate the production possibility set as the intersection of the standard DEA technology and the set of non-negative integers. Unfortunately, they did not provide any theoretical justification for their integer DEA (henceforth IDEA) technology, even though it is obvious that the proposed technology does not satisfy the standard axioms of free disposability or convexity. To address this problem Kuosmanen and Kazemi Matin (2009) (henceforth KKM) introduced two new axioms of *natural disposability* and *natural divisibility*. Imposing the classic *additivity* axiom (Koopmans 1951), KKM proved that LV's constant returns to scale (CRS) technology has a sound axiomatic foundation. Specifically, they showed that the IDEA technology is the smallest set that contains all observed data points and satisfies the axioms of additivity, natural disposability, and natural divisibility. Subsequent paper by Kazemi Matin and Kuosmanen (2009) (henceforth KMK) extended the result to the variable returns to scale (VRS) case, introducing the axiom of *natural convexity*.

Another contribution of LV is the development of a mixed integer linear programming (MILP) DEA formulation to measure efficiency of DMUs relative to the IDEA technology using Farrell's (1957) radial input-oriented measure. KKM argue that the classic Farrell measure needs to be modified in the context of integer-valued input-output data, and propose to measure efficiency as the radial distance to the monotonic hull of the IDEA technology. They further argue that LV's MILP formulation over-estimates efficiency, and they demonstrate their argument by means of a numerical example and an application.

Following the pioneering works by LV and KKM, a number of extensions and applications of integer DEA have been published (see, e.g., Wu et al. 2009, 2010; Lozano et al. 2011; Kazemi Matin and Emrouznejad 2011; Alirezaee and Sani 2011; Chen et al. 2012; Du et al. 2012; Nöhren and Heinzl 2012; Lozano 2013; Chen et al. 2013). We will survey the extensions and applications in more detail Sect. 4.8 of this chapter.

Unfortunately, the axiomatic foundation and the MILP formulation of integer DEA have also caused serious confusion since the original works by LV. Recently, a series of papers by Khezrimotlagh et al. (2012, 2013a, 2013b) (henceforth KSM) have contributed to further confusion by discrediting the contributions of KKM and disregarding both the importance of a sound axiomatic foundation and rigorous mathematical formulations. While the bogus critique by KSM is not worth serious consideration, the naïve mistakes of KSM provided us some further motivation to elaborate our arguments and shed some new light on the intimate connection between the axioms of production theory and the implementation through MILP.

[2] Previous studies such as Banker and Morey (1986), Kamakura (1988), and Rousseau and Semple (1993) (among others) consider inputs and outputs measured on the *categorical* or *ordinal* scale, which are obviously integer valued. However, input-output variables defined on the *interval* or *ratio* scales can be integer valued as well.

The purposes of this chapter are three-fold. First, we re-examine the axioms and MILP formulations of integer DEA, elaborating some aspects that have apparently caused confusion in the literature. Emphasizing the importance of the axiomatic foundation, we demonstrate that LV's MILP formulations fail to satisfy the axioms of free disposability of continuous inputs and outputs, and natural disposability of discrete inputs and outputs. We illustrate the inconsistency of LV's MILP formulation with the IDEA technology they suggested through detailed numerical examples, which demonstrate the differences between the LV's formulation and those developed by KKM and KMK.

Second, we critically examine alternative efficiency metrics available for integer DEA. We complement the MILP formulations for the radial input oriented Farrell (1957) measure proposed by KKM and KMK with the radial output oriented measure, and the general directional distance function (Chambers et al. 1996, 1998). We then critically discuss the additive efficiency metrics considered by LV (2007), demonstrating that the optimal slacks are not necessarily unique. The same problem applies to the range adjusted additive measure proposed by Cooper et al. (1999). The non-uniqueness of slacks can make the application of the slack based measure by Tone (2001) problematic in the context of integer DEA.

Third, attributing all deviations from the frontier to inefficiency, ignoring stochastic noise, is generally recognized as the main limitation of DEA (see Kuosmanen, Johnson and Saastamoinen, in this volume, (henceforth KJS) for a review of recent advances in modeling noise). To address this shortcoming, we examine the estimation of the IDEA technology in the single output setting under stochastic noise. Modeling inefficiency and noise as Poisson distributed random variables, we outline the first extension of stochastic nonparametric envelopment of data (StoNED) approach by Kuosmanen and Kortelainen (2012) to discrete output variables.

The rest of this chapter is organized as follows. Section 4.2 introduces and discusses the axioms for a DEA problem with integer-valued inputs and outputs. Section 4.3 derives the associated DEA production sets that satisfy the fundamental minimum extrapolation principle,[3] and generalize the method to the hybrid case where both real and integer valued inputs and outputs are present. Section 4.4 modifies the Farrell input efficiency measure to the integer DEA setting, and show how the efficiency score can be computed by solving a MILP problem. Section 4.5 discusses new developments on integer DEA and some extensions. Section 4.6 presents concluding discussion with some potential avenues for future research. The paper includes several theorems: proofs of all theorems and lemmas are presented in the Appendix.

[3] The minimum extrapolation principle was formally introduced by Banker et al. (1984), but formal minimum extrapolation theorems (and proofs) date back at least to Afriat (1972).

## 4.2 Axioms

The axiomatic approach to constructing production possibility sets as a combination of observed activities has a long history in economics, dating back at least to Von Neumann (1945–1946) and Koopmans (1951). Afriat (1972) was the first to prove the minimum frontier production functions that envelop all observed data and satisfy the following sets of axioms: i) free disposability, ii) convexity and free disposability, and iii), CRS, convexity and free disposability. Banker et al. (1984) extended Afriat's result to the multi-output production possibility sets, and formally introduced the fundamental *minimum extrapolation principle*.

Multi-output production technology can be generally characterized by the production possibility set $T$ defined as

$$T = \left\{(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in \mathbb{R}_+^m \text{ can produce } \mathbf{y} \in \mathbb{R}_+^s\right\},$$

where $\mathbf{x}$ is a $m$-dimensional vector of input quantities and $\mathbf{y}$ is a $s$-dimensional vector of output quantities.[4] Intuitively, the set $T$ can be understood as a list of feasible input-output combinations. Even if we restrict to discrete or integer valued input-output vectors, in general, there are infinitely many feasible input-output vectors, which makes the list infinitely long. It is worth emphasizing that, in many applications, the production possibility set $T$ is interpreted as the *benchmark technology* that forms a reference for performance comparisons and efficiency analysis. In this interpretation, the boundary of set $T$ characterizes standards for good performance, not only the production possibilities from the strictly technical point of view.

Observed DMUs are characterized by a pair of non-negative input and output vectors $(\mathbf{x}_j, \mathbf{y}_j) j \in J = \{1, \ldots, n\}$. Conventional DEA approaches implicitly assume that all inputs and outputs are continuous, real-valued variables. However, observed data are always discrete as the number of decimal digits is necessarily finite. This forms the motivation for integer DEA. Note that any *discrete* data that cannot be meaningfully approximated as continuous data can easily be converted to integers by a simple multiplicative transformation. Suppose, for example, that a continuous output variable is measured at the precision of one decimal digit (e.g., 0, 0.1, 0.2, ...), but rounding the DEA targets to the nearest decimal digit seems problematic for one reason or another. This discrete output variable an be harmlessly multiplied by factor 10 (amounting to a change of units of measurement), which results as an integer valued output variable.

In the following we will focus on integer-valued inputs and outputs $(\mathbf{x}, \mathbf{y}) \in \mathbb{Z}_+^{m+s}$, which lead us to integer DEA (IDEA) introduced by LV. In the following sub-sections we will adapt the classic axioms of DEA to allow for integer valued inputs and outputs, following KKM and KMK.

[4] For clarity, we denote vectors by bold lower case letters (e.g., $\mathbf{x}$) and matrices by bold capital letters (e.g., $\mathbf{X}$).

### 4.2.1 Free Disposability and Natural Disposability

Free disposability is an intuitive and widely used axiom. It is closely related to monotonicity of functional representations of technology: free disposability implies that the production function is monotonic increasing in inputs and the cost function is monotonic increasing in outputs. It is possible to assess efficiency relying solely on the free disposability axiom, using the free disposable hull (FDH) method (Deprins et al. 1984; Tulkens 1993). However, free disposability is not always a meaningful axiom. For example, if the output vector $\mathbf{y}$ includes undesirable outputs (bads) such as waste or pollution, the free disposability axiom can be replaced by the weak disposability axiom.[5] Free disposability is also relaxed for modeling congestion. [6]

The axiom of free disposability is conventionally stated as follows:

(A1) *Free disposability*: $(\mathbf{x}, \mathbf{y}) \in T \quad \text{and} \quad (\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^{m+s}, \quad \mathbf{y} \geq \mathbf{v} \quad \Rightarrow (\mathbf{x} + \mathbf{u}, \mathbf{y} - \mathbf{v}) \in T$.

This axiom states that it is always possible to produce less output with a given level of inputs, or alternatively, use more inputs to produce the same amount of output. Vector $\mathbf{u}$ can be interpreted as the amount of excess inputs used, and vector $\mathbf{v}$ represents the foregone output. If we interpret this axiom literally, it seems impossible to consume infinite amounts of inputs in a finite production process. Hence axiom (A1) is not necessarily valid from a purely technical point of view. However, it does have a compelling economic interpretation: (A1) essentially states that inefficient production (in the sense of Koopmans 1951) is feasible. Stated differently, if our objective is to assess technical efficiency in the sense of Koopmans (1951), and we interpret $T$ as a benchmark technology rather than as a list of technically feasible points, then (A1) is a completely harmless axiom irrespective of whether it is technically feasible or not.

Axiom (A1) implies continuity. Clearly, if this axiom holds, then there are feasible real-valued input-output vectors $(\mathbf{x}, \mathbf{y}) \in T$ that are not included in $\mathbb{Z}_+^{m+s}$. Stated conversely, if the production possibility set $T$ contains only integer-valued input-output vectors, then it cannot satisfy the standard free disposability axiom. Therefore, it is necessary to adapt this axiom to be consistent with integer-valued inputs and outputs. KKM propose the following axiom:

(B1) *Natural disposability*: $(\mathbf{x}, \mathbf{y}) \in T \quad \text{and} \quad (\mathbf{u}, \mathbf{v}) \in \mathbb{Z}_+^{m+s}, \quad \mathbf{y} \geq \mathbf{v} \quad \Rightarrow (\mathbf{x} + \mathbf{u}, \mathbf{y} - \mathbf{v}) \in T$.

The economic rationale of axiom (B1) is exactly the same as that of the standard free disposability axiom (A1): inefficient production is feasible. However, (B1) only allows for integer-valued disposal of outputs through vector v and integer-valued excess inputs through vector u. Therefore, axiom (B1) is a suitable counterpart of (A1) that applies for integer valued inputs and outputs.

---

[5] The correct way of implementing weak disposability in DEA has caused some confusion in the literature: see Kuosmanen (2005), Färe and Grosskof (2009), Kuosmanen and Podinovski (2009), and Podinovski and Kuosmanen (2011) for an interesting debate on this issue.

[6] This is another issue that has caused confusion: see Cherchye et al. (2001).

### 4.2.2 Convexity and Natural Convexity

The classic DEA approaches (Farrell 1957; Charnes et al. 1978; Banker et al. 1984) impose convexity in addition to free disposability. The standard convexity axiom can be stated as follows:

(A2) *Convexity*: $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathrm{T},\ (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$
$= \lambda(\mathbf{x}, \mathbf{y}) + (1 - \lambda)(\mathbf{x}', \mathbf{y}'),\ 0 \leq \lambda \leq 1 \ \Rightarrow\ (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T.$

This axiom states that convex combinations of observed DMUs are always feasible. The weights assigned to the observations are characterized by parameter $\lambda$. In general, we can form convex combinations of all $n$ observations in $J$ using a $n$-dimensional parameter vector $\lambda$.

Convexity does not necessarily have a strong justification from the technical point of view, but it is a fundamental axiom in economic theory. For example, convexity is critically important for establishing duality results between alternative representations of technology (Shephard 1970; Färe and Primont 1995). For example, if the profit function of a firm is known, we can always recover the convex hull of its production possibility set $T$ (see Kuosmanen 2003, for details). If we interpret $T$ as a benchmark technology for competitive profit maximizing firms that take prices as given, then convexity is an equally harmless axiom as free disposability. However, if we consider nonprofit firms or monopolistic competition, convexity may be a restrictive assumption as it assumes away economies of scale (see, e.g., Kuosmanen 2001). Weaker forms of quasi-convexity (i.e., convex input or output sets) have also been considered in the DEA literature (e.g., Petersen 1990; Bogetoft 1996; Bogeoft et al. 2000; Post 2001).

Clearly, if axiom (A2) holds, then there are feasible real-valued input-output vectors $(\mathbf{x}, \mathbf{y}) \in T$ that are not integer-valued. Conversely, if the production possibility set $T$ contains only integer-valued input-output vectors, then it violates convexity. Therefore, it is necessary to adapt this axiom to be consistent with integer-valued inputs and outputs. KMK propose the following axiom:

(B2) *Natural convexity*: $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in T,\ (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \lambda(\mathbf{x}, \mathbf{y}) + (1 - \lambda)(\mathbf{x}', \mathbf{y}')$, $0 \leq \lambda \leq 1$ and $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in \mathbb{Z}_+^{m+s} \ \Rightarrow\ (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T$.

Analogous to the pair of axioms (A1) and (B1), the rationale of axiom (B2) is to adapt (A2) to the context of integer-valued inputs without changing its meaning. Note that (B2) only adds to (A2) the requirement that $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in \mathbb{Z}_+^{m+s}$, that is, the resulting convex combination must itself be integer-valued. Note that KMK allow the weights $\lambda$ used for forming convex combinations to be real valued. They do not see a problem in using real valued numbers in the mathematical operations involved in the axioms as far as the resulting input-output vectors are integer-valued.

KSM (2012) criticize KMK for the use of real valued weights $\lambda$ for forming convex combinations.[7] They propose to substitute weights $\lambda$ in (B2) by the ratio $u/v$,

[7] KSM (2012) state: "*Now, if it has been supposed that only the integer numbers set is considered, then it should not have been used the real number variable in the integer axioms! In fact, a new*

such that $u \leq v,\ u, v \in \mathbb{Z}_+$. Mathematically, this restricts the domain of weights $\lambda$ from the real numbers to the set of rational numbers. Therefore, the alternative axiom proposed by KSM does not expand the production possibility set, it can only contract it. In fact, we can prove the following:

**Lemma 1** *Assume Axiom (B2) is satisfied. Then for any given* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in T$, *if there exists a real valued* $\lambda$ *such that* $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \lambda(\mathbf{x}, \mathbf{y}) + (1 - \lambda)(\mathbf{x}', y') \in T$, *then there exist integers* $u, v \in \mathbb{Z}_+,\ u \leq v$, *such that*

$$\lambda = u/v$$

This lemma shows that the alternative convexity axiom proposed by KSM makes no difference whatsoever. If one finds the axiom by KSM more aesthetic or elegant, one can harmlessly use it, without a need to revise the theory developed by KKM. However, for the sake of intuition and transparency, we prefer to maintain a close connection between the axioms for real valued and integer valued variables (i.e., axioms A and axioms B). Since there is no real benefit from restricting the domain of weights $\lambda$ from the set of real numbers to the set of rational numbers, this is only a matter of subjective preference. In this light, the claims about "*major shortcomings*" that KSM repeatedly express in their papers are completely irrational.

### *4.2.3 Returns to scale*

Returns to scale concerns radial contraction or expansion of all inputs and outputs by the same factor. Note that if no axioms concerning returns to scale are imposed, then the technology is said to exhibit *variable returns to scale* (VRS). To implement VRS in DEA, the weights $\lambda$ employed for forming convex combinations of observed DMUs must sum to one (i.e., $\sum_{j=1}^{n} \lambda_j = 1$). When further axioms concerning returns to scale are imposed, this constraint can be relaxed.

Consider first the radial contraction possibilities. The conventional axiom of non-increasing returns to scale (NIRS) can be stated as follows:

(A3) *Non-increasing returns to scale*: $(\mathbf{x}, \mathbf{y}) \in T$ and $0 \leq \lambda \leq 1 \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

This axiom allows one to scale down any observed input-output vector by factor $\lambda$. Note that axiom (A3) implies that inactivity is feasible: the origin (**0**,**0**) is included in the production possibility set $T$ because, starting from any observed $(\mathbf{x}, \mathbf{y})$, we can set factor $\lambda = 0$. If we simply insert the origin (**0**,**0**) as one of the observed points in the data set, then the variable returns to scale DEA technology will automatically satisfy axiom (A3). This provides an implicit way of implementing NIRS, which may be useful in some context (see Kuosmanen 2005). A more standard way of

*axiom must not have any doubts or parallel affects with those previous axioms. In other words, an axiom is an evident premise as to be accepted as true without controversy.*" This discussion reveals that KSM do not understand the economic meaning of axioms in DEA. In fact, none of the standard DEA axioms can meet the requirements of KSM.

implementing NIRS in DEA is to set a constraint that the sum of intensity weights must be less than or equal to one (i.e., $\sum_{j=1}^{n} \lambda_j \leq 1$).

Clearly, even if we start from an inter-valued input-output vector $(\mathbf{x}, \mathbf{y}) \in \mathbb{Z}_+^{m+s}$, the rescaled vector $(\lambda\mathbf{x}, \lambda\mathbf{y})$ is not necessarily integer valued. Therefore, axiom (A3) is not directly applicable for integer DEA. KKM propose to modify axiom (A3) as

(B3) *Natural divisibility*: $(\mathbf{x}, \mathbf{y}) \in T$ and $0 \leq \lambda \leq 1$ and $(\lambda\mathbf{x}, \lambda\mathbf{y}) \in \mathbb{Z}_+^{m+s} \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

Natural divisibility simply introduces an additional restriction that the downward rescaled version of the original input-output vector must result as an integer valued production plan to be feasible.

Consider next the radial expansion. The conventional axiom of non-decreasing returns to scale (NDRS) can be stated as follows:

(A4) *Non-decreasing returns to scale*: $(\mathbf{x}, \mathbf{y}) \in T \quad \text{and} \quad \lambda \geq 1 \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

This axiom allows for radial expansion of any observed input-output vector away from the origin by factor $\lambda \geq 1$. The NRDS axiom is implemented in DEA by enforcing the sum of intensity weights to be greater than or equal to one (i.e., $\sum_{j=1}^{n} \lambda_j \geq 1$).

Obviously, the rescaled vector $(\lambda\mathbf{x}, \lambda\mathbf{y})$ does not have to be integer valued. Therefore, KMK propose to adapt this axiom for integer DEA as

(B4) *Natural augmentability*: $(\mathbf{x}, \mathbf{y}) \in T \quad \text{and} \quad \lambda \geq 1 \quad \text{and} \quad (\lambda\mathbf{x}, \lambda\mathbf{y}) \in \mathbb{Z}_+^{m+s} \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

Natural augmentability requires that the radial expansion must result as an integer valued input-output vector in order to be feasible.

Note that in both (B3) and (B4), KMK assume a real-valued multiplier $\lambda$. In both cases, we could equally well express $\lambda$ as a ratio of two integers.

**Lemma 2** *For any given* $(\mathbf{x}, \mathbf{y}) \in T$, *if there exists a real valued* $\lambda$ *such that* $(\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$, *then there exist integers* $u, v \in \mathbb{Z}_+$, $u \leq v$, *such that*

$$\lambda = u/v.$$

This result again shows that the alternative formulations of the KMK axioms suggested by KSM do not make any practical difference whatsoever.

Finally, if both (A3) and (A4) hold, then the technology is said to satisfy constant returns to scale (CRS):

(A5) *Constant returns to scale*: $(\mathbf{x}, \mathbf{y}) \in T \quad \text{and} \quad \lambda \geq 0 \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

In the CRS case, the sum of intensity weights $\lambda$ is unrestricted. Observe that imposing additional axioms on returns to scale implies less restrictive constraints for the intensity weights $\lambda$, which leads to the expansion of the estimated production possibility set.

From a pure technical point of view, the CRS axiom appears totally unrealistic. However, it does have compelling economic justification in many applications. If the objective of the firm is to maximize profitability (i.e., the ratio of revenue to cost) at given prices, then the CRS axiom is completely harmless (see Kuosmanen et al. 2004; Lemma 1).

KMK did not introduce an integer equivalent of (A5): note that if both (A3) and (A4) hold, then (A5) holds. The converse is also true. Therefore, the CRS case is obtained in integer DEA by imposing (B3) and (B4). For the sake of completeness, we can state the integer version of (A5) as

(B5) *Natural radial rescaling*: $(\mathbf{x}, \mathbf{y}) \in T$ and $\lambda \geq 0$ and $(\lambda\mathbf{x}, \lambda\mathbf{y}) \in \mathbb{Z}_+^{m+s} \Rightarrow (\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$.

In fact, KKM examine the CRS case in detail, imposing the axiom of additivity (adopted from Koopmans 1951) in addition to natural divisibility (B3).

(A6) *Additivity*: $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in T \Rightarrow (\mathbf{x} + \mathbf{x}', \mathbf{y} + \mathbf{y}') \in T$.

Since axiom (A6) was first introduced in the context of continuous variables, we label it as type-A axiom. Note, however, that the additivity axiom does not require or imply continuity, and hence it applies equally well to integer valued inputs and outputs. Interestingly, we can build the IDEA technology under CRS to the axioms of additivity and natural divisibility axioms, as shown by the following result:

**Lemma 3** *If the axioms* (B2) *Natural convexity and* (B5) *Natural radial rescaling are satisfied, then the axioms of* (B3) *Natural divisibility and* (A6) *Additivity must also hold. Conversely, if axioms* (B3) *and* (A6) *are satisfied, then axioms* (B2) *and* (B5) *must also hold. In other words, these two pairs of axioms are equivalent in the following sense:*

$$[\text{(B2) and (B5)}] \Leftrightarrow [\text{(B3) and (A6)}].$$

### *4.2.4 Envelopment*

In addition to the standard axioms of production theory (e.g., Shephard 1970; Färe and Primont 1995), the classic DEA article by Banker et al. (1984) imposes the following axiom:

(E1) *Envelopment*: all observed data points $(\mathbf{x}_j, \mathbf{y}_j)$ are feasible: $(\mathbf{x}_j, \mathbf{y}_j) \in T \ \forall j \in J$.

For clarity, we label this assumption as type-E postulate, as (E1) not really an axiom in the same sense as (A1)–(A6) and (B1)–(B5) considered above. Note that all axioms introduced before are conditional statements expressed using $\Rightarrow$ (i.e., if condition "A" holds, then "B" is feasible). In contrast, (E1) is an unconditional statement about the observed data. In our interpretation, the minimum extrapolation principle together with (E1) form the estimation principle of DEA analogous to the minimization of least squares or the maximization of the log-likelihood function in regression analysis.

In a technical sense, (E1) is a natural and intuitive axiom: if point $(\mathbf{x}_j, \mathbf{y}_j)$ is observed, then it clearly must be feasible. One could argue that this axiom is proved by empirical evidence.

However, the fact that $(\mathbf{x}_j, \mathbf{y}_j)$ is observed once does not necessarily guarantee that DMU $j$ can replicate $(\mathbf{x}_j, \mathbf{y}_j)$ again in the future, or that other DMUs can achieve the

**Table 4.1** Axioms considered in this paper

| Axioms that imply continuity | | Corresponding axioms for discrete variables | |
|---|---|---|---|
| (A1) | Free disposability | (B1) | Natural disposability |
| (A2) | Convexity | (B2) | Natural convexity |
| (A3) | Non-increasing returns to scale | (B3) | Natural divisibility |
| (A4) | Non-decreasing returns to scale | (B4) | Natural augmentability |
| (A5) | Constant returns to scale | (B5) | Natural radial rescaling |
| *Other axioms/conditions* | | | |
| (A6) | Additivity | | |
| (E1) | Envelopment | | |

point $(\mathbf{x}_j, \mathbf{y}_j)$. In many applications of efficiency analysis, production process is subject to uncontrollable random elements, including technological risks (e.g., machine failure). There are also economic risks (e.g., variation in demand and input-output prices), and risks related to the operating environment (e.g., competition, regulation, weather conditions). In practice, DEA can handle a limited number of input and output variables,[8] and hence one often needs to either omit some relevant inputs or outputs, or resort to aggregated inputs and outputs (e.g., monetary cost or revenue aggregates) that are subject to errors of aggregation. While DEA implicitly assumes homogenous DMUs that operate in a homogenous environment, in reality, evaluated DMUs tend to be heterogenous and operate in heterogenous environments. The random variations, omitted variables, data errors, and hetereogeneity are some of the possible reasons for why the envelopment condition (E1) is not valid in applications.

The recent works by Kuosmanen (2008), Kuosmanen and Johnson (2010), and Kuosmanen and Kortelainen (2012) demonstrate that it is possible to relax the envelopment condition (E1), and estimate production technologies subject to some of the axioms (A1)–(A6) in a nonparametric or semi-nonparametric fashion (see KJS for a review). We consider the CNLS (*convex nonparametric least squares*) and StoNED (*stochastic nonparametric envelopment of data*) developed in these papers a promising way forward. An extension of StoNED method to IDEA technology will be developed in Sect. 4.6. To pave a way for the stochastic extension, we will maintain the assumption (E1) in Sects. 4.3–4.5.

To summarize this section, Table 4.1 lists the axioms considered, indicating the standard axioms (A1)–(A5) that imply continuity and the corresponding axioms (B1)–(B5) for discrete, integer-valued variables, and the other axioms/ conditions.

[8] DEA is a nonparametric estimator subject to the *curse of dimensionality*. This implies that the precision of DEA estimator deteriorates rapidly as the number of input and output variables increases. Also the discriminating power of DEA is affected: when the dimensionality is large, almost all DMUs appear as inefficient.

## 4.3 Continuous, Integer-Valued and Hybrid DEA Technologies

Having introduced the axioms, we will next examine the continuous and integer-valued DEA estimators of the production possibility set $T$, and a hybrid case where some inputs and outputs are integer-valued while others are continuous.

Applying the fundamental *minimum extrapolation principle*, any DEA technology can be constructed as the intersection of such sets $S \subset \mathbb{R}_+^{m+s}$ that contain all observed DMUs (E1) and satisfy the stated axioms (Banker et al. 1984). In the case of continuous input-output variables, the DEA estimator of the production possibility set $T$ can be stated as

$$\begin{aligned} T_{DEA}^{RTS} &= \{(\mathbf{x},\mathbf{y}) \in \mathbb{R}_+^{m+s} \\ &\text{subject to} \\ &\mathbf{x} \geq \sum\nolimits_{j=1}^{n} \mathbf{x}_j \lambda_j; \\ &\mathbf{y} \leq \sum\nolimits_{j=1}^{n} \mathbf{y}_j \lambda_j; \\ &\boldsymbol{\lambda} \in \Lambda_{RTS}\}, \end{aligned}$$

where $\Lambda_{RTS}$ denotes the generic domain of intensity weights under alternative RTS specifications. Specifically, $\Lambda_{RTS}$ can be specified by choosing one of the four options below:

$$\begin{aligned} \Lambda_{VRS} &= \left\{\sum\nolimits_{j=1}^{n} \lambda_j = 1;\ \boldsymbol{\lambda} \geq 0\right\} \\ \Lambda_{NIRS} &= \left\{\sum\nolimits_{j=1}^{n} \lambda_j \leq 1;\ \boldsymbol{\lambda} \geq 0\right\} \\ \Lambda_{NDRS} &= \left\{\sum\nolimits_{j=1}^{n} \lambda_j \geq 1;\ \boldsymbol{\lambda} \geq 0\right\} \\ \Lambda_{CRS} &= \{\boldsymbol{\lambda} \geq 0\} \end{aligned}$$

This generic domain allows for alternative specifications of RTS known in the DEA literature. The connection to the axioms introduced in Sect. 4.2 is the following. Under axioms (A1) + (A2), we have the VRS specification. $\Lambda_{VRS}$ Axioms (A1) + (A2) + (A3) imply the NIRS specification $\Lambda_{NIRS}$, while axioms (A1) + (A2) + (A4) imply the NDRS specification $\Lambda_{NDRS}$. Under axioms (A1) + (A2) + (A5), we have the CRS specification $\Lambda_{CRS}$.

Banker et al. (1984) formally show that set $T_{DEA}^{RTS}$ satisfies the envelopment condition (E1) and the stated axioms, and that $T_{DEA}^{RTS}$ is the intersection of all such sets that satisfy those axioms. In this sense, $T_{DEA}^{RTS}$ is the smallest set that satisfies the stated axioms.[9] Note that the axiom of convexity is implemented through the use

[9] In the single output case, Afriat (1972) proves the similar minimum extrapolation result for the smallest production function satisfies axioms (A1), (A1) + (A2), or (A1) + (A2) + (A5).

of intensity weights $\lambda_j$ (compare with axiom (A2)), which allow for any convex combination of observed DMUs. Restricting weights $\lambda_j$ to be integers relaxes the convexity axiom (A2), leading to the free disposable hull (Deprins et al. 1984) and free replicable hull (Tulkens 1993) technologies. The axiom of free disposability is implemented through the inequality constraints for inputs and outputs. Replacing the inequality constraints by equality constraints would relax the free disposability axiom (A1), leading to the DEA formulations of weak disposability (e.g. Kuosmanen 2005) and congestion (e.g. Cherchye et al. 2001).

In the case of integer-valued inputs and outputs, the generic IDEA technology first proposed by LV can be similarly stated as

$$\begin{aligned} T^{RTS}_{IDEA} &= \{(\mathbf{x},\mathbf{y}) \in \mathbb{Z}^{m+s}_{+} \\ &\text{subject to} \\ &\mathbf{x} \geq \sum\nolimits_{j=1}^{n} \mathbf{x}_j \lambda_j; \\ &\mathbf{y} \leq \sum\nolimits_{j=1}^{n} \mathbf{y}_j \lambda_j; \\ &\boldsymbol{\lambda} \in \Lambda_{RTS}\}, \end{aligned}$$

where $\Lambda_{RTS}$ is the generic domain of intensity weights under alternative RTS specifications introduced above. In the case of the IDEA technology, axioms (B1) + (B2) imply the VRS specification $\Lambda_{VRS}$. Under axioms (B1) + (B2) + (B3) we have the NIRS specification $\Lambda_{NIRS}$, and axioms (B1) + (B2) + (B4) imply the NDRS specification $\Lambda_{NDRS}$. Finally, the CRS specification $\Lambda_{CRS}$ is obtained under axioms (B1) + (B2) + (B5).

Comparing the sets $T^{RTS}_{IDEA}$ and $T^{RTS}_{DEA}$, it is obvious that $T^{RTS}_{IDEA} \subset T^{RTS}_{DEA}$. LV correctly note that

$$T^{RTS}_{IDEA} = T^{RTS}_{DEA} \cap \mathbb{Z}^{m+s}_{+}.$$

In words, IDEA technology is the intersection of the set of integer vectors and the conventional DEA technology, and the latter set is further an intersection of all such sets that satisfy (E1), (A1), (A2), and the specified RTS axioms. However, it is easy to see that $T^{RTS}_{IDEA}$ itself does not satisfy any of the axioms (A1) – (A4). This is the reason why KKM criticized LV for the lack of axiomatic foundation. It is not enough that a benchmark technology is an intersection of some arbitrary sets: the minimum extrapolation principle requires that $T^{RTS}_{IDEA}$ itself satisfies the stated axioms, and is the smallest set that does so.

Fortunately, the axiomatic foundation can be established using the parallel set of axioms (B1) – (B4), as shown by KKM and KMK. The minimum extrapolation theorems by KKM and KMK can be formally summarized as follows:

**Theorem 1** *Production possibility set $T^{RTS}_{IDEA}$ is the intersection of all sets $S \subset \mathbb{Z}^{m+s}_{+}$ that satisfy the envelopment (E1), axioms (B1) and (B2), the RTS axioms ((B3), (B4), (B5), or none) corresponding to the specified returns to scale.*

Note that axioms (B1) and (B2) could be relaxed in the same way as in the standard DEA technology. If the inequality constraints for inputs and outputs are replaced by equality constraints, this amounts to relaxing axiom (B1). Similarly, if intensity weights $\lambda_j$ are restricted to be binary integers, the convexity axiom (A2) is relaxed. We emphasize the direct correspondence between the axioms and the mathematical formulations of alternative DEA technologies.

In addition to the settings where all inputs and outputs are either continuous or integer valued, in many applications of IDEA some of the input-output variables are integer valued while others can be meaningfully approximated as continuous variables. Following LV, KKM, and KMK, this case will be henceforth referred to as the *hybrid integer DEA* (HIDEA). In general, we can partition the set of input variables as $I = I^I \cup I^N$ and the set of output variables as $O = O^I \cup O^N$, where subsets $I^I$ and $O^I$ contain the integer valued inputs and outputs, respectively, whereas subsets $I^{NI}$ and $O^{NI}$ include the real valued inputs and outputs. Without loss of generality, subsets $I^I$ and $I^{NI}$, as well as $O^I$ and $O^{NI}$, are assumed to be mutually disjoint, and $\left|I^I\right| = p \leq m$ and $\left|O^I\right| = q \leq s$. Applying these notations, we can state any non-negative input and output vectors $(\mathbf{x}, \mathbf{y})$ as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^I \\ \mathbf{x}^{NI} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}^I \\ \mathbf{y}^{NI} \end{pmatrix}.$$

In the hybrid setting, we can impose type-A axioms for the continuous inputs and outputs included in $I^{NI}$ and $O^{NI}$, while type-B axioms are used for integer-valued inputs and outputs included in $I^I$ and $O^I$. In practice, we can formulate the HIDEA technology as

$$T^{RTS}_{HIDEA} = \left\{ \begin{pmatrix} \mathbf{x}^I & \mathbf{y}^I \\ \mathbf{x}^{NI} & \mathbf{y}^{NI} \end{pmatrix} \right.$$

subject to

$$(\boldsymbol{x}^I, \mathbf{y}^I) \in \mathbb{Z}^{p+q}_+;$$

$$\begin{pmatrix} \mathbf{x}^I \\ \mathbf{x}^{NI} \end{pmatrix} \geq \sum\nolimits_{j=1}^{n} \begin{pmatrix} \mathbf{x}^I_j \\ \mathbf{x}^{NI}_j \end{pmatrix} \lambda_j;$$

$$\begin{pmatrix} \mathbf{y}^I \\ \mathbf{y}^{NI} \end{pmatrix} \leq \sum\nolimits_{j=1}^{n} \begin{pmatrix} \mathbf{y}^I_j \\ \mathbf{y}^{NI}_j \end{pmatrix} \lambda_j;$$

$$\left. \boldsymbol{\lambda} \in \Lambda_{RTS} \right\},$$

where $\Lambda_{RTS}$ are specified for VRS, NIRS, NRDS, or CRS as noted above. Note that the same set of intensity weights $\lambda_j$ are used for both integer-valued and continuous

input-output variables. However, the constraint $(\boldsymbol{x}^I, \boldsymbol{y}^I) \in \mathbb{Z}_+^{p+q}$ only applies to the subset of integer-valued input-output variables.

The next theorem generalizes the axiomatic foundation established in Theorem 1 to this hybrid setting.

**Theorem 2** *Production possibility set $T_{HIDEA}^{RTS}$ is the intersection of all sets S that satisfy the envelopment (E1), axioms (A1) and (A2) for the subsets $(I^{NI}, O^{NI})$, axioms (B1) and (B2) for the subsets $(I^I, O^I)$, and the RTS axioms ((A3), (A4), (A5), or none for the subsets $(I^{NI}, O^{NI})$, and (B3), (B4), (B5), or none for for the subsets $(I^I, O^I)$) corresponding to the specified returns to scale.*

In addition to these symmetric cases where the real-valued and integer-valued variables exhibit the same type of returns to scale, it could be interesting to allow the returns to scale differ for the real-valued and integer-valued variables, in the spirit of the *hybrid returns to scale* technology by Podinovski (2004). For example, in some applications it might be reasonable to assume the real-valued variables are subject to VRS, while the integer-valued variables exhibit CRS. Extending the HIDEA problem to hybrid returns to scale specifications falls beyond the scope of the present paper, and is left as an interesting topic for future research.

## 4.4 Efficiency Measures and Distance Functions

### *4.4.1 Modified Farrell Input Efficiency Measure*

Having introduced the IDEA and HIDEA technologies, we will next examine the measurement of efficiency as a distance from the observed input-output vector of the evaluated DMU to the efficient boundary of the benchmark technology. Before proceeding, we must stress that the standard efficiency measures (including the radial Farrell input and output measures, the additive Pareto-Koopmans efficiency measures, and the directional distance functions) all implicitly assume continuous, real-valued inputs and outputs. Consider, for example, the classic Farrell input efficiency measure, defined as

$$Eff^{In}(\mathbf{x}_0, \mathbf{y}_0) = \min\{\theta | (\theta \mathbf{x}_0, \mathbf{y}_0) \in T\},$$

where vector $(\mathbf{x}_0, \mathbf{y}_0)$ is the input-output vector of the DMU under evaluation (which can be one of the observed DMUs or a hypothetical unit of interest). Value $\theta = 1$ indicates full efficiency, and values $\theta < 1$ imply the evaluated DMU is inefficient: $100\% \times (1 - \theta)$ indicates the degree of inefficiency. Unfortunately, applying the standard Farrell measure directly to $T_{IDEA}$ is likely problematic because $T_{IDEA}$ is essentially a discrete set of disconnected points. Hence, applying the standard Farrell measure as such can yield very strange, counterintuitive results. For example, it is possible that $Eff^{In}(\mathbf{x}_0, \mathbf{y}_0) = 1$ for input-output vector $(\mathbf{x}_0, \mathbf{y}_0)$ that is strictly dominated by another point in $T_{IDEA}$.

To avoid complications due the discrete nature of IDEA and HIDEA technologies, KKM propose to modify the Farrell input efficiency measure as:

$$Eff^{In+}(\mathbf{x}_0, \mathbf{y}_0) = \min\left\{\theta \in \mathbb{R}_+ | \exists(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T \ : \widetilde{\mathbf{x}}^I \in \mathbb{Z}_+^p \ ; \theta\mathbf{x}_0 \geq \widetilde{\mathbf{x}}; \mathbf{y}_0 \leq \widetilde{\mathbf{y}}\right\}.$$

This modified Farrell measure gauges radial distance to the monotonic hull of the benchmark technology, requiring that the reference point $(\theta\mathbf{x}_0, \mathbf{y}_0)$ must be dominated by a feasible input-output vector $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ which has integer-valued inputs for the subset $I^I$. For the sake of completeness, note that we could also add a requirement that $\widetilde{\mathbf{y}}^I \in \mathbb{Z}_+^q$, but this would be completely redundant in the case of input-oriented efficiency measure.

The modified input efficiency measure $Eff^{In+}$ preserves the usual interpretation of the Farrell measure as a downward scaling potential in inputs at the given output level. It guarantees that DMUs assigned the efficiency score one are weakly efficient in the Pareto-Koopmans sense. Unfortunately, the original papers by LV suggested MILP formulations for computing the radial input- and output-oriented efficiency measures without explicit recognition of the need to modify the efficiency metric. Therefore, it is not immediately clear what LV intend to measure in the first place, and how the constraints of LV's MILP formulations should be interpreted: do the inequality constraints of LV represent the disposability axioms of the benchmark technology or the measurement of distance to the monotonic hull of the benchmark technology? This appears to be the source of confusion for KSM, who similarly overlook the modification of the Farrell efficiency measured clearly stated in both KKM and KMK articles. We return to this point in more detail in the numerical examples considered below.

### 4.4.2 MILP Formulation

In the case of the general $T_{HIDEA}$ benchmark technology, the modified input efficiency measure $Eff^{In+}$ can be computed by solving the following MILP problem:

$$Eff^{In+}(\mathbf{x}_0, \mathbf{y}_0) = \min_{\theta, \lambda, \widetilde{\mathbf{x}}} \theta$$

subject to

$$\begin{cases} \sum_{j=1}^n x_{ij}\lambda_j \leq \widetilde{x}_i, & \forall i \in I^I \\ \widetilde{x}_i \leq \theta x_{i0}, & \forall i \in I^I \\ \widetilde{x}_i \in \mathbb{Z}_+, & \forall i \in I^I \end{cases}$$

$$\sum_{j=1}^n x_{ij}\lambda_j \leq \theta x_{i0}, \quad \forall i \in I^{NI},$$

$$\sum_{j=1}^n y_{rj}\lambda_j \geq y_{r0}, \quad \forall r \in O,$$

$$\boldsymbol{\lambda} \in \Lambda_{RTS}.$$

To clarify some key issues that continue to cause confusion in the IDEA literature, it is worth to examine the interpretation and the rationale of the constraints of the MILP problem in detail, highlighting the direct connections between the axioms introduced in Sect. 4.2 and their implementation in the MILP formulation.

For clarity, the constraints of the above MILP problem have been stated as inequalities, and the non-radial slacks have been omitted. In contrast, KKM and KMK state their MILP formulations using equality constraints and slacks.[10] This is one potential source of confusion prevailing in the IDEA literature. In particular, KSM (2013b) have criticized the MILP formulations by KKM and KMK for producing sub-optimal slacks. We find this critique misplaced because KKM and KMK were mainly interested in measuring efficiency using the radial metric $Eff^{In+}$: the slacks were used merely as instruments for imposing the free disposability and natural disposability axioms (A1), (B1), and for measuring the distance to the monotonic hull of the benchmark technology. In the case of continuous variables, we see the non-radial slacks merely as artifacts of the DEA technology, which do not necessarily have any relation to the underlying production technology: even if the true technology is smooth, the piece-wise linear DEA technology will have slacks. The presence of slacks does not imply the true technology is non-smooth. In the case of IDEA technology, the non-radial slacks may be meaningful. However, the slacks determined by the MILP problem are not necessarily unique, and not even Pareto-Koopmans efficient, as will be demonstrated below by means of numerical examples. For these reasons, we do not consider the non-radial slacks to be particularly interesting or useful.

To further clarify the MILP formulation, we use curly bracket { to identify the constraints associated with integer-valued inputs (subset $I^I$). The first constraint introduces a vector of integer-valued variables $\widetilde{\mathbf{x}} \in \mathbb{Z}_+^p$. Variables $\widetilde{\mathbf{x}}$ represent the integer-valued benchmark introduced in the definition of $Eff^{In+}$: note that the elements of $\widetilde{\mathbf{x}}$ are optimized subject to the first and the second constraint of the MILP problem. The first constraint states that the convex combination of the observed DMUs must dominate the benchmark $\widetilde{\mathbf{x}}$. Note that the inequality sign stated in the first constraint imposes the natural disposability axiom (B1). If the first inequality constraint is stated as the equality, then we effectively relax axiom (B1).

The second constraint states that the benchmark $\widetilde{\mathbf{x}}$ must dominate the radial contraction of the evaluated DMU $\theta \mathbf{x}_0$. Note that the inequality sign of the constraint is due to the fact that $Eff^{In+}$ is defined as a distance to the monotonic hull of the benchmark technology (i.e., the HIDEA technology in this case). Relaxing the natural disposability axiom (B1) does not affect this inequality constraint because the inequality represents a property of the efficiency metric, and not the benchmark technology.

The third constraint states that the benchmark $\widetilde{\mathbf{x}}$ must be integer-valued for the subset of inputs $I^I$. The fourth constraint is the standard envelopment constraint

[10] In their original manuscript, KKM stated their MILP formulation using inequality constraints. They later introduced slacks by request of a reviewer.

for outputs. Note that the distinction of continuous versus integer-valued outputs is redundant for the input-oriented efficiency index that keeps the output vector $\mathbf{y}_0$ as constant. Finally, the optional returns to scale constraints are expressed using the generic domain $\Lambda_{RTS}$ introduced above.

It is worth to note that our MILP formulation stated above differs from that of LV (2006) in one critical respect. In the original MILP formulation by LV, the envelopment constraint for the integer-valued inputs is stated as an equality: $\sum_{j=1}^{n} x_{ij}\lambda_j = \widetilde{x}_i$. KKM state that, as a result of this equality constraint, "*the intensity weights* $\lambda_j$ *need not be optimal.*" The detailed examination of the constraints of the MILP formulation discussed above allows us to pinpoint the axiomatic consequences of the LV and KKM formulations, revealing the source of the problem explicitly. Specifically, we noted above that the inequality sign in our first constraint imposes the natural disposability axiom (B1). By stating the first constraint as an equality, the LV formulation effectively relaxes the natural disposability axiom for the integer-valued inputs. Therefore, the MILP implementation by LV (2006) is not consistent with the specification of their IDEA and HIDEA technologies.

LV (2007) introduced the VRS formulation, where they correctly specify the envelopment constraint for the integer-valued inputs as an inequality $\sum_{j=1}^{n} x_{ij}\lambda_j \leq \widetilde{x}_i$, in contrast to their original CRS formulation in LV (2006). LV do not justify or explain where the inequality sign comes from in the VRS case, but instead they claim that "*In the CRS model that distinction is not necessary and the integer DEA target is always equal to the linear combination of the existing DMU.*" (LV 2007, p. 15) This claim is obviously not true. As emphasized above, the inequality sign of the envelopment constraint for integer inputs is due to the natural disposability axiom, which LV seem to ignore in their statement quoted above. Obviously, the natural disposability axiom is completely unrelated to the RTS specification. The misleading and erroneous statement by LV may be one source of confusion.

Another important difference between the VRS specifications of LV and KMK concerns the treatment of continuous inputs (i.e., the subset $I^{NI}$). KKM apply the radial contraction by factor $\theta$ to both integer-valued and continuous variables, whereas LV (2007) restrict the radial projection to the subset of integer-valued inputs, keeping continuous inputs at constant level. This is not a problem as such, it just implies a different orientation of efficiency measurement.[11] A more problematic feature of the LV (2007) formulation is the use of equality constraints for the continuous inputs and outputs, specifically,

$$\sum_{j=1}^{n} x_{ij}\lambda_j = x_{i0}, \quad \forall i \in I^{NI}$$
$$\sum_{j=1}^{n} y_{rj}\lambda_j = y_{r0} \quad \forall r \in O$$

[11] In most applications we can think of, it would seem more natural to treat integer-valued inputs as quasi-fixed factors, and project DMUs to the frontier in the direction of continuous inputs.

These constraints obviously do not allow for free disposability of the continuous inputs and outputs. However, LV (2007) do allow for free disposability of integer-valued inputs and outputs, which seems contradictory. Unfortunately, LV (2007) do not explicitly state the specific axioms imposed.

Recently, KSM (2012, 2013a, 2013b) confuse the readership further by claiming that the MILP formulations by LV and KKM are equivalent in the CRS case and that the formulations of LV and KMK are equivalent in the VRS case. In light of the observations above, these claims are obviously not true.[12] Indeed, detailed examination of the constraints of alternative MILP formulations presented in the literature clearly underlines the importance of stating the axioms explicitly and formulating the DEA problems rigorously, consistent with the maintained axioms.

### *4.4.3 Numerical Examples*

The following simple numerical example illustrates the problem in LV's (2006) MILP formulation and the line of argument presented in LV (2007), which KSM (2012, 2013a, 2013b) fail to recognize. Consider a CRS technology with two inputs and one output, and assume the input-output vector $(x_1, x_2, y)$ is integer-valued. Assume two DMUs with the following data: A = (5, 12, 3), and B = (10, 12, 2).

Figure 4.1 illustrates the boundary of the IDEA technology in the three-dimensional space. The observed DMUs are indicated by black circles labeled as A and B. In this example, the efficient subset of the IDEA technology is characterized by DMU A and any virtual units obtained by applying axioms (B1), (B2), and (B5) or any combination thereof. DMU B lies in the interior of the IDEA technology, and is hence inefficient.

Suppose we are interested in measuring efficiency of DMU B. The white circles in Fig. 4.1 indicate the benchmarks for DMU B, obtained by using the MILP formulations of KKM and LV, respectively, applying the radial input orientation and the CRS specification. Figure 4.1 indicates that the benchmarks are different. To better visualize the benchmarks, we next turn to the two-dimensional diagram of the input isoquants presented in Fig. 4.2.

Figure 4.2 illustrates the input isoquants at output levels 1, 2, and 3, and the radial projection of the evaluated DMU B to the frontier. As in Fig. 4.1, the benchmarks obtained with the KKM and LV formulations are indicated by white circles.

The KKM benchmark is obtained from DMU A using the stated axioms as follows. Firstly, we can use natural disposability axiom (B1) and add one unit of input 1 to DMU A, to obtain a feasible point A' = (6, 12, 3). Secondly, we can use natural divisibility axiom (B5) to rescale point A' downward by factor 2/3 to obtain the point A" = (4, 8, 2). Note that A" produces two units of output, similar to DMU B. Indeed,

[12] An interested reader can easily verify the empirical results reported by KKM and KMK. For transparency, the data and the computational codes for GAMS and LINGO are freely available on the website:http://nomepre.net/index.php/integerdea.

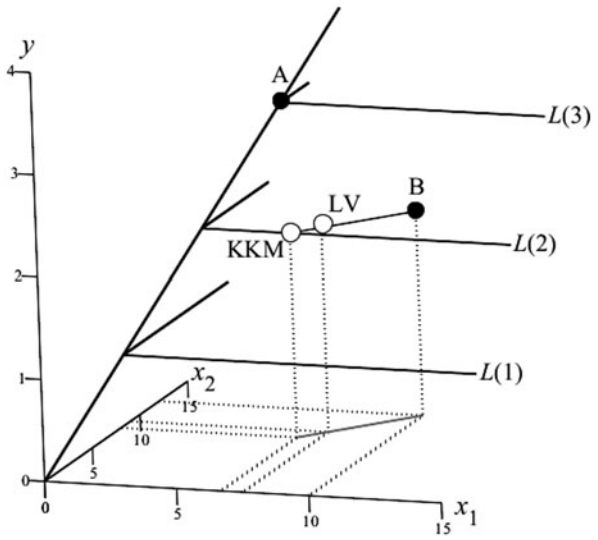



**Fig. 4.1** Three-dimensional illustration of the IDEA technology considered in the numerical example. Observed DMUs A and B are indicated by black circles. The white circles indicate the benchmarks for DMU B obtained using the KKM and LV formulations

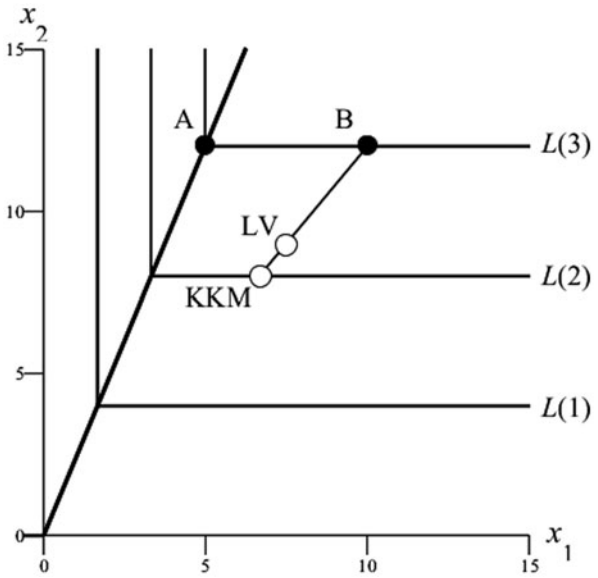



**Fig. 4.2** Two-dimensional illustration of the numerical example. Three input isoquants $L(1)$, $L(2)$, and $L(3)$ correspond to the output levels 1, 2, and 3, respectively. The line between point B and KKM indicates the radial projection of the evaluated DMU B

point A" provides a valid benchmark for DMU B. Contracting the input vector of DMU B in radial manner, we see that input 2 proves the limiting factor: the radial input efficiency of DMU B is

$$Eff_{KKM}^{In+}(10,\ 12,\ 2) = \frac{x_{2A''}}{x_{2B}} = \frac{8}{12} = \frac{2}{3}.$$

Note that there remains non-radial slack in input 1: $\frac{2}{3}x_{1B} = 6\frac{2}{3} > 4$. In addition to the radial contraction, input 1 could be further decreased by $6\frac{2}{3} - 4 = 2\frac{2}{3}$ units. Note that this input slack is due to the fact that the efficiency metric $Eff^{In+}$ measures distance to the monotonic hull of the IDEA technology: it has nothing to do with the natural disposability axiom used for obtaining the benchmark point A". This is the reason why KKM introduce two types of slack variables, and in the MILP formulation of Sect. 4.2 we use two sets of inequality constraints to ensure that

$$\sum\nolimits_{j=1}^{n} x_{ij}\lambda_j \leq \widetilde{x}_i \leq \theta x_{i0}, \quad \forall i \in I^I.$$

The essential problem of the LV formulation is that it does not include separate slacks for the natural disposability and for the monotonic efficiency metric. Hence, LV do not allow the use of natural disposability axiom (B1): we can only apply axioms (B2) and (B5) in this case. The benchmark of LV's formulation can be constructed as follows. Firstly, use the natural divisibility axiom (B5) to rescale DMU B downward by factor 1/2 to obtain the point B' = (5, 6, 1). Secondly, apply the axiom (B2) to form the convex combination of points A and B' as

$$\mathrm{B} = \frac{1}{2}\mathrm{A} + \frac{1}{2}\mathrm{B} = \frac{1}{2}(5,12,3) + \frac{1}{2}(5,6,1) = (5,9,2).$$

This convex combination provides the benchmark according to the MILP formulation of LV. Note we did not use natural disposability or non-radial slack until this point. Note further that the KKM benchmark A" dominates the LV benchmark B" in this example: $(4, 8) < (5, 9)$. Contracting the input vector of DMU B in radial manner, input 2 is the limiting factor also in the LV formulation: the radial input efficiency of DMU B is

$$Eff_{LV}^{In+}(10,\ 12,\ 2) = \frac{x_{2B''}}{x_{2B}} = \frac{9}{12} = \frac{3}{4} > Eff_{KKM}^{In+}(10,\ 12,\ 2) = \frac{2}{3}.$$

Besides the radial contraction, the LV formulation has non-radial slack in input 1, similar to the KKM case. This slack allows LV to project evaluated DMUs to the monotonic hull of the IDEA technology. However, a single slack variable is insufficient for utilizing the natural disposability axiom. In the LV formulation, the input constraints become

$$\sum\nolimits_{j=1}^{n} x_{ij}\lambda_j = \widetilde{x}_i \leq \theta x_{i0}, \quad \forall i \in I^I.$$

The use of equality constraint eliminates the natural disposability axiom.

Before proceeding to the extensions, it is worth to note that the optimal $\widetilde{x}_i$ identified by the KKM method need not be unique. Indeed, there may be multiple integer-valued $\widetilde{x}_i$ that fall within the interval characterized by the inequality constraints:

$$\sum\nolimits_{j=1}^{n} x_{ij}\lambda_j \leq \widetilde{x}_i \leq \theta x_{i0}, \quad \forall i \in I^I.$$

To illustrate the non-uniqueness in terms of the previous numerical example, note that we could equally well add four units of input 1 to DMU A (rather than just one unit), to obtain a feasible point C' = (9, 12, 3). Next, we can use natural divisibility axiom (B5) to rescale point C' downward by factor 2/3 to obtain the point C" = (6, 8, 2). Although point A" considered above dominates C", point C" provides an equally valid reference point for assessing radial input efficiency of DMU B. Contracting the input vector of DMU B radially, input 2 remains the limiting factor, and the radial input efficiency of DMU B is $\frac{2}{3}$ even if we use C" as the benchmark. However, the second non-radial slack in input 1 is now $6\frac{2}{3} - -6 = \frac{2}{3}$. This example illustrates that the integer programming algorithm applied for solving the MILP problem may well return sub-optimal target points, as KSM (2013b) have noted. We must stress there is no guarantee that the optimal intensity weights, multiplier weights, or slacks are unique even in the standard DEA formulations. In the case of discrete inputs and outputs, it should be nothing surprising to find non-unique slacks and non-unique targets. To conclude, we emphasize that the MILP formulations presented by KKM and KMK were developed for measuring radial input efficiency, and can only be guaranteed to serve that purpose. Since the non-radial slacks obtained as the optimal solution to the MILP problem are not necessarily unique, adjusting the radial projection for the non-radial slacks may result as sub-optimal target points. We return to this issue in Sect. 4.5.3 below.

## 4.5 Alternative Efficiency Metrics

The sound axiomatic foundation of IDEA technology based on the minimum extrapolation principle makes several extensions of the conventional DEA readily available to IDEA. LV (2007) consider several alternative efficiency metrics, including the input and output oriented radial Farrell measures, additive and range-adjusted slack based measures, and the Russell measure. Du et al. (2012) consider the additive super-efficiency measure in the context of DEA. In this section we review some alternative efficiency measures, starting from the radial output oriented efficiency measure, and proceeding to the general directional distance function. We complete this section with a critical review of additive and slack based measures, noting some problems in these approaches in the context of IDEA technology.

### *4.5.1 Modified Farrell Output Efficiency Measure and its Implementation*

In Sect. 4.4 we restricted attention to the radial input-oriented efficiency measure by Farrell (1957), modified by KKM to the IDEA context. In this section we briefly extend the discussion to the radial output measure.

Farrell's output efficiency measure is defined as

$$Eff^{Out}(\mathbf{x}_0, \mathbf{y}_0) = \max\{\gamma | (\mathbf{x}_0, \gamma \mathbf{y}_0) \in T\}.$$

Note that in this case $\gamma = 1$ indicates full efficiency, and values $\gamma > 1$ indicate that the evaluated DMU is inefficient (note that we can convert the output efficiency measures to the interval (0, 1] by using the inverse $\gamma^{-1}$). To avoid complications due the discrete nature of IDEA and HIDEA technologies, we can modify the radial output efficiency measure as:

$$Eff^{Out+}(\mathbf{x}_0, \mathbf{y}_0) = \max\left\{\gamma \in \mathbb{R}_+ | \exists (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T \ : \widetilde{\mathbf{y}}^I \in \mathbb{Z}_+^q \ ; \mathbf{x}_0 \geq \widetilde{\mathbf{x}}; \gamma \mathbf{y}_0 \leq \widetilde{\mathbf{y}}\right\}.$$

This modified Farrell measure gauges radial distance to the monotonic hull of the benchmark technology, requiring that the reference point $(\mathbf{x}_0, \gamma \mathbf{y}_0)$ must be dominated by a feasible input-output vector $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ which has integer-valued inputs for the subset $O^I$.

The modified output efficiency measure $Eff^{Out+}$ has the usual interpretation of the radial expansion potential of the evaluated output vector at the given level of inputs. It guarantees that DMUs assigned the efficiency score one are weakly efficient in the Pareto-Koopmans sense.

In the case of the $T_{HIDEA}$ benchmark technology, the modified output efficiency measure $Eff^{Out+}$ can be computed by solving the following MILP problem:

$$Eff^{Out+}(\mathbf{x}_0, \mathbf{y}_0) = \max_{\gamma, \lambda, \widetilde{\mathbf{y}}} \gamma$$

subject to

$$\begin{cases} \sum_{j=1}^n y_{rj}\lambda_j \geq \widetilde{y}_r, & \forall r \in O^I \\ \widetilde{y}_r \geq \gamma y_{r0}, & \forall r \in O^I \\ \widetilde{y}_r \in \mathbb{Z}_+, & \forall r \in O^I \end{cases}$$

$$\sum_{j=1}^n y_{rj}\lambda_j \geq \gamma y_{r0}, \quad \forall r \in O^{NI},$$

$$\sum_{j=1}^n x_{ij}\lambda_j \leq x_{i0}, \quad \forall i \in I,$$

$$\lambda \in \Lambda_{RTS}.$$

In this case, we indicate the constraints of the integer-valued outputs (subset $O^I$) by curly bracket {. Note that it is unnecessary to introduce integer-valued input targets $\widetilde{\mathbf{x}}^I \in \mathbb{Z}_+^p$ because the inputs are held constant at their observed levels. This reduces computational complexity compared with the MILP formulation by LV (2007) because our formulation excludes $p$ integer-valued model variables as redundant (here $p$ is the number of input factors). Note further that we set two inequality constraints for the integer-valued outputs to ensure that

$$\sum_{j=1}^n y_{rj}\lambda_j \geq \widetilde{y}_i \geq \gamma y_{r0}, \quad \forall r \in O^I.$$

The first inequality imposes the natural disposability axiom for the integer-valued outputs, whereas the latter inequality is due to the fact that we measure distance to the monotonic hull of the discrete HIDEA benchmark technology.

### 4.5.2 *Modified Directional Distance Function and its Implementation*

We next consider a modified version of the directional distance function (DDF) by Chambers et al. (1996, 1998). To our knowledge, this is the first application of DDF to the IDEA context.

DDF allows us to project the observed DMUs to the frontier in non-radial manner, allowing for simultaneous contraction of inputs and expansion of outputs. DDF indicates the distance from a given input-output vector to the boundary of the benchmark technology in some pre-assigned direction $(\mathbf{g}_x, \mathbf{g}_y) \in \mathbb{R}_+^{m+s}$. DDF can be formally defined as

$$DDF(\mathbf{x}_0, \mathbf{y}_0, \mathbf{g}_x, \mathbf{g}_y) = \sup\left\{\delta | (\mathbf{x}_0 - \delta \mathbf{g}_x, \mathbf{y}_0 + \delta \mathbf{g}_y) \in T\right\}.$$

Note that in this case $\delta = 0$ indicates full efficiency, and values $\delta > 0$ indicate that the evaluated DMU is inefficient. Note further that DDF contains the radial input and output efficiency measures as its special cases. For example, setting $(\mathbf{g}_x, \mathbf{g}_y) = (\mathbf{0}, \mathbf{y}_0)$, we obtain

$$DDF(\mathbf{x}_0, \mathbf{y}_0, \mathbf{g}_x, \mathbf{g}_y) = 1 - \ Eff^{Out+}(\mathbf{x}_0, \mathbf{y}_0).$$

To avoid complications due the discrete nature of IDEA and HIDEA technologies, we can modify the original DDF as:

$$DDF^{+}(\mathbf{x}_0, \mathbf{y}_0, \mathbf{g}_x, \mathbf{g}_y) =$$
$$\sup\left\{\delta | \exists (\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T \ : \widetilde{\mathbf{x}}^I \in \mathbb{Z}_+^p; \widetilde{\mathbf{y}}^I \in \mathbb{Z}_+^q \ ; (\mathbf{x}_0 - \delta \mathbf{g}_\mathbf{x}) \geq \widetilde{\mathbf{x}}; (\mathbf{y}_0 + \delta \mathbf{g}_y) \leq \widetilde{\mathbf{y}}\right\}.$$

This modified DDF gauges directional distance to the monotonic hull of the benchmark technology, requiring that the reference point $(\mathbf{x}_0 - \delta \mathbf{g}_x, \mathbf{y}_0 + \delta \mathbf{g}_y)$ must be dominated by a feasible input-output vector $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ which has integer-valued inputs and outputs for the subsets $I^I$ and $O^I$.

In the case of the $T_{HIDEA}$ benchmark technology, the modified DDF can be computed by solving the following MILP problem:

$$DDF^{+}(\mathbf{x}_0, \mathbf{y}_0, \mathbf{g}_x, \mathbf{g}_y) = \max_{\delta, \lambda, \widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}} \delta$$

subject to

$$\begin{cases} \sum_{j=1}^{n} x_{ij} \lambda_j \leq \widetilde{x}_i, & \forall i \in I^I \\ \widetilde{x}_i \leq x_{i0} - \delta g_{xi}, & \forall i \in I^I \\ \widetilde{x}_i \in \mathbb{Z}_+, & \forall i \in I^I \end{cases}$$

$$\begin{cases} \sum_{j=1}^{n} y_{rj}\lambda_j \geq \widetilde{y}_r, & \forall r \in O^I \\ \widetilde{y}_i \geq y_{r0} + \delta g_{yr}, & \forall r \in O^I \\ \widetilde{y}_i \in \mathbb{Z}_+, & \forall r \in O^I \end{cases}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{r0} + \delta g_{yr}, \quad \forall r \in O^{NI},$$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \leq x_{i0} - \delta g_{xi}, \quad \forall i \in I^{NI}.$$

$$\boldsymbol{\lambda} \in \Lambda_{RTS}.$$

In general, DDF requires that we introduce integer-valued targets for both inputs (i.e.,) $\widetilde{\mathbf{x}}^I \in \mathbb{Z}_+^p$) and outputs ($\widetilde{\mathbf{y}}_i \in \mathbb{Z}_+^q$) because DDF can adjust all inputs and outputs simultaneously. However, if the direction vector ($\mathbf{g_x}, \mathbf{g_y}$) contains any zero elements, we can harmlessly exclude the integer-valued targets for the corresponding inputs and outputs, and treat those inputs and outputs as fixed factors, similar to the treatment of outputs in the radial input efficiency measure considered in Sect. 4.4.2, and the treatment of inputs in the radial input efficiency measure considered in Sect. 4.5.1.

### *4.5.3 Additive and Slack Based Measures*

LV (2007) introduced the additive IDEA formulation, applying the Pareto-Koopmans measure by Charnes et al. (1985) to the IDEA technology. A slightly modified version of LV's additive formulation can be presented as follows:

$$\max_{\mathbf{s}^+,\mathbf{s}^-,\boldsymbol{\lambda},\widetilde{\mathbf{x}},\widetilde{\mathbf{y}}} \sum_{i\in I^I} s_i^- + \sum_{r\in O^I} s_r^+$$

subject to

$$\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \leq \widetilde{x}_i, & \forall i \in I^I \\ \widetilde{x}_i \leq x_{i0} - s_i^-, & \forall i \in I^I \\ \widetilde{x}_i \in \mathbb{Z}_+, & \forall i \in I^I \end{cases}$$

$$\begin{cases} \sum_{j=1}^{n} y_{rj}\lambda_j \geq \widetilde{y}_r, & \forall r \in O^I \\ \widetilde{y}_i \geq y_{r0} + s_r^+, & \forall r \in O^I \\ \widetilde{y}_i \in \mathbb{Z}_+, & \forall r \in O^I \end{cases}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \geq y_{r0}, \quad \forall r \in O^{NI},$$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \leq x_{i0}, \quad \forall i \in I^{NI}.$$

$$\boldsymbol{\lambda} \in \Lambda_{RTS}.$$

Following LV, our MILP formulation minimizes the sum of slacks in the integer-valued inputs and outputs (subsets $O^I$ and $I^I$, respectively), keeping the continuous inputs and outputs at constant level. We could easily introduce slacks to the continuous inputs and outputs as well (see Du et al. 2012).

Our MILP formulation of the additive IDEA differs from that of LV (2007) in that we use inequality constraints for the continuous inputs and outputs (subsets $O^{NI}$ and $I^{NI}$, respectively), allowing for free disposability of these inputs and outputs. In contrast, LV do not allow for free disposability of continuous inputs and outputs, but they do implicitly assume free disposability of integer-valued inputs and outputs, which may seem confusing.

We noted at the end of Sect. 4.4 that the optimal solution of the KKM and KMK MILP formulations for computing the modified Farrell input efficiency may yield sub-optimal benchmarks. Specifically, the optimal integer-valued reference point $\widetilde{x}$ need not be unique, and it is possible that $\widetilde{\mathbf{x}}$ is dominated by another feasible point. If one is interested in computing efficient benchmarks, one can first compute the radial or directional projection to the IDEA frontier using the MILP formulations presented in Sects. 4.4, 4.5.1, or 4.5.3, and subsequently apply the additive MILP formulation presented in this section to maximize the sum of slacks. While the additive formulation ensures benchmarks that are efficient in the Pareto-Koopmans sense, a unique solution cannot be guaranteed.

Consider a simple example of three DMUs that use a single integer-valued input $x$ to produce a single integer-valued output $y$. Suppose the observed data of DMUs, presented as vectors $(x, y)$ is the following: A = (2,1), B = (3,2), and C = (3,1). Note that A and B lie on the efficient boundary of the IDEA technology, whereas C is dominated by both A and B. Now, apply the additive IDEA formulation to assess efficiency of DMU C. The optimal value of the objective function is unique, equal to 1. However, neither the benchmark $(\widetilde{x}, \widetilde{y})$ nor the slacks $(s^-, s^+)$ are unique. It is possible to identify DMU A as the benchmark (i.e., $(\widetilde{x}, \widetilde{y}) = (2,1)$), which yields $(s^-, s^+) = (1,0)$. It is equally possible to identify DMU B as the benchmark (i.e., $(\widetilde{x}, \widetilde{y}) = (3,2)$), which yields $(s^-, s^+) = (0,1)$. The MILP algorithm will arbitrarily identify one of these two alternatives to be presented as the optimal solution. While even standard DEA does not guarantee a unique optimum for the slacks, benchmarks, or multiplier weights, alternate optima are likely to occur in the context of integer-valued inputs and outputs. Therefore, it is important to be aware of the fact that the optimal slacks need not be unique. It seems some of the critiques by KSM are based on misunderstanding this fact.

The additive measure can be used for testing whether the evaluated DMU is on the Pareto-Koopmans efficient frontier (i.e., $\sum_{i \in I^I} s_i^- + \sum_{r \in O^I} s_r^+ = 0$) or not. However, the use of the additive measure for gauging efficiency is problematic. Non-uniqueness of the optimal slacks noted above is not the only problem. LV (2007) note that the interpretation of the additive measure as an efficiency index is meaningful only when the inputs and outputs are measured in the same units of measurements (e.g., in money), which is not usually the case. Indeed, an appealing feature of DEA is that inputs and outputs can be measured in different units without a need to convert them to money metric or other measure of relative values prior to the analysis.

Several attempts to adjust the additive measure to different units of measurement have been presented in the DEA literature, most notably the range adjusted measure (RAM) by Cooper et al. (1999) and the slack-based model (SBM) by Tone (2001). In RAM formulation, the objective function of the additive IDEA formulation is replaced by

$$\max_{\mathbf{s}^+,\mathbf{s}^-,\boldsymbol{\lambda},\tilde{\mathbf{x}},\tilde{\mathbf{y}}} \sum_{i\in I^I} \frac{s_i^-}{R_i} + \sum_{r\in O^I} \frac{s_r^+}{R_r},$$

where $R_i = \max_j x_{ij} - \min_j x_{ij}$ is the observed range of input $i$, and $R_r = \max_j y_{rj} - -\min_j y_{rj}$ is the observed range of output $r$, respectively. Many variants of SBM (Tone 2001) are known in the DEA literature. In SBM we first compute the additive MILP formulation, or its range adjusted variant. The main idea of SBM is to aggregate thus obtained slacks to a single efficiency metric. Given the additive IDEA formulation presented above, the SBM measure can be stated as

$$SBM = \frac{1 - \frac{1}{p}\left(\sum_{i\in I^I} \frac{s_i^-}{x_{i0}}\right)}{1 + \frac{1}{q}\left(\sum_{r\in O^I} \frac{s_i^+}{y_{i0}}\right)}.$$

These examples consider slacks in the integer-valued inputs and outputs (similar to LV 2007), but one could equally well include slacks to continuous inputs and outputs as well.

In the context of IDEA technology, the fact that the optimal slacks $(\mathbf{s}^-, \mathbf{s}^+)$ are not necessarily unique can be problematic. Reconsider the numerical example with three DMUs A, B, and C, and consider SBM efficiency of DMU C. If the MILP algorithm identifies DMU A as the benchmark, then

$$SBM = (1 \ - 1/3)/\,(1 \ + 0) = 2/3.$$

However, the MILP algorithm can equally well identify DMU B as the benchmark, resulting with

$$SBM = (1 \ - 0)/\,(1 \ + 1) = 1/2.$$

This example illustrate that the SBM measure is not invariant or robust to alternate optimal of $(\mathbf{s}^-, \mathbf{s}^+)$, and indeed, there is no guarantee that the optimal slacks are unique. To avoid this problem, one could enumerate the SBM measure for all alternate optima, and choose the slacks that maximize or minimize the SBM measure. However, identifying all alternate optima of $(\mathbf{s}^-, \mathbf{s}^+)$ seems challenging if not computationally prohibitive in practice. To our knowledge, non-uniqueness of slacks and its potential problems have not been duly addressed in the DEA literature. In our view, non-uniqueness of slacks in DEA is one rational argument for why the radial or directional distance functions are preferred over the slack based approaches.

We conclude this section by noting that the numerical examples used in this section for illustrating the non-uniqueness problem may seem overly simplistic. We deliberately used the simplest thinkable examples to illustrate. If non-uniqueness can occur and cause problems in a simple example, it would be foolish to assume the problem disappears as one proceeds to more complex examples or real applications.

## 4.6 Stochastic Noise

In Sect. 4.2.4 we examined the envelopment condition (E1), noting that the best observed performance level may not be achievable by all DMUs due to unobserved heterogeneity of DMUs and their operating environments, technological and economic risks and uncertainty, omitted factors such as quality differences, errors in measurement and data processing, and other sources of noise. In this section we will briefly extend the StoNED framework introduced by Kuosmanen and Kortelainen (2012) to the present context of integer valued inputs and outputs.

To maintain direct contact with the conventional stochastic frontier analysis (SFA) and StoNED, we consider the single-output case, and model the production technology using the production function $f(\mathbf{x})$, which indicates the maximum output that can be produced with input vector $\mathbf{x}$ (for a general multi-output model, see Sect. 7.4.6.3 in KJS). Thus, the production possibility set $T$ can be stated as

$$T = \left\{(\mathbf{x}, y) \in \mathbb{R}_+^{m+1} | y \leq f(\mathbf{x}) \right\}.$$

We do not impose any particular functional form for $f$: we only assume the production possibility set $T$ satisfies axiom (A1) for continuous inputs and (B1) for integer-valued inputs, axiom (B2), and possibly some RTS axioms. Inputs $\mathbf{x}$ can be integer-valued or continuous. The main challenge in this setting concerns the modeling integer-valued output $y \in \mathbb{Z}_+$. To our knowledge, all previous studies on stochastic frontier estimation in the single-output case assume a continuous output variable.

To model stochastic noise explicitly, the following data generating process will be assumed. The observed outputs of DMUs $i = 1, \ldots, n$, denoted as $y_i$, are assumed to be generated from equation

$$y_i = f(\mathbf{x}_i) - u_i + v_i,$$

Where $\mathbf{x}_i$ is the input vector of DMU $i$ (which may contain both discrete or continuous inputs), $u_i$ is a random inefficiency term, and $v_i$ is a random noise term. Random variables $u_i$ and $v_i$ are assumed to be independent of inputs $\mathbf{x}$ and of each other. More specific assumptions regarding $u_i$ and $v_i$ are the following.

The inefficiency term $u_i$ is assumed to be a discrete, Poisson distributed random variable:[13]

$$u_i \sim Pois(\lambda_u),$$

where parameter $\lambda_u = E(u_i) = Var(u_i)$ characterizes both the expected value and variance of the random inefficiency term (note: $\lambda_u$ should not be confused with the

[13] The Poisson distribution is the most widely used discrete probability distribution in statistics. It can be derived from the probability of a given number of events occurring in a fixed interval of time and/or space when the events occur with a known average rate and independently of the time since the last event. Note that the Poisson distribution can be derived as a limiting case to the binomial distribution as the number of trials approaches to infinity and the expected number of successes is fixed.

intensity weights of DEA). In this model, inefficiency $u_i$ is always a non-negative integer, with a known probability mass function

$$\Pr(u_i = k) = \frac{\lambda_u^k e^{-\lambda}}{k!}.$$

Note that a DMU is fully efficient with probability $\Pr(u_i = 0) = e^{-\lambda_u}$.

The noise term $v_i$ is specified as

$$v_i = \widetilde{v}_i - \lfloor \lambda_v \rfloor,$$

where

$$\widetilde{v}_i \sim Pois(\lambda_v),$$

and $\lfloor \lambda_v \rfloor$ denotes the largest integer less than or equal to $\lambda_v$. Parameter $\lambda_v = E(\widetilde{v}_i) = Var(\widetilde{v}_i)$ characterizes both the expected value and variance of the random variable $\widetilde{v}_i$, while $\lfloor \lambda_v \rfloor$ is the mode of $\widetilde{v}_i$. Note that while random variable $\widetilde{v}_i$ is always non-negative, the noise term $v_i$ has zero mode and it can take either positive and negative values. As parameter $\lambda_v$ increases, the noise term $v_i$ approaches to the normal distribution with zero mean and variance. Note that in this model the impact of noise term has the lower bound $\lfloor -\lambda_v \rfloor$ .

To estimate the frontier production function $f$ and the parameters $\lambda_u$ and $\lambda_v$, we can modify the stepwise StoNED estimator developed by Kuosmanen and Kortelainen (2012) as follows. In the first step, we estimate conditional mean output, which can be written as

$$E(y_i|\mathbf{x}_i) = f(\mathbf{x}_i) - \lambda_u + \lambda_v - \lfloor \lambda_v \rfloor = g(\mathbf{x}_i).$$

Note that function $g$ differs from $f$ only by constant $-\lambda_u + \lambda_v - \lfloor \lambda_v \rfloor$ . Note further that even though the observed outputs $y_i$ are assumed to be integer valued, the conditional mean $E(y_i|\mathbf{x}_i)$ does not need to be an integer. Therefore, convex nonparametric least squares (CNLS) provides an unbiased and consistent estimator of function $g(\mathbf{x}_i)$. Kuosmanen (2008) shows that the CNLS estimator can be computed by solving the following quadratic programming problem

$$\min \sum_{i=1}^{n} \varepsilon_i^2$$

Subject to

$$\begin{aligned} y_i =& \alpha_i + \boldsymbol{\beta}_i^{'} \mathbf{x}_i + \varepsilon_i \quad i = 1, \ldots, n, \\ & \alpha_i + \boldsymbol{\beta}_i^{'} \mathbf{x}_i \leq \alpha_j + \boldsymbol{\beta}_j^{'} \mathbf{x}_i \quad i, j = 1, \ldots, n, \\ & \boldsymbol{\beta}_i \geq \mathbf{0} \quad i = 1, \ldots, n \end{aligned}$$

Where $\varepsilon_i$ are the CNLS residuals that represent the deviations of observed DMUs from the conditional mean function $g(\mathbf{x}^i)$, and $\boldsymbol{\beta}_i$ are vectors of nonnegative slope

coefficients that together with intercepts $\alpha_i$ characterize a supporting hyper plane of the unknown concave function to be estimated in point $\mathbf{x}_i$.[14] See KJS, Sect. 7.4.3, for a more detailed exploration of the CNLS formulation, its interpretation, and computation.

Having solved the CNLS problem, we can estimate the conditional mean function $g(\mathbf{x}_i)$ in the observed data points by

$$\widehat{g}\,(\mathbf{x}_i) = \widehat{\alpha}_i + \widehat{\boldsymbol{\beta}}_i'\mathbf{x}_i.$$

Further, we have the CNLS residuals $\widehat{\varepsilon}_i$ that are nonparametric estimators of

$$(v_i - \lambda_v + \lambda_v) - (u_i - \lambda_u) = (\widetilde{v}_i - \lambda_v) - (u_i - \lambda_u) = (\widetilde{v}_i - u_i) - (\lambda_v - \lambda_u).$$

To estimate the parameters $\lambda_u$ and $\lambda_v$, we can utilize the CNLS residuals and the assumption of Poisson distributed inefficiency and noise.

Before proceeding to step two, consider random variable $\widetilde{\varepsilon}_i = \widetilde{v}_i - u_i$. Since $\widetilde{\varepsilon}_i$ is a difference of two independent Poisson distributed random variables, it follows the Skellam distribution (Skellam 1946). The mean, variance and skewness of the Skellam distributed random variable are related to the central moments of the distribution as follows. Define

$$\begin{aligned} \Delta &= \lambda_v - \lambda_u, \quad \text{and} \\ \mu &= (\lambda_v + \lambda_u)/2. \end{aligned}$$

Using these notations, the variance and skewness of $\widetilde{\varepsilon}_i$ can be stated as

$$\begin{aligned} Var(\widetilde{\varepsilon}_i) &= 2\mu, \\ Skew(\widetilde{\varepsilon}_i) &= \Delta/(2\mu)^{3/2}. \end{aligned}$$

Note that the CNLS residuals are consistent estimators of $\widetilde{\varepsilon}_i$ minus a constant. Therefore, we can use the sample variance and skewness of CNLS residuals as estimators of $Var(\widetilde{\varepsilon}_i)$ and $Skew(\widetilde{\varepsilon}_i)$, to obtain estimates $\widehat{\Delta}$ and $\widehat{\mu}$.

Step 2 of the StoNED estimation is the following. Using the above moment equations, we obtain the following estimators for parameters $\lambda_u$ and $\lambda_v$:

$$\begin{aligned} \widehat{\lambda}_u &= \widehat{\mu} - \frac{1}{2}\widehat{\Delta} = \frac{1}{2}(Var(\varepsilon_i) - Skew(\varepsilon_i)(Var(\varepsilon_i))^{3/2}), \\ \widehat{\lambda}_v &= \widehat{\mu} + \frac{1}{2}\widehat{\Delta} = \frac{1}{2}(Var(\varepsilon_i) + Skew(\varepsilon_i)(Var(\varepsilon_i))^{3/2}), \end{aligned}$$

Where $Var(\varepsilon_i)$ and $Skew(\varepsilon_i)$ are the sample variance and skewness of the CNLS residuals, respectively. Using the parameter estimates $\widehat{\lambda}_u$ and $\widehat{\lambda}_v$, we can estimate the

[14] The second set of constraints imposes convexity, applying the Afriat theorem (Afriat 1972). The convexity axiom can be relaxed by replacing CNLS with isotonic regression, see Keshvari and Kuosmanen (2013), for details.

probability distributions of inefficiency and noise terms. Recall that expected value of inefficiency is simply $\lambda_u$, and hence we can use $\widehat{\lambda}_u$ directly as the estimator of mean inefficiency. Note that in the stochastic frontier model $Skew(\varepsilon_i)$ is generally expected to be negative. Therefore, negative skewness of residuals increases the mean of the inefficiency term relative to that of the noise term in the Poisson model. If skewness is zero, then $\widehat{\lambda}_u = \widehat{\lambda}_v$. Positive skewness is also allowed: "wrong skewness" increases the mean of the noise term compared to that of the inefficiency term. This is an attractive feature of the Poisson model and the proposed method of moments estimator: wrong skewness does not cause major problems in this framework.[15]

In step 3 we adjust the CNLS estimate of the conditional mean $\widehat{g}(\mathbf{x}_i)$ to estimate the frontier. Note that we need to shift the CNLS estimator upward by the mean inefficiency, but in this case, also the noise term may have non-zero mean (recall we assumed $v_i$ has zero mode, which does not imply zero mean). Further, we need to take into account that values of the production function must be integers. Using the equation of the conditional mean $E(y_i|\mathbf{x}_i)$, the integer-valued StoNED frontier estimator can be stated as

$$\widehat{f}(\mathbf{x}_i) = \left\lfloor \widehat{g}(\mathbf{x}_i) + \widehat{\lambda}_u - \widehat{\lambda}_v + \left\lfloor \widehat{\lambda}_v \right\rfloor \right\rfloor,$$

where symbol $\lfloor a \rfloor$ is denotes the largest integer less than or equal to $a$. Function $\widehat{f}$ can be proved to satisfy the axioms of natural convexity, natural disposability of output and integer-valued inputs, free disposability of continuous inputs, and any RTS axioms postulated. Function $\widehat{f}$ does not necessarily envelope all observed DMUs, and hence the StoNED frontier will typically lie below the corresponding IDEA frontier. Note that enveloping noisy data will generally result as biased and inconsistent estimates. Provided that the assumed doubly-Poisson model of inefficiency and noise is correctly specified, the StoNED estimator $\widehat{f}$ described above can be shown to be statistically consistent.

To obtain DMU specific efficiency estimates, we must first recognize that the observed departures from the estimated frontier, that is, $y_i - \widehat{f}(\mathbf{x}_i)$ or $y_i / \widehat{f}(\mathbf{x}_i)$, cannot be used directly for measuring efficiency. We can write the observed distance from the estimated frontier as

$$y_i - \widehat{f}(\mathbf{x}_i) = (f(\mathbf{x}_i) - u_i + v_i) - \widehat{f}(\mathbf{x}_i) = (f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i)) - u_i + v_i.$$

Even if our estimate is precise, that is $f(\mathbf{x}_i) - \widehat{f}(\mathbf{x}_i) = 0$, the distance to the estimated frontier consists of two components: inefficiency and noise. To make DMU specific efficiency assessments, we need the conditional distribution of $u_i$ for a given level of $y_i - \widehat{f}(\mathbf{x}_i)$, analogous to Jondrow et al. (1982).

In the discrete case of two Poisson distributed random variables, deriving the conditional distribution of $u_i$ for given $y_i - \widehat{f}(\mathbf{x}_i)$ is relatively straightforward. Firstly, note that we can calculate the unconditional probabilities $\Pr(u_i = k)$ for each $k =$

[15] See, e.g., Simar and Wilson (2010) for a more detailed discussion about the wrong skewness problem in stochastic frontier estimation.

$0,1,\ldots,K$, where $k$ denotes the index of possible values of $u_i$, and $K$ is the smallest integer that satisfies $\Pr(u_i = K) < \ddot{\varepsilon}$ for some pre-specified threshold probability $\ddot{\varepsilon}$ (e.g., we can set $\ddot{\varepsilon} = 10^6$). Secondly, we know that if $u_i = k$, then the noise term must be equal to $v_i = y_i - \widehat{f}(\mathbf{x}_i) + k$. Hence, we can calculate the unconditional probabilities $\Pr(v_i = y_i - \widehat{f}(\mathbf{x}_i) + k)$ associated with each $k = 0,1,\ldots,K$. Note that if $y_i - \widehat{f}(\mathbf{x}_i) + k < -\left\lfloor \widehat{\lambda}_v \right\rfloor$, then the value of $k$ falls below the minimum bound of noise $v_i$, and hence we need to set $\Pr(v_i = y_i - \widehat{f}(\mathbf{x}_i) + k) = 0$ in such cases.

Having calculated the unconditional probability distributions of $u_i$ and $v_i$ for $k = 0,1,\ldots,K$, we calculate the sum product

$$\tau_i = \sum\nolimits_{k=0}^{y_i - \widehat{f}(\mathbf{x}_i) - \widehat{\lambda}_v} \Pr(u_i = k) \times \Pr(v_i = y_i - \widehat{f}(\mathbf{x}_i) + k)$$

The conditional distribution of $u_i$ for given $y_i - \widehat{f}(\mathbf{x}_i)$ is then obtained as

$$\Pr(u_i = k | y_i - \widehat{f}(\mathbf{x}_i)) = \frac{\Pr(u_i = k) \times \Pr(v_i = y_i - \widehat{f}(\mathbf{x}_i) + k)}{\tau_i}.$$

As a point estimator of $u_i$, one could use the mean of the conditional distribution

$$E(u_i | y_i - \widehat{f}(\mathbf{x}_i)) = \sum\nolimits_{k=0}^{K} \Pr(u_i = k | y_i - \widehat{f}(\mathbf{x}_i)) \times k,$$

following the common practice in the SFA literature. Another possibility is to use the median of the conditional distribution. However, whichever point estimator might be used, it is important to keep in mind that $u_i$ is essentially a random variable, and hence point estimation of a single realization of this random variable may be a pointless exercise. We emphasize that the knowledge of the conditional distribution of $u_i$ at given $y_i - \widehat{f}(\mathbf{x}_i)$ provides means for more useful statistical inferences beyond computing point estimates for efficiency rankings. For example, one could apply the conditional distributions for assessing the probability that DMU $i$ is more efficient than another DMU $j$, or the probability that a group of DMUs is more efficient than another group.

While the double-Poisson model and the associated StoNED estimator appear to be well suited for estimating IDEA technology under noise, one important caveat is worth noting. We assumed the observed outputs to be non-negative integers, and we would typically assume some observed $y_i$ to be small, as large integers can be reasonably approximated as continuous variables. While took the lower bound $y_i \geq 0$ explicitly into account in the conditional distribution of $u_i$, we assumed parameters $\lambda_u$ and $\lambda_v$ to be constant at all input levels. This is not necessarily a realistic assumption as the range of possible output values typically depends on the input levels, and hence the variances represented by parameters $\lambda_u$ and $\lambda_v$ are not constant. Therefore, it would be important to take heteroscedasticity of inefficiency and noise explicitly into account by modeling these parameters explicitly as functions of inputs, that is, $\lambda_u(\mathbf{x})$ and $\lambda_v(\mathbf{x})$. However, we need to walk before we can run. We leave explicit

modeling of heteroscedasticity as an interesting topic for future research, noting that there exists extensive econometric literature on this topic.[16]

## 4.7 Conclusion and Directions for Future Research

The main insights of this chapter can be classified in three categories. First, a detailed examination of the axioms of integer DEA and the associated MILP formulations was presented in order to clarify some points of confusion prevailing in the literature. The key insight gained through this analysis is the intimate connection between the axioms and the formulation of the MILP problem. Without a proper understanding of explicitly stated axioms, the MILP formulation will likely produce erroneous or misleading results. For example, we demonstrated that LV's MILP formulations fail to satisfy such axioms as free disposability of continuous inputs and outputs, and natural divisibility of discrete inputs and outputs. We illustrated through simple numerical examples that the MILP formulations by LV and KKM yield different results, in contrast to what KSM have recently claimed. The numerical examples also explain how the differences arise from the inconsistency of LV's MILP formulations with their definition of IDEA technology. These observations underline the critical importance of the sound axiomatic foundation.

Second, we examined alternative efficiency metrics available for integer DEA, complementing the KKM and KMK formulations for the modified radial input oriented measure with the modified versions of the radial output oriented measure and the directional distance function. We also critically discussed the additive efficiency measures, demonstrating by simple numerical examples that the optimal slacks are not necessarily unique. The non-uniqueness of slacks can be particularly problematic for the slack based measures of efficiency in the context of integer DEA.

Third, we introduced a new model of IDEA technology in the single output setting under stochastic noise. Modeling both inefficiency and noise as Poisson distributed random variables, we developed the first extension of the StoNED method to the discrete setting. We developed the method of moments estimator for identifying the parameters of the double-Poisson model, and discussed how the conditional distribution of inefficiency at the given distance from the estimated frontier can be computed and applied for statistical inferences.

In conclusion, we hope that this chapter helps to clarify some issues that have caused confusion, but also identify some interesting avenues for future research. The basic axioms of DEA are already well understood in the context of IDEA, but there are other axioms such as weak disposability (e.g., Kuosmanen 2005) and selective proportionality (Podinovski 2004) that deserve to be examined in the context of IDEA technology. For real applications, probabilistic modeling of noisy data appears to be the main challenge. In this chapter we presented the first attempt to

[16] KJS, Sect. 7.4.8, discusses some of this literature in the context of StoNED estimation.

modeling stochastic noise in the discrete setting assuming Poisson distributed noise. Further work is obviously needed to operationalize these ideas to be applicable to real applications. For example, the truncated distribution of observed outputs above zero and the associated heteroskedasticity deserve to be addressed explicitly.

## 4.8 Appendix: Proofs of theorems and lemmas

**Lemma 1** *Assume Axiom (B2) is satisfied. Then for any given* $(\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in T$, *if there exists a real valued* $\lambda$ *such that* $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = \lambda(\mathbf{x}, \mathbf{y}) + (1 - \lambda)(\mathbf{x}', \mathbf{y}') \in T$, *then there exist integers* $u, v \in \mathbb{Z}_+, \quad u \leq v$, *such that*

$$\lambda = u/v.$$

Proof.

We can write $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ equivalently as $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) = (\mathbf{x}', \mathbf{y}') + \lambda(\mathbf{x} - \mathbf{x}', \mathbf{y} - \mathbf{y}')$. The first term $(\mathbf{x}', \mathbf{y}')$ is an integer-valued vector by assumption, and the second term is the product of another vector of integers $(\mathbf{x} - \mathbf{x}', \mathbf{y} - \mathbf{y}')$ and multiplier $\lambda$. Since $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in T$ implies $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}}) \in \mathbb{Z}_+^{m+s}$, then obviously $\lambda$ cannot be an irrational number. Therefore, there must exist integers $u, v \in \mathbb{Z}_+, \quad u \leq v$, such that $\lambda = u/v$. ■

**Lemma 2** *For any given* $(\mathbf{x}, \mathbf{y}) \in T$, $(\mathbf{x}, \mathbf{y}) \neq (0,0)$, *if there exists a real valued* $\lambda$ *such that* $(\lambda\mathbf{x}, \lambda\mathbf{y}) \in T$, *then there exist integers* $u, v \in \mathbb{Z}_+, \quad u \leq v$, *such that*

$$\lambda = u/v.$$

Proof.

Analogous to Proof of Theorem 1, we note that $(\mathbf{x}, \mathbf{y}) \in T$ implies $(\mathbf{x}, \mathbf{y}) \in \mathbb{Z}_+^{m+s}$. For any $(\mathbf{x}, \mathbf{y}) \neq (0,0)$, it is clear that multiplier $\lambda$ cannot be an irrational number. ■

**Lemma 3** *If the axioms (B2) Natural convexity and (B5) Natural radial rescaling are satisfied, then the axioms of (B3) Natural divisibility and (A6) Additivity must also hold. Conversely, if axioms (B3) and (A6) are satisfied, then axioms (B2) and (B5) must also hold. In other words, these two pairs of axioms are equivalent in the following sense:*

$$[\text{(B2) and (B5)}] \Leftrightarrow [\text{(B3) and (A6)}]$$

Proof. Follows directly from Theorem 1 in KKM and Theorem 4 in KMK.

**Theorem 1** *Production possibility set* $T_{IDEA}^{RTS}$ *is the intersection of all sets* $S \subset \mathbb{Z}_+^{m+s}$ *that satisfy the envelopment condition (E1), the axioms (B1) and (B2), the RTS axioms ((B3), (B4), (B5), or none) corresponding to the specified returns to scale.*

Proof.

See KMK, Theorem 1 (VRS), Theorem 2 (NIRS), Theorem 3 (NDRS), and Theorem 4 (CRS), proved in Appendix A. ■

**Theorem 2** *Production possibility set $T^{RTS}_{HIDEA}$ is the intersection of all sets S that satisfy the envelopment (E1), axioms (A1) and (A2) for the subsets $(I^{NI}, O^{NI})$, axioms (B1) and (B2) for the subsets $(I^{I}, O^{I})$, and the RTS axioms ((A3), (A4), (A5), or none for the subsets $(I^{NI}, O^{NI})$, and (B3), (B4), (B5), or none for for the subsets $(I^{I}, O^{I})$) corresponding to the specified returns to scale.*

Proof.

See KMK, Theorem 5 (VRS), Theorem 6 (NIRS), Theorem 7 (NDRS), and Theorem 8 (CRS), proved in Appendix A. ■

## References


Afriat SN (1972) Efficiency estimation of production functions. Int Econ Rev 13(3):568–598

Alirezaee MR, Sani MRR (2011) An enumeration algorithm for integer-valued data envelopment analysis. Int Trans Oper Res 18(6):729–740

Banker RD, Morey RC (1986) The use of categorical variables in data envelopment analysis. Manage Sci 32(12):1613–1627

Banker RD, Charnes A, Cooper WW, (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30(9):1078–1092

Bogetoft P (1996) DEA on relaxed convexity assumptions. Manage Sci 42(3):457–465

Bogetoft P, Tama JM, Tind J (2000) Convex input and output projections of nonconvex production possibility sets. Manage Sci 46(6):858–869

Chambers RG, Chung YH, Färe R (1996) Benefit and distance functions. J Econ Theory 70(2):407–419

Chambers RG, Chung YH, Färe R (1998) Profit, directional distance function, and Nerlovian efficiency. J Optim Theory Appl 98(2):351–364

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Charnes A, Cooper WW, Golany B, Seiford LM, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econom 30(1):91–107

Cherchye L, Kuosmanen T, Post T (2001). Alternative treatments of congestion in DEA. Eur J Oper Res 132:75–80

Chen CM., Du J, Huo J, Zhu J (2012) Undesirable factors in integer-valued DEA: evaluating the operational efficiencies of city bus systems considering safety records. Decis Support Syst 54(1):330–335

Chen Y, Djamasbi S, Du J, Lim S (2013) Integer-valued DEA super-efficiency based on directional distance function with an application of evaluating mood and its impact on performance. Int J Prod Econ 146(2):550–556

Cooper WW, Park KS, Pastor JT (1999) RAM: a range adjusted measure of inefficiency for use with additive models, and relations to other models and measures in DEA. J Prod Anal 11:5–42

Deprins D, Simar L, Tulkens H, (1984) Measuring labor-efficiency in post offices. In: Marchand M, Pestieau P, Tulkens H, (eds) The performance of public enterprises, concepts and measurement. Elsevier Science Ltd., North Holland

Du J, Chen CM, Chen Y, Cook WD., Zhu J (2012) Additive super-efficiency in integer-valued data envelopment analysis. Eur J Oper Res 218(1):186–192

Farrell M (1957) The measurement of productive efficiency. J Royal Stat Soc Ser A 120(3):253–290

Färe R, Grosskopf S (2009) A Comment on weak disposability in nonparametric production analysis. Am J Agric Econ 91(2):535–538

Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On estimation of technical inefficiency in the stochastic frontier production function model. J Econom 19:233–238

Kamakura WA (1988) A note on the use of categorical variables in data envelopment analysis. Manage sci 34(10):1273–1276
Kazemi Matin, R., Emrouznejad, A. 2011. An integer-valued data envelopment analysis model with bounded outputs. Int Trans Oper Res 18(6):741–749
Kazemi Matin R, Kuosmanen T (2009) Theory of integer-valued data envelopment analysis under alternative returns to scale axioms. Omega 37(5):988–995
Keshvari A Kuosmanen T (2013) Stochastic non-convex envelopment of data: applying isotonic regression to frontier estimation. Eur J Oper Res 231(2):481–491
Khezrimotlagh D, Salleh S, Mohsenpour Z (2012) A comment on theory of integer-valued data envelopment analysis. Appl Math Sci 6(116):5769–5774
Khezrimotlagh D, Salleh S, Mohsenpour Z (2013a) A new robust mixed integer-valued model in DEA. Appl Math Model 37(24):9885–9897
Khezrimotlagh D, Salleh S, Mohsenpour Z (2013b) A note on integer-valued radial model in DEA. Comput Ind Eng 66(1):199–200
Koopmans TC (1951a) Analysis of production as an efficient combination of activities. In: Koopmans TC (ed) Activity analysis of production and allocation. Wiley, New York, pp 33–97
Koopmans TC (ed) (1951b) Activity analysis of production and allocation, cowles commission for research in economics. Wiley, New York
Kuosmanen T (2001) DEA with efficiency classification preserving conditional convexity. Eur J Oper Res 132(2):326–342
Kuosmanen T (2003) Duality theory of non-convex technologies. J Prod Anal 20:273–304
Kuosmanen T (2005) Weak disposability in nonparametric production analysis with undesirable outputs. Am J Agric Econ 87:1077–1082
Kuosmanen T (2008) Representation theorem for convex nonparametric least squares. Econom J 11:308–325
Kuosmanen T, Johnson AL (2010) Data envelopment analysis as nonparametric least squares regression. Oper Res 58(1):149–160
Kuosmanen T, Kazemi Matin R (2009) Theory of integer-valued data envelopment analysis. Eur J Oper Res 192(2):658–667
Kuosmanen T, Kortelainen M (2012) Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. J Prod Anal 38(1):11–28
Kuosmanen T, Podinovski VV (2009) Weak disposability in nonparametric production analysis: reply to Färe and Grosskopf. Am J Agric Econ 91(2):539–545
Kuosmanen T, Post GT, Sipiläinen T (2004) Shadow price approach to total factor productivity measurement: with an application to Finnish grass-silage production. J Prod Anal 22(1):95–121
Kuosmanen T, Johnson AL, Saastamoinen A (2014) Stochastic nonparametric approach to efficiency analysis: a unified framework. In: Zhu J (ed) data envelopment analysis, Springer
Lozano S (2013) Using DEA to find the best partner for a horizontal cooperation. Comput Ind Eng 66(2):286–292
Lozano S, Villa G (2006) Data envelopment analysis of integer-valued inputs and outputs. Comput Oper Res 33(10):3004–3014
Lozano S, Villa G (2007) Integer DEA models. In Zhu J, Cook WD (eds) Modeling data irregularities and structural complexities in data envelopment analysis. Springer, New York, pp 271–290
Lozano S, Villa G, Canca D (2011) Application of centralised DEA approach to capital budgeting in Spanish ports. Comput Ind Eng 60(3):455–465
Nöhren M, Heinzl A (2012) Measuring the relative efficiency of global delivery models in IT outsourcing. Lect Notes Bus Inf Process 130:61–75
Petersen, N. C. (1990) Data Envelopment analysis on a relaxed set of assumptions, *Management Science* 20(3), 305–314
Podinovski VV (2004) Bridging the gap between the constant and variable returns-to-scale models: selective proportionality in data envelopment analysis. J Oper Res Soc 55(3):265–276
Post GT (2001) Transconcave data envelopment analysis. Eur J Oper Res 132(2):374–389

Rousseau JJ, Semple JH (1993) Notes: categorical outputs data envelopment analysis. Manage Sci 39(3):384–386
Shephard R, (1970) Theory of cost and production functions. Princeton University Press, Princeton
Simar L, Wilson PW (2010) Inferences from cross-sectional, stochastic frontier models. Econom Rev 29(1):62–98
Skellam JG (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. J Royal Stat Soc Ser A 109(3):296–296
Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509
Tulkens H (1993) On FDH efficiency analysis: some methodological issues and applications to retail banking, courts, and urban transit. J Prod Anal 4:183–210
Von Neumann J (1945–1946) A model of general economic equilibrium. Rev Econ Stud 13(1):1–9
Wu J Zhou Z, Liang L (2009) Measuring the performance of nations at Beijing summer olympics using integer-valued DEA model. J Sports Econ 11(5):549–566
Wu J, Liang L, Song H (2010) Measuring hotel performance using the integer DEA model. Tour Econ 16(4):867–882

# Chapter 5
# DEA Models with Production Trade-offs and Weight Restrictions

**Victor V. Podinovski**

**Abstract** There is a large literature on the use of weight restrictions in multiplier DEA models. In this chapter we provide an alternative view of this subject from the perspective of dual envelopment DEA models in which weight restrictions can be interpreted as production trade-offs. The notion of production trade-offs allows us to state assumptions that certain simultaneous changes to the inputs and outputs are technologically possible in the production process. The incorporation of production trade-offs in the envelopment DEA model, or the corresponding weight restrictions in the multiplier model, leads to a meaningful expansion of the model of production technology. The efficiency measures in DEA models with production trade-offs retain their traditional meaning as the ultimate and technologically realistic improvement factors. This overcomes one of the known drawbacks of weight restrictions assessed using other methods. In this chapter we discuss the assessment of production trade-offs, provide the corresponding theoretical developments and suggest computational methods suitable for the solution of the resulting DEA models.

**Keywords** Data envelopment analysis · Production trade-offs · Weight restrictions

## 5.1 Introduction

The conventional variable and constant returns-to-scale (VRS and CRS) DEA models can each be stated as two mutually dual linear programs: as an envelopment or multiplier model (Charnes et al. 1978; Banker et al. 1984). The envelopment model is based on an explicit representation of the production technology. The efficiency of decision making units (DMUs) in this model is obtained by their input or output radial projection on the boundary of the technology. The dual multiplier model is stated in terms of variable vectors of input and output weights. This model assesses the efficiency of DMUs in terms of the ratio of their aggregated weighted outputs to

V. V. Podinovski (✉)
Warwick Business School, University of Warwick,
CV4 7AL Coventry, UK
e-mail: v.podinovski@warwick.ac.uk

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_5

aggregated weighted inputs, in relation to similar ratios calculated for all observed DMUs.

One common modification of the multiplier model is based on the use of weight restrictions—the incorporation among its constraints of additional inequalities on the input and output weights (Thanassoulis et al. 2008; Cooper et al. 2011a). Weight restrictions are attractive because of their apparent managerial meaning and also because their use can significantly improve the efficiency discrimination of DEA models (Allen et al. 1997; Thanassoulis et al. 2004).

A well-known drawback of weight restrictions arises from the fact that their use in the multiplier model implicitly changes the model of production technology in the envelopment form (Allen et al. 1997). Specifically, weight restrictions enlarge the model of technology and generally shift the efficient frontier to a more demanding level, as illustrated by Roll et al. (1991). An obvious problem with this is that the efficient projections of inefficient DMUs located on the expanded frontier may not be producible (technologically realistic). Furthermore, the traditional meaning if efficiency as the ultimate and *technologically feasible* improvement factor generally becomes unsubstantiated (Podinovski 2004a; Førsund 2013).

The purpose of this chapter is to describe an approach to the construction of weight restrictions that *by definition* does not have the above drawback. The idea is to consider the dual forms of weight restrictions induced in the envelopment models. These are additional terms that are simultaneously added to, or subtracted from, the inputs and outputs of the units in the production technology. Following Podinovski (2004a), we refer to these terms as *production trade-offs*.

Weights restrictions and production trade-offs are mathematically equivalent. From the practical point of view they may, however, be regarded as different tools. While the terminology of weight restrictions is a natural language for the elicitation and communication of value judgements, the notion of production trade-offs makes us think in terms of production technology and possible substitutions between its inputs and outputs.

Production trade-offs do not generally follow from the data—instead, they are additional assumptions that we (or experts) are willing to make about the production technology: that a certain simultaneous change (substitution) of inputs and outputs is technologically possible, at all units.

In this respect production trade-offs should not be confused with *marginal rates* of transformation and substitution between the inputs and outputs. The latter represent the slopes of the supporting hyperplanes to the technology and are generally different at different boundary units. Changing the inputs and outputs of a boundary unit in the proportions based on the marginal rates (calculated at this unit) would keep the resulting unit on the supporting hyperplane—this does not mean that the resulting unit is producible. In other words, marginal rates represent the movements (changes to inputs and outputs) that are tangent to the technology and are not supposed to result in producible units. In contrast, production trade-offs represent movements that are not necessarily tangent to the boundary of the true technology (that we are attempting to model), but are assumed to keep the resulting units technologically possible.

The use of production trade-offs for the construction of weight restrictions has been illustrated in different contexts. These include the assessment of efficiency of university departments (Podinovski 2007a), secondary schools (Khalili et al. 2010), primary health care providers (Amado and Santos 2009), primary diabetes care providers (Amado and Dyson 2009), electricity distributors (Santos et al. 2011) and agricultural farms (Atici 2012). The following are a few examples of production trade-offs employed in the above studies.

1. *Primary health care provision:* the hospital outputs should not deteriorate if the number of nurses is reduced by 1 and the number of doctors is increased by 1 (Amado and Santos 2009). This corresponds to the weight restriction stating that the weight attached to the number of doctors is at least as large as the weight attached to the number of nurses.
2. *Electricity distribution:* a distribution utility should be able to increase the delivery of electricity by at least 40 KWh per Euro of increase of operating expenses—the latter is chosen as a representative measure for all distribution costs (Santos et al. 2011). This implies that the weight attached to operating expenses (in Euros) is greater than or equal to 40 times the weight attached to the number of KWh delivered.
3. *Agricultural farms:* the resources required for the production of 1 tonne of wheat are sufficient for the production of at least 0.75 tonnes of barley, at any farm in the given region (Atici 2012). This implies that the weight attached to wheat is greater than or equal to 0.75 times the weight attached to barley.

Production trade-offs have exactly the same effect on the model of technology as weight restrictions: the technology expands but, in contrast with the latter case, in *a controlled way* that we explicitly assume to be technologically possible. Because the expanded technology and, therefore, its efficient frontier are realistic in the production sense, this further implies that the radial targets are producible and the efficiency measures retain their conventional technological meaning as possible improvement factors.

The use of production trade-offs overcomes the known drawbacks of weight restrictions not because they are different: as noted, both are equivalent concepts. The advantage of trade-offs is that their assessment explicitly refers to the technology and requires our judgement to be stated in the language of possible changes to inputs and outputs. The assessed trade-offs can be incorporated either in the envelopment model (which currently requires the use of a general linear optimiser), or as equivalent weight restrictions in the multiplier model (which can be performed in most current DEA solvers). In the latter case, the weight restrictions do not have the above known general drawbacks because they are constructed by transformation of production trade-offs. We call this method *the trade-off approach* to the construction of weight restrictions.

It is worth mentioning that several earlier studies came close to the notion of production trade-offs. Charnes et al. (1989), Roll et al. (1991) and Halme and Korhonen (2000) show that the incorporation of weight restrictions in multiplier models induce

dual terms that change the technology but do not explore this relation as a basis for the assessment of weight restrictions that have a production meaning.

The assessment of weight restrictions in some earlier applications of DEA can also be viewed as being *implicitly* based on (or consistent with) the idea of production trade-offs. Dyson and Thanassoulis (1988) consider a DEA model with a single input. In this study the lower bound on each output weight is related to the minimum amount of the input required per unit of the output. This is essentially a statement of a production trade-off, although in a specific DEA model that cannot be easily generalised to the case of multiple outputs. In the assessment of bank branch performance, Schaffnit et al. (1997) and Cook and Zhu (2008) incorporate limits on the ratios of the weights of different transaction and maintenance activities. Such limits are based on the lower and upper bounds on the amounts of time that such activities require and effectively express production trade-offs between the activities.

## 5.2 Production Trade-offs

Following Podinovski (2004a), in this section we introduce production trade-offs as the dual forms of weight restrictions. It is also straightforward to introduce production trade-offs independently and establish their dual relationship to weight restrictions afterwards. We prefer the former approach because it builds up on the already well-established concept of weight restrictions in the DEA literature.

Consider technology $\mathcal{T} \subset \mathbb{R}^m_+ \times \mathbb{R}^s_+$ with $m \geq 1$ inputs and $s \geq 1$ outputs. The elements of $\mathcal{T}$ are DMUs stated as the pairs $(X, Y)$, where $X$ and $Y$ are the input and output vectors, respectively. Let $J = \{1, \ldots, n\}$ be the set of observed DMUs. Each observed DMU can also be stated as $(X_j, Y_j)$, where $j \in J$. Denote $(X_o, Y_o)$ the unit in $\mathcal{T}$ whose efficiency is being assessed.

In order for the DEA models to be well-defined and avoid the consideration of special cases, we make the following standard data assumption: at least one input and one output of each observed DMU is strictly positive. We also assume that every output $r = 1, \ldots, s$ is strictly positive for at least one observed DMU $j_1 \in J$, and every input $i = 1, \ldots, m$ is strictly positive for at least one observed DMU $j_2 \in J$.

Let $v \in \mathbb{R}^m_+$ and $u \in \mathbb{R}^s_+$ be, respectively, the vectors of input and output weights used in the multiplier DEA models. Consider the following $K \geq 1$ homogeneous weight restrictions:

$$u^\top Q_t - v^\top P_t \leq 0, \; t = 1, \ldots, K, \tag{5.1}$$

where $P_t \in \mathbb{R}^m$ and $Q_t \in \mathbb{R}^s$ are some constant vectors, for all $t$, and symbol$^\top$ denotes transposition. Components of vectors $P_t$ and $Q_t$ can be positive, negative or zero. If, for some $t$, both $P_t \neq 0$ and $Q_t \neq 0$, the corresponding weight restriction $t$ in (5.1) is called *linked* and is often referred to as Assurance Region II (Thompson et al. 1990). Otherwise, the weight restriction is *not linked* and is termed Assurance Region I, or polyhedral cone ratio (Charnes et al. 1989, 1990).

Suppose we wish to assess the efficiency of some DMU $(X_o, Y_o)$ using the multiplier model with weight restrictions (5.1). To be specific, we consider the case of CRS first, and comment on the case of VRS afterwards.

The input radial efficiency of DMU $(X_o, Y_o)$ is obtained as the optimal value $\theta^*$ in the following multiplier model that incorporates weight restrictions (5.1):

*Model* $\mathbb{M}^1_{CRS}$:

$$\theta^* = \max \quad u^\top Y_o, \tag{5.2}$$

subject to

$$v^\top X_o = 1,$$

$$u^\top Y_j - v^\top X_j \le 0, \quad j = 1, \ldots, n,$$

$$u^\top Q_t - v^\top P_t \le 0, \quad t = 1, \ldots, K,$$

$$u, v \ge 0.$$

(In model (5.2) and below, the vector inequalities $\le$ and $\ge$ mean that the corresponding inequality is true for each component.)

Note that, although model (5.2) maximises the aggregated output $u^\top Y_o$ of DMU $(X_o, Y_o)$, its dual envelopment form (5.4) presented below projects the latter unit on the boundary of the technology by the radial contraction of the input vector $X_o$. This explains why model (5.2) and its dual are conventionally referred to as input-minimisation, or input-oriented, models (Cooper et al. 2011b).

Similarly, the output radial efficiency of DMU $(X_o, Y_o)$ is equal to the inverse $1/\eta^*$ of the optimal value $\eta^*$ in the following output-maximisation (or output-oriented) multiplier model:

*Model* $\mathbb{M}^2_{CRS}$:

$$\eta^* = \min \quad v^\top X_o, \tag{5.3}$$

subject to

$$u^\top Y_o = 1,$$

$$u^\top Y_j - v^\top X_j \le 0, \quad j = 1, \ldots, n,$$

$$u^\top Q_t - v^\top P_t \le 0, \quad t = 1, \ldots, K,$$

$$u, v \ge 0.$$

The notion of production trade-offs and their relation to weight restrictions becomes apparent when we consider the envelopment models dual to (5.2) and (5.3). Using vectors $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $\pi = (\pi_1, \ldots, \pi_K)$, the dual to (5.2) can be stated as follows:

*Model* $\mathbb{E}^1_{CRS}$:

$$\theta^* = \min \quad \theta, \tag{5.4}$$

$$\begin{aligned} \text{subject to} \quad & \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t \geq Y_o, \\ & \sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t \leq \theta X_o, \\ & \lambda \geq 0, \pi \geq 0, \theta \text{ sign free.} \end{aligned}$$

Similarly, the dual to (5.3) is the envelopment model

*Model* $\mathbb{E}^2_{CRS}$:

$$\begin{aligned} & \eta^* = \max \quad \eta, \\ \text{subject to} \quad & \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t \geq \eta Y_o, \\ & \sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t \leq X_o, \\ & \lambda \geq 0, \pi \geq 0, \eta \text{ sign free.} \end{aligned} \tag{5.5}$$

In both envelopment models (5.4) and (5.5) the first group of terms on the left-hand side of their constraints defines a composite unit $(\bar{X}\lambda, \bar{Y}\lambda)$ in the conventional CRS technology. This unit is further modified by the pairs of vectors

$$(P_t, Q_t),\ t = 1, \ldots, K, \tag{5.6}$$

used in proportions $\pi_t \geq 0$. In particular, vector $P_t$ represents changes to the inputs, and vector $Q_t$ shows changes to the outputs. Therefore, each pair $(P_t, Q_t)$ in (5.6) can be referred to as *a production trade-off*.

It is clear that for some weight restrictions (5.1) the corresponding production trade-offs (5.6) may not represent a technologically possible substitution between the inputs and outputs. In this case the unit obtained on the left-hand side of models (5.4) and (5.5) is generally not producible. An obvious way to overcome this problem is to construct technologically realistic trade-offs (5.6) in the first place. The efficiency of DMU $(X_o, Y_o)$ can then be assessed by solving either the envelopment models (5.4) and (5.5) or their dual multiplier models (5.2) and (5.3). In the latter case, the trade-offs (5.6) should be converted to weight restrictions (5.1).

The above process describes *the trade-off approach* to the construction of weight restrictions. Its idea is that the weight restrictions (5.1) are assessed in the dual envelopment space where they take on the form of production trade-offs (5.6). The latter are essentially additional production assumptions based on our understanding of the technology. Examples illustrating the trade-off approach are discussed in Sect. 5.3 below.

In the case of VRS, the dual relationship between weight restrictions (5.1) and production trade-offs (5.6) is the same as above. The VRS analogues of the CRS envelopment models are the programs (5.4) and (5.5) with the additional normalising condition

$$\sum_{j=1}^{n} \lambda_j = 1. \tag{5.7}$$

Below we denote to the resulting envelopment VRS models as $\mathbb{E}^1_{VRS}$ and $\mathbb{E}^2_{VRS}$, respectively. The corresponding dual multiplier models are referred to as $\mathbb{M}^1_{VRS}$ and $\mathbb{M}^2_{VRS}$. Both VRS multiplier models utilise an additional sign free variable $u_0$ dual to equality (5.7).

**Remark 1** The case of non-homogeneous weight restrictions is considered in Podinovski (2004a, 2005). Such restrictions have a non-zero constant on the right-hand side of inequalities (5.1), an example of which is absolute weight bounds (Dyson and Thanassoulis 1988). Non-homogeneous weight restrictions can also be related to production trade-offs in the envelopment model, but formula (5.6) is no longer valid. The exact trade-off induced by a non-homogeneous weight restriction depends on the DMU $(X_o, Y_o)$ under the assessment and the orientation (input minimisation or output maximisation) of the model. This complicates the assessment of non-homogeneous weight restrictions and makes them less attractive in practical applications.

A further difficulty arising in DEA models with non-homogeneous weight restrictions is that the managerial meaning of the resulting efficiency obtained via the multiplier DEA model may be unclear. In particular, the optimal input and output weights in the resulting models do not generally represent the assessed DMU in the best light compared to the other DMUs (Podinovski and Athanassopoulos 1998; Podinovski 1999, 2004b).

## 5.3 Illustrative Example

Below we consider an example that illustrates the use of production trade-offs in the assessment of efficiency of academic departments from different universities using a hypothetical data set. The departments are assumed to be from the same academic area (e.g., economics). The choice of inputs and outputs in this example is the same as in Podinovski (2007a) but the data set is different.

Table 5.1 shows seven hypothetical university departments denoted D1, D2, ..., D7. The two inputs include full academic staff and research staff. The three outputs include undergraduate students, master (postgraduate) students and academic publications.

To be specific, we consider the case of output radial efficiency. Using the two conventional CRS and VRS output-maximisation DEA models, we obtain the efficiency scores as shown in the second left columns in Tables 5.2 and 5.3 titled "CRS" and "VRS", respectively. It is not surprising that, given the small set of observed DMUs,

**Table 5.1** University departments

| | | Departments | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | D1 | D2 | D3 | D4 | D5 | D6 | D7 |
| Outputs | Undergraduates | 800 | 1200 | 1680 | 630 | 1070 | 1450 | 1550 |
| | Master students | 200 | 500 | 250 | 410 | 120 | 230 | 0 |
| | Publications | 90 | 21 | 2 | 97 | 11 | 109 | 3 |
| Inputs | Full academic staff | 92 | 104 | 64 | 75 | 62 | 98 | 63 |
| | Research staff | 15 | 11 | 0 | 12 | 1 | 32 | 0 |

**Table 5.2** Output radial efficiency (%) of departments in the CRS models with different sets of production trade-offs/weight restrictions

| Department | CRS | CRS 1 | CRS 2 | CRS 3 | CRS 4 | CRS 5 | CRS 6 | CRS 7 | CRS 8 |
|---|---|---|---|---|---|---|---|---|---|
| D1 | 83.27 | 76.82 | 76.82 | 76.47 | 76.47 | 76.47 | 76.47 | 76.47 | 76.47 |
| D2 | 97.37 | 97.37 | 73.30 | 69.59 | 69.59 | 69.59 | 69.59 | 69.59 | 69.59 |
| D3 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| D4 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| D5 | 100.00 | 100.00 | 100.00 | 100.00 | 77.61 | 75.92 | 75.92 | 71.78 | 71.78 |
| D6 | 100.00 | 100.00 | 100.00 | 86.63 | 86.63 | 86.63 | 86.63 | 86.63 | 86.63 |
| D7 | 100.00 | 100.00 | 100.00 | 100.00 | 85.33 | 84.61 | 84.61 | 83.02 | 83.02 |

**Table 5.3** Output radial efficiency (%) of departments in the VRS models with different sets of production trade-offs/weight restrictions

| Department | VRS | VRS 1 | VRS 2 | VRS 3 | VRS 4 | VRS 5 | VRS 6 | VRS 7 | VRS 8 |
|---|---|---|---|---|---|---|---|---|---|
| D1 | 95.68 | 91.09 | 91.09 | 90.07 | 90.07 | 81.52 | 81.52 | 81.52 | 81.52 |
| D2 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 95.41 | 95.41 | 95.41 | 92.61 |
| D3 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| D4 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| D5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 79.66 | 79.66 |
| D6 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 94.54 | 94.54 | 94.54 | 94.54 |
| D7 | 100.00 | 100.00 | 100.00 | 100.00 | 99.36 | 99.36 | 97.89 | 87.03 | 87.03 |

the efficiency discrimination is low: in the case of CRS only two departments are inefficient, and in the case of VRS only one is inefficient.

Table 5.4 shows the optimal input and output weights obtained in the standard CRS model. The weights $u_1$, $u_2$ and $u_3$ correspond to the three outputs: undergraduate students, master students and publications, respectively. The weights $v_1$ and $v_2$ correspond to the two inputs: academic and research staff, respectively. Although optimal weights are generally not unique, the weights in Table 5.4 are consistent

**Table 5.4** Optimal output and input weights in the standard output-oriented multiplier CRS model

| Department | $u_1$ | $u_2$ | $u_3$ | $v_1$ | $v_2$ |
|---|---|---|---|---|---|
| D1 | 0.0004 | 0 | 0.0073 | 0.0116 | 0.0088 |
| D2 | 0 | 0.002 | 0 | 0.0078 | 0.0195 |
| D3 | 0.0002 | 0.0025 | 0 | 0.0156 | 0 |
| D4 | 0 | 0 | 0.0103 | 0.0005 | 0.08 |
| D5 | 0.0001 | 0.0003 | 0.0762 | 0.0066 | 0.59 |
| D6 | 0.0003 | 0 | 0.0051 | 0.0082 | 0.0062 |
| D7 | 0.0005 | 0 | 0.0543 | 0.0159 | 0.3683 |

with the known drawback of conventional DEA models: the complete flexibility of weights often results in zero weights attached to some of the inputs and outputs. These represent the areas in which the DMU under the assessment is relatively weak (Thanassoulis et al. 1987; Dyson and Thanassoulis 1988).

For example, department D4 has a relatively low number of students per member of staff but the highest number of publications per staff. This is reflected in the optimal weights attached to these outputs: both undergraduate and master students have a zero weight attached to them. This implies that the DEA model used for the assessment of department D4 effectively ignores the first two outputs. Exactly the same efficiency score for department D4 is obtained if we remove the two types of student from model specification and assess the efficiency of D4 based on the two inputs and publications only.

Let us show that both the CRS and VRS DEA models can be improved using simple production trade-offs.

### *5.3.1 Undergraduate and Master Students*

We start by comparing the resources (academic staff) that are used by the departments for the teaching of undergraduate and master students.

**Assumption 1** The teaching of one undergraduate student does not require more resources (academic staff time) than the teaching of one master student.

We can restate the above assumption as the following trade-off that all departments should accept:

$$P_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, Q_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}. \tag{5.8}$$

The meaning of the above trade-off is straightforward: it is possible to replace one master student (the value $-1$ in the second component of vector $Q_1$) by one undergraduate student (the value 1 in the first component of vector $Q_1$). For this replacement, no change of the inputs (resources) is needed: vector $P_1$ is a zero vector. There should also be no change to the third output (publications): the third component of vector $Q_1$ is zero.

Production trade-off (5.8) can be restated as a weight restriction using formula (5.1):

$$u_1 - u_2 \leq 0. \tag{5.9}$$

This inequality implies that the weight attached to master students cannot be less than the weight attached to undergraduate students. The same weight restriction may possibly be obtained by a value judgement but it is the original production trade-off (5.8) that makes this weight restriction meaningful in the production sense.

**Assumption 2** The teaching of a master student may require more resources than an undergraduate student, however, by no more than a factor of 3.

Note that the above assumption is not a precise measure of the relative amount of resources required by the two types of output, and it should not be for two reasons. First, the estimates of this ratio may vary depending on the methodology used for its calculation even for one particular department. Second, even if the precise ratio were possible to assess, this would most likely vary between the departments. Because of these uncertainties, Assumption 2 is supposed to be a safe conservative estimate (an upper bound of different possible estimates) that all departments should agree on.

We state Assumption 2 as the following production trade-off:

$$P_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, Q_2 = \begin{pmatrix} -3 \\ 1 \\ 0 \end{pmatrix}. \tag{5.10}$$

The above trade-off means that no extra resources should be claimed ($P_2$ is a zero vector) and there should be no detriment to the publications if the number of undergraduate students is reduced by 3 and the number of master students is increased by 1. Using formula (5.1), production trade-off (5.10) is restated as the weight restriction

$$-3u_1 + u_2 \leq 0. \tag{5.11}$$

According to (5.11), the weight attached to master students cannot be more than 3 times larger than the weight attached to undergraduate students. Note that the factor 3 does not reflect the perceived importance of master students compared to undergraduates, as both outputs may be deemed equally important for the departments or the decision maker who is assessing their efficiency. The factor 3 is obtained as (the upper bound on) the ratio of the resources that these two outputs require. This should be acceptable to all departments.

### 5.3.2 Research Staff and Publications

Consider the role of research staff in producing academic publications. Because the rate of publications may vary between different departments and individual researchers, the following two assumptions are intended to be sufficiently conservative.

**Assumption 3** Each researcher should be able to publish at least one paper in two years.

The above statement can be stated as the following production trade-off:

$$P_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, Q_3 = \begin{pmatrix} 0 \\ 0 \\ 0.5 \end{pmatrix}. \tag{5.12}$$

This trade-off implies that if the number of researchers is increased by 1, it should be possible to increase the number of papers by 0.5 per year. Equivalently, using formula (5.1), the above trade-off translates to the following linked weight restriction:

$$0.5u_3 - v_2 \leq 0.$$

**Assumption 4** No department can justify a reduction of the number of papers by more than 6 per year by referring to a loss of one research staff.

The number 6 in the above statement is purely speculative and is simply used as an illustration of a reasonably high research output. In real applications this can be revised either way. Assumption 4 is stated as the following production trade-off:

$$P_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, Q_4 = \begin{pmatrix} 0 \\ 0 \\ -6 \end{pmatrix}. \tag{5.13}$$

Equivalently, Assumption 4 can be stated as the following weight restriction:

$$-6u_3 + v_2 \leq 0.$$

Production trade-offs (5.12) and (5.13) effectively specify the lower and upper bounds on the number of papers per average researcher at any department. Any publication rate below the lower bound of 0.5 is treated as evidence of inefficiency. A publication rate above the upper bound of 6 is regarded as unrealistically high.

### 5.3.3 Academic Staff and Students

There are different ways in which the link between academic staff and their outputs (students and publications) can be expressed. Below we consider two statements that link one input and two outputs simultaneously in a single trade-off.

The idea of these two assumptions is based on the common use of student-to-staff ratios at academic departments and the expectation of certain publication rates.

**Assumption 5** One full academic post is a sufficient resource for the number of undergraduate students at the department to increase by 10 and the number of publications to increase by 0.5.

This assumption is stated as the following production trade-off:

$$P_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, Q_5 = \begin{pmatrix} 10 \\ 0 \\ 0.5 \end{pmatrix}. \tag{5.14}$$

It further translates to the linked weight restriction:

$$10u_1 + 0.5u_3 - v_1 \leq 0. \tag{5.15}$$

**Assumption 6** A loss of one academic post should not lead to a reduction of more than 20 undergraduate students and 5 publications per year.

This assumption is represented by the following trade-off

$$P_6 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, Q_6 = \begin{pmatrix} -20 \\ 0 \\ -5 \end{pmatrix}, \tag{5.16}$$

and the following equivalent weight restriction:

$$-20u_1 - 5u_3 + v_1 \leq 0. \tag{5.17}$$

### *5.3.4 Students and Publications*

Three university departments in our data set, D1, D4 and D6, can be regarded as research-intensive. They have a moderate teaching-to-staff ratio and a relatively high publication rate. Departments D3 and D7 are focused primarily on the teaching. They have a high student-to-staff ratio and a low number of publications. Overall, this suggests that the departments in Table 5.1 can be viewed as having *different specialisations*.

Highly specialised DMUs are often shown as efficient by DEA models. This is because the peer groups of units to which specialised units can be compared have to show a similar specialisation, which is a limiting factor. Below we overcome the above problem by relating the "production" of students and publications by means of production trade-offs. The latter are based on the evaluation of the resources (staff time) that are needed for the generation of the two outputs.

**Assumption 7** The reduction of the number of undergraduate students by 20 releases the academic staff time sufficient to write one academic paper.

As a justification of the above statement, we can think of an academic member of staff being on a one-year study leave. This involves no teaching load and an expectation of several research outputs. The reduction of undergraduate students by 20 can be approximately equated to one year of staff time, and the publication of just one paper is a conservative estimate of the publication output achievable within one year. This assumption is stated as the following production trade-off:

$$P_7 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, Q_7 = \begin{pmatrix} -20 \\ 0 \\ 1 \end{pmatrix}. \tag{5.18}$$

It further translates to the weight restriction:

$$-20u_1 + u_3 \leq 0.$$

**Assumption 8** The reduction of the number of publications by 5 releases the academic staff time sufficient to increase the number of undergraduate students by 20.

This assumption is stated as the following trade-off:

$$P_8 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, Q_8 = \begin{pmatrix} 20 \\ 0 \\ -5 \end{pmatrix}. \tag{5.19}$$

It further translates to the weight restriction:

$$20u_1 - 0.5u_3 \leq 0.$$

Taken together, trade-offs (5.18) and (5.19) put the bounds on the ratio between the resources (staff time) required to teach undergraduate students and publish papers. Namely, the teaching of 20 undergraduate students may, depending on the department, equate to the writing of between 1 and 5 papers. If the number of students is reduced by 20, any department should be able to compensate for this by increasing the number of publications by at least 1 paper per year. If the number of students is increased by 20 (and the staff number is kept constant), this may be used to justify the reduction of publications by no more than 5 papers per year.

### *5.3.5 Computational Results*

Tables 5.2 and 5.3 show the output radial efficiency of all departments in the CRS and VRS DEA models with different sets of production trade-offs. We obtained these

**Table 5.5** Optimal output and input weights in the final output-oriented multiplier CRS model with all eight production trade-offs/weight restrictions

| Department | $u_1$ | $u_2$ | $u_3$ | $v_1$ | $v_2$ |
|---|---|---|---|---|---|
| D1 | 0.0004 | 0.0004 | 0.0068 | 0.0120 | 0.0135 |
| D2 | 0.0003 | 0.0010 | 0.0047 | 0.0128 | 0.0093 |
| D3 | 0.0005 | 0.0005 | 0.0088 | 0.0156 | 0.0176 |
| D4 | 0.0003 | 0.0003 | 0.0067 | 0.0103 | 0.0188 |
| D5 | 0.0007 | 0.0007 | 0.0142 | 0.0218 | 0.0397 |
| D6 | 0.0003 | 0.0003 | 0.0049 | 0.0086 | 0.0097 |
| D7 | 0.0006 | 0.0006 | 0.0124 | 0.0191 | 0.0348 |

results using a common commercial solver. Obviously, solving the envelopment and corresponding multiplier models led to the same efficiency scores.

As noted above, in these two tables, the columns titled "CRS" and "VRS" correspond to the standard DEA models without production trade-offs. Models CRS $k$ and VRS $k$, where $k = 1, \ldots, 8$, incorporate all production trade-offs $(P_t, Q_t)$, $t = 1, \ldots, k$ stated above. For example, models CRS 1 and VRS 1 incorporate the single trade-off $(P_1, Q_1)$ as stated in (5.8). Models CRS 3 and VRS 3 incorporate three trade-offs $(P_1, Q_1)$, $(P_2, Q_2)$ and $(P_3, Q_3)$. Models CRS 8 and VRS 8 incorporate all eight production trade-offs.

Both tables allow us to observe the gradual improvement of efficiency discrimination as additional trade-offs are progressively incorporated. The final columns CRS 8 and VRS 8 show a significant improvement over the conventional CRS and VRS models.

Table 5.5 shows the optimal input and output weights in model CRS 8. These weights were obtained by solving the dual multiplier model with the eight weight restrictions equivalent to the production trade-offs. In comparison to Table 5.4, all optimal weights in Table 5.5 are strictly positive.

In this respect it should be noted that in practical applications of production trade-offs the aim of making all optimal weights strictly positive may be a goal that is hard to achieve. The incorporation of realistic production trade-offs (or weight restrictions based on them) is a worthwhile improvement to the DEA model, even if this does not completely eliminate all zero weights in the optimal solution.

## 5.4 Graphical Illustrations

To illustrate the effect of production trade-offs on the technology, consider the following two examples. Both are concerned with the assessment of efficiency of university departments. Note that these departments are different from those in Table 5.1.

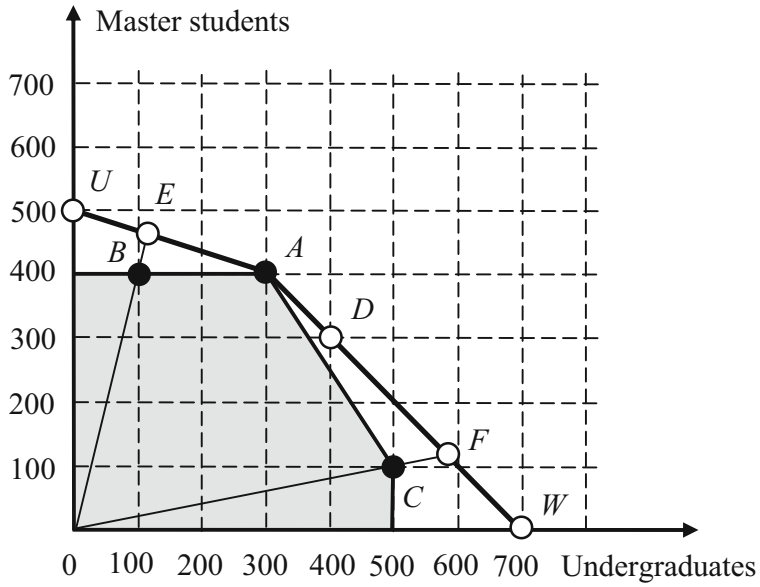



**Fig. 5.1** Production trade-offs expanding the technology in output dimensions

**Example 1** Let units $A$, $B$ and $C$ shown in Fig. 5.1 be observed departments. These departments are assumed to have the same level of a single input (staff) which is not depicted, and different levels of two outputs: undergraduate and master students. Because the input is equal, the shaded area represents both the VRS and CRS technology induced by the three units. More precisely, the shaded area is the section of either technology for the given level of input. For simplicity, we still refer to this section as the technology.

The efficient frontier of this technology is the line segment $AC$. Department $B$ is located on the boundary of technology but is dominated by $A$. It is therefore only weakly efficient. The output radial efficiency of all three departments is equal to 1.

Consider production trade-off (5.8). (We ignore the publications and research staff that are not present in this example.) By the assumption made, this trade-off can be applied to any department. For example, starting at $A$, we can increase the number of its undergraduate students by 1 and simultaneously reduce the number of master students by 1. This procedure can be repeated multiple times. Increasing the number of undergraduate students of department $A$ by 100 and reducing the number of master students also by 100, we arrive at the hypothetical department $D$. Continuing this process, we induce the straight line $AW$.

We have shown that the line $AW$ consists of producible units and should therefore be regarded as part of the technology. Using the free disposability of outputs, we should also add the nonnegative area below $AW$ to the technology. Note that, if we start at any other unit, e.g., at $B$ or $C$, the application of trade-off (5.8) does not add any further new points to the technology.

The use of trade-off (5.8) allows us to add new units in the scenario in which the number of undergraduate students is increased. To consider the reduction of this input, we need to refer to production trade-off (5.10). Starting at point $A$ and using the same logic as above, we move away from $A$ to point $U$. All points on the line $AU$ are producible because we are replacing 3 undergraduate students by 1 master student in this process, as in trade-off (5.10). This adds the line $AU$ to the technology, along with the dominated nonnegative region below it.

Overall, the specification of two trade-offs (5.8) and (5.10) results in the expansion of the technology from the shaded area in Fig. 5.1 to the area below the broken line $UAW$, and the latter is the new efficient frontier. Department $A$ remains efficient, while departments $B$ and $C$ are no longer efficient and are projected on the units $E$ and $F$, respectively. Note that, because of the assumptions about production trade-offs (5.8) and (5.10), both target units $E$ and $F$ are technologically feasible. Therefore, the output radial efficiency of the units $B$ and $C$ retains its traditional technological meaning. Namely, for each unit the inverse of its output radial efficiency is the ultimate improvement factor by which both of its outputs can be improved.

**Example 2** In this example we illustrate the effect of linked production trade-offs on the production technology. For simplicity we consider the case of VRS with a single input (academic staff) and a single output (undergraduate students). The shaded area in Fig. 5.2 corresponds to the VRS technology induced by two departments $A$ and $B$. Both departments are efficient in this technology.

Consider the following variants of linked production trade-offs (5.14) and (5.16) adapted to our example:

$$P_5^* = (1), Q_5^* = (10), \tag{5.20}$$

$$P_6^* = (-1), Q_6^* = (-20). \tag{5.21}$$

We use the same logic as in Example 1. Starting from unit $A$ and applying trade-off (5.20) in different proportions, we add the ray $AW$ to the technology. Similarly, the application of trade-off (5.21) to unit $A$ induces the line $AK$. Using free disposability of input and output, the VRS technology expands to the nonnegative area below the broken line $KAW$, and the latter is its new efficient frontier.

Note that unit $B$ is no longer efficient in the expanded technology. Its output radial efficiency is assessed by its projection on the unit $E$. Because the latter unit is producible according to the stated trade-off assumptions, it is a technologically feasible efficient target for department $B$.

## 5.5 CRS and VRS Technology with Production Trade-offs

Above we defined production trade-offs as the dual forms of weight restrictions. Their use in the example involving university departments resulted in a meaningful expansion of the CRS and VRS technology and led to a significant improvement of efficiency discrimination.

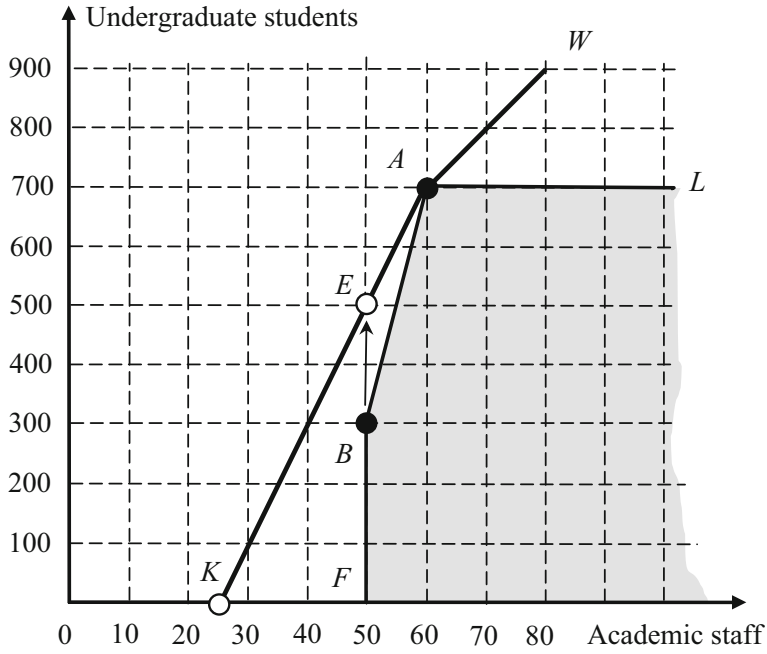



**Fig. 5.2** Linked production trade-offs expanding the VRS technology

The missing link in the above development is the definition of technology with production trade-offs. Below we address this gap using the axiomatic approach to the definition of technology pioneered by Banker et al. (1984). The main definitions and results of this section are based on the results of Podinovski (2004a).

### *5.5.1 Axiomatic Definitions*

The first three axioms are the standard production assumptions that define the conventional VRS technology $\mathcal{T}_{VRS}$. Adding the fourth axiom defines the CRS technology $\mathcal{T}_{CRS}$.

**Axiom 1 (Feasibility of observed data)** $(X_j, Y_j) \in \mathcal{T}$, for any $j \in J$.

**Axiom 2 (Convexity)** Technology $\mathcal{T}$ is a convex set.

**Axiom 3 (Free disposability)** If $(X, Y) \in \mathcal{T}$, $Y \geq Y' \geq 0$ and $X \leq X'$, then $(X', Y') \in \mathcal{T}$.

**Axiom 4 (Proportionality)** If $(X, Y) \in \mathcal{T}$ and $\alpha \geq 0$, then $(\alpha X, \alpha Y) \in \mathcal{T}$.

The following axiom states that each of the production trade-offs $(P_t, Q_t)$ in (5.6) can be applied to any unit in technology $\mathcal{T}$, and any number of times (in any proportion) $\pi_t \geq 0$ as long as the resulting unit has nonnegative inputs and outputs.

**Axiom 5 (Feasibility of production trade-offs)** Let $(X, Y) \in \mathcal{T}$. Then, for each trade-off $(P_t, Q_t)$ in (5.6) and for any $\pi_t \geq 0$, the unit

$$(\tilde{X}, \tilde{Y}) = (X + \pi_t P_t, Y + \pi_t Q_t) \in \mathcal{T},$$

provided $\tilde{X} \geq 0$ and $\tilde{Y} \geq 0$.

The next, and last, axiom states that the production technology should be a closed set. This is a standard property of production technologies (Shephard 1974; Färe et al. 1985) that is often automatically satisfied and needs not to be stated—this is true in the cases of CRS, VRS and free disposal hull technology of Deprins et al. (1984). However, as shown by an example in Podinovski (2004a), this is not so for technologies that incorporate production trade-offs as stated in Axiom 5. Therefore, the following axiom needs to be explicitly stated.

**Axiom 6 (Closedness)** Technology $\mathcal{T}$ is a closed set.

The following definition is based on *the minimum extrapolation principle* introduced to DEA by Banker et al. (1984).

**Definition 1** The CRS technology $\mathcal{T}_{CRS-TO}$ with trade-offs (5.6) is the intersection of all technologies $\mathcal{T}$ that satisfy Axioms 1–6.

It is straightforward to verify that technology $\mathcal{T}_{CRS-TO}$ satisfies all Axioms 1–6. For example, Axiom 2 is satisfied because the intersection of convex sets is a convex set. Definition 1 implies that $\mathcal{T}_{CRS-TO}$ is the smallest technology that satisfies all Axioms 1–6. This means that it contains only those DMUs that are required to satisfy the axioms and no other arbitrary units.

The above definition is not constructive, and its equivalent operational statement is given by the following theorem.

**Theorem 1 (Podinovski 2004a)** Technology $\mathcal{T}_{CRS-TO}$ is the set of all units $(X, Y) \in \mathbb{R}^m_+ \times \mathbb{R}^s_+$ that can be stated in the form

$$Y = \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t - e, \tag{5.22}$$

$$X = \sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t + d, \tag{5.23}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n_+$, $\pi = (\pi_1, \ldots, \pi_K) \in \mathbb{R}^K_+$, $e \in \mathbb{R}^s_+$ and $d \in \mathbb{R}^m_+$.

Theorem 1 provides a meaningful interpretation to the envelopment models (5.4) and (5.5). It shows that the radial improvement of the input and, respectively, output vectors of the unit $(X_o, Y_o)$ is performed within the technology $\mathcal{T}_{CRS-TO}$. Note, however, that this interpretation is correct only if the improved unit has nonnegative

input and output vectors, as required by Theorem 1. This requirement is automatically satisfied in model (5.5) because the output-improvement factor $\eta$ is maximised. In model (5.4) the input-improvement factor $\theta$ is minimised and may in some cases become negative. It may appear that we need to add the condition $\theta \geq 0$ to the constraints of model (5.4)—this would remedy the problem and guarantee that the minimisation of $\theta$ is performed in technology $\mathcal{T}_{CRS-TO}$. While this is possible, there are two reasons why this may not be a good idea.

First, adding the condition $\theta \geq 0$ to the constraints of model (5.4) would invalidate its duality with the multiplier model (5.2). The second and, perhaps, more important consideration is that the feasibility of negative values of $\theta$ in model (5.4) indicates an inconsistency within the trade-offs (5.6) or, equivalently, weight restrictions (5.1). Allowing $\theta$ to take on negative values in the envelopment models make them self-testing for errors in the construction of trade-offs (or weight restrictions). We consider this issue in detail in the next section.

Generally though, the nonnegativity conditions $(X, Y) \in \mathbb{R}^m_+ \times \mathbb{R}^s_+$ are important in the statement of technology $\mathcal{T}_{CRS-TO}$ and should not be omitted unless proved redundant in a particular DEA model. This is discussed further in Sect. 5.7 (see Remark 2) in relation to the additive DEA model based on the above technology.

In the case of VRS, we follow the same logic as above and give the following definition.

**Definition 2** The VRS technology $\mathcal{T}_{VRS-TO}$ with trade-offs (5.6) is the intersection of all technologies $\mathcal{T}$ that satisfy Axioms 1–3, 5 and 6.

As in the above case, it is straightforward to verify that technology $\mathcal{T}_{VRS-TO}$ satisfies Axioms 1–3, 5 and 6 and is, therefore, the smallest technology that satisfies them.

**Theorem 2 (Podinovski 2004a)** Technology $\mathcal{T}_{VRS-TO}$ is the set of all units $(X, Y) \in \mathbb{R}^m_+ \times \mathbb{R}^s_+$ that can be stated in the form (5.22) and (5.23), subject to the additional normalising equality (5.7) and the same nonnegativity conditions on vectors $\lambda$, $\pi$, $e$ and $d$ as in Theorem 1.

The duality of weight restrictions and production trade-offs allows us to give a positive answer to the long-standing question of whether the use of weight restrictions in VRS DEA models is theoretically sound (Thanassoulis and Allen 1998). The counter-argument is that in CRS models the marginal rates of transformation and substitution between inputs and outputs (that define the slopes of facets on the boundary of the technology) are invariant with respect to the scaling (or the size) of the unit, while in the VRS technology this is not so. The main concern is then that the weight restrictions that specify bounds on the marginal rates would be inappropriate in the VRS technology because such rates change with the scale of operations. This argument is weakened by the fact that the marginal rates in the CRS technology are still generally different at any two units, unless one is a scaled variant of the other.

The above problem does not arise if we interpret weight restrictions as the dual forms of production trade-offs and assess the latter in the first place. Indeed, if production trade-offs (5.6) are assumed technologically feasible in the CRS technology $\mathcal{T}_{CRS-TO}$ (in the sense of Axiom 5), then they must be technologically feasible in the

VRS technology $\mathcal{T}_{VRS-TO}$ because the latter is a subset of $\mathcal{T}_{CRS-TO}$. Therefore, any production trade-offs (or weight restrictions based on them) that are deemed realistic and appropriate in the CRS model, are also acceptable and can be used in the VRS model.

### 5.5.2 *Some Properties of CRS and VRS Technologies with Trade-offs*

Below we establish two properties of technologies $\mathcal{T}_{CRS-TO}$ and $\mathcal{T}_{VRS-TO}$.

**Theorem 3** Technologies $\mathcal{T}_{CRS-TO}$ and $\mathcal{T}_{VRS-TO}$ are polyhedral sets. In particular, $\mathcal{T}_{CRS-TO}$ is a polyhedral cone.

**Proof of Theorem 3** The set $P$ of all solutions $\{X, Y, \lambda, \pi, e, d\}$ to the set of linear equations (5.22), (5.23) and inequalities $X, Y, \lambda, \pi, e, d \geq 0$ is a polyhedral set in $\mathbb{R}^{2(m+s)+n+K}$. Technology $\mathcal{T}_{CRS-TO}$ in Definition 1 is the projection of $P$ on its input and output dimensions $X$ and $Y$. By the known projection lemma (see, e.g., Jones et al. 2008, Lemma 3.1), $\mathcal{T}_{CRS-TO}$ is a polyhedral set. Because $\mathcal{T}_{CRS-TO}$ satisfies Axiom 4, it is a cone. The case of technology $\mathcal{T}_{VRS-TO}$ is considered in a similar way. ■

The second property is somewhat more subtle. Without production trade-offs, the conventional CRS technology is the cone extension of the VRS technology. This means that any unit $(X, Y)$ in the CRS technology is obtained by the scaling of some unit $(\tilde{X}, \tilde{Y})$ in the VRS technology by some factor $\alpha \geq 0$. This result is generally incorrect for the CRS and VRS technologies that incorporate production trade-offs (although it is "almost correct" in the sense defined below).

**Example 3** Consider the CRS and VRS technologies with a single input and single output induced by the single observed unit $A = (2,1)$. Suppose we specified the linked trade-off: $(P, Q) = (1,2)$. Figure 5.3 shows the resulting VRS technology $\mathcal{T}_{VRS-TO}$ as the shaded area below the broken line $GAF$. Note that the ray $AF$ is obtained by the application of trade-off $(P, Q)$ to the unit $A$ following the same logic as in Example 2. Furthermore, the CRS technology $\mathcal{T}_{CRS-TO}$ is the cone under the ray $OE$: the ray $OE$ is obtained by the application of trade-off $(P, Q)$ to the zero unit—the latter is included in the original CRS technology. This implies that, for example, unit $B = (1,2)$ is in technology $\mathcal{T}_{CRS-TO}$. (As an alternative argument, unit $B$ satisfies the conditions of Theorem 1 with $\lambda_1 = 0$ and $\pi_1 = 1$.) It is, however, straightforward to show that there exists no unit $(\tilde{X}, \tilde{Y}) \in \mathcal{T}_{VRS-TO}$ and $\alpha \geq 0$ such that $B = \alpha(\tilde{X}, \tilde{Y})$.

Example 3 shows that technology $\mathcal{T}_{CRS-TO}$ is generally not the cone extension of $\mathcal{T}_{VRS-TO}$. Below we prove that $\mathcal{T}_{CRS-TO}$ is the *closed* cone extension of $\mathcal{T}_{VRS-TO}$. To state this formally, denote the cone extension of $\mathcal{T}_{VRS-TO}$ as

$$\text{cone } \mathcal{T}_{VRS-TO} = \left\{(X, Y) \in \mathbb{R}^m \times \mathbb{R}^s \,|\, \exists (\tilde{X}, \tilde{Y}) \in \mathcal{T}_{VRS-TO}, \alpha \geq 0 : (X, Y) = \alpha(\tilde{X}, \tilde{Y})\right\}.$$

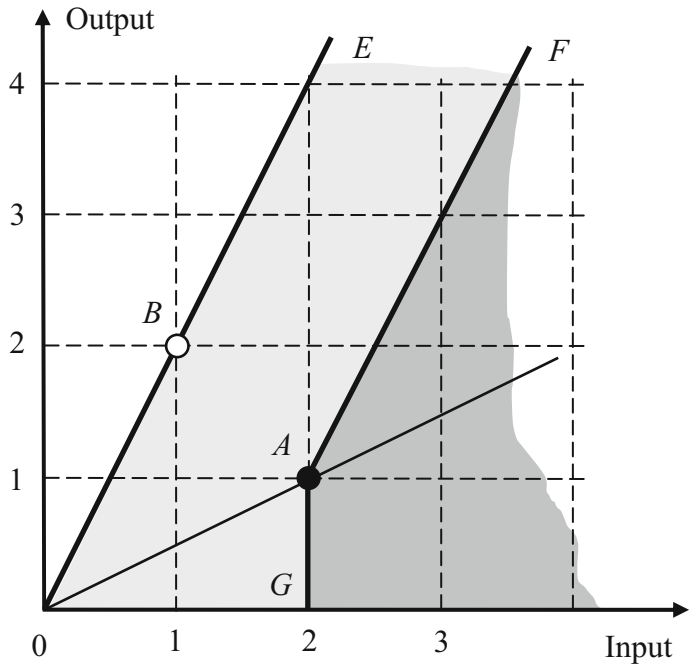



**Fig. 5.3** The VRS (*dark grey*) and CRS (*light grey*) technologies induced by unit $A$ and production trade-off $(P, Q) = (1,2)$

Denote $cl(cone\ \mathcal{T}_{VRS-TO})$ the closure of the set $cone\ \mathcal{T}_{VRS-TO}$ (intersection of all closed sets containing $cone\ \mathcal{T}_{VRS-TO}$).

**Theorem 4** Technology $\mathcal{T}_{CRS-TO}$ is the closed cone induced by $\mathcal{T}_{VRS-TO}$:

$$\mathcal{T}_{CRS-TO} = cl(cone\ \mathcal{T}_{VRS-TO}).$$

**Proof of Theorem 4** By Theorem 2, any $(\tilde{X}, \tilde{Y}) \in \mathcal{T}_{VRS-TO}$ satisfies (5.22), (5.23) and (5.7) with some vectors $\tilde{\lambda}$, $\tilde{\pi}$, $\tilde{e}$ and $\tilde{d}$. For any $\alpha \geq 0$, $\alpha(\tilde{X}, \tilde{Y})$ satisfies (5.22) and (5.23) with the vectors $\alpha\tilde{\lambda}$, $\alpha\tilde{\pi}$, $\alpha\tilde{e}$ and $\alpha\tilde{d}$. By Theorem 1, $(\tilde{X}, \tilde{Y}) \in \mathcal{T}_{CRS-TO}$. Therefore, $cone\ \mathcal{T}_{VRS-TO} \subseteq \mathcal{T}_{CRS-TO}$, and $cl(cone\ \mathcal{T}_{VRS-TO}) \subseteq cl\ \mathcal{T}_{CRS-TO} = \mathcal{T}_{CRS-TO}$. (The last equality is true because $\mathcal{T}_{CRS-TO}$ satisfies Axiom 6.)

Conversely, let $(X, Y) \in \mathcal{T}_{CRS-TO}$. Then $(X, Y)$ satisfies (5.22) and (5.23) with some vectors $\lambda'$, $\pi'$, $e'$ and $d'$. Let $\lambda^* = \sum_{j=1}^{n} \lambda'_j$. Two cases arise.

*Case 1* Assume that $\lambda^* > 0$. Define $(\tilde{X}, \tilde{Y}) = (1/\lambda^*)(X, Y)$. Then $(\tilde{X}, \tilde{Y}) \in \mathcal{T}_{VRS-TO}$ because it satisfies (5.22), (5.23) and (5.7) with $\lambda = \lambda'/\lambda^*$, $\pi = \pi'/\lambda^*$,

$e = e'/\lambda^*$ and $d = d'/\lambda^*$. Because $(X, Y) = \alpha(\tilde{X}, \tilde{Y})$ where $\alpha = \lambda^*$, we have $(X, Y) \in cone\ \mathcal{T}_{VRS-TO} \subseteq cl(cone\ \mathcal{T}_{VRS-TO})$.

*Case 2* Assume that $\lambda^* = 0$. Therefore, $\lambda' = 0$. (This is the case for unit $B$ in Example 3.) Consider the sequence of units $(X_k, Y_k)$, $k = 1,2,\ldots$, defined as follows:

$$(X_k, Y_k) = \sum_{j=1}^{n} \left( \frac{1}{n}(X_j, Y_j) \right) + k(X, Y). \tag{5.24}$$

Because both terms on the right-hand side of (5.24) are nonnegative, each unit $(X_k, Y_k)$ is nonnegative. It is straightforward to verify that $(X_k, Y_k)$ satisfies conditions (5.22), (5.23) and (5.7) with the vector $\lambda_k$ whose components are $(\lambda_k)_j = 1/n$, $j = 1,\ldots,n$, and vectors $\pi_k = k\pi'$, $e_k = ke'$ and $d_k = kd'$. Therefore, $(X_k, Y_k) \in \mathcal{T}_{VRS-TO}$, for all $k = 1,2,\ldots$

Define the sequence of units $(\tilde{X}_k, \tilde{Y}_k) = (1/k)(X_k, Y_k)$. Obviously, we have $(\tilde{X}_k, \tilde{Y}_k) \in cone\ \mathcal{T}_{VRS-TO}$, for all $k$. Note that $(X, Y)$ is the limit unit of the sequence of units $(\tilde{X}_k, \tilde{Y}_k)$. Indeed, based on (5.24),

$$(\tilde{X}_k, \tilde{Y}_k) = \frac{1}{k} \sum_{j=1}^{n} \left( \frac{1}{n}(X_j, Y_j) \right) + (X, Y) \underset{k \to +\infty}{\to} (X, Y).$$

Therefore $(X, Y) \in cl(cone\ \mathcal{T}_{VRS-TO})$. Because $(X, Y)$ is an arbitrary unit in $\mathcal{T}_{CRS-TO}$, in both cases 1 and 2 we have $\mathcal{T}_{CRS-TO} \subseteq cl(cone\ \mathcal{T}_{VRS-TO})$. Taking into account the inverse embedding obtained in the first part of the proof, we have $\mathcal{T}_{CRS-TO} = cl(cone\ \mathcal{T}_{VRS-TO})$. ■

Theorem 4 states that the CRS technology $\mathcal{T}_{CRS-TO}$ is obtained from the VRS technology $\mathcal{T}_{VRS-TO}$ by the scaling of its units by all factors $\alpha \geq 0$, and subsequently adding all limit points (units) to the resulting set.

## 5.6 Weight Restrictions and the Infeasibility Problem

It is well-known that the use of weight restrictions in multiplier models (5.2) and (5.3), and in their VRS analogues, may result in their infeasibility (see, e.g., Allen et al. 1997; Pedraja-Chaparro et al. 1997). A similar problem may occur when production trade-offs are incorporated in envelopment DEA models. By duality, if a multiplier model with weight restrictions is infeasible, its dual envelopment model (which is always feasible) must have an unbounded objective function.

The unboundness of the objective function $\eta$ in the output-maximisation CRS model (5.5) and its VRS analogue indicates that the incorporation of weight restrictions (production trade-offs) has created an unlimited production of the output vector $Y_o$. (Because $\eta Y_o$ can be taken to infinity while keeping the input vector $X_o$ constant.) This is inconsistent with the established properties of production technologies (Shephard 1974; Färe et al. 1985) and indicates that an error has occurred in the construction of weight restrictions or trade-offs.

The unboundness of the objective function $\theta$ in the input-minimisation model (5.4) or its VRS analogue implies that $\theta = 0$ is feasible in the model. Consequently, the technology allows free production of the output vector $Y_o$ from the zero vector of inputs $\theta X_o = 0X_o$. This is an equally problematic situation that indicates that weight restrictions should be reconsidered.

In the author's experience based on teaching DEA to a large class of undergraduate students for many years, who were asked to use weight restrictions in their work, the above infeasibility problems are not unusual. These are more likely to happen if the model incorporates a relatively large number of weight restrictions of complex structure: those that involve several input and output weights in one inequality, as in (5.15) and (5.17). The use of trade-offs for the assessment of weight restrictions facilitates and often encourages the formulation of complex weight restrictions. For example, weight restrictions (5.15) and (5.17) that have a clear meaning as stated in Assumptions 5 and 6 are unlikely to be stated using value judgements, because it may not even be clear what they mean in value terms.

Podinovski and Bouzdine-Chameeva (2013) show that free and unlimited production of output vectors may occur *even if all multiplier models are feasible and all efficiency scores appear plausible*. In such cases, the technology is modelled incorrectly and the efficiency scores are also incorrect. One cannot therefore rely on the fact that the efficiency scores appear unproblematic—there may still be an undetected underlying problem with weight restrictions that invalidates the results of analysis and needs correcting.

Below we outline the results presented in Podinovski and Bouzdine-Chameeva (2013). These include a description of the infeasibility (and unboundness) problem caused by weight restrictions and the forms it can take, depending on the assumption of returns to scale (VRS or CRS) and the orientation of the model (input minimisation or output maximisation). This leads to the formulation of analytical and computational tests that give us a conclusive answer as to whether there is a problem with weight restrictions.

### 5.6.1 Definitions and Examples

We start with the following two definitions. Let $Y_o \in \mathbb{R}^s_+$, $Y_o \neq 0$, be a vector of outputs.

**Definition 3** Technology $\mathcal{T}$ allows *free production* of vector $Y_o$ if $(0, Y_o) \in \mathcal{T}$.

**Definition 4** Technology $\mathcal{T}$ allows *unlimited production* of vector $Y_o$ if there exists a vector of inputs $X_o$ such that $(X_o, \alpha Y_o) \in \mathcal{T}$ for all $\alpha \geq 0$.

Podinovski and Bouzdine-Chameeva (2013) prove that the above two notions are equivalent in any cone technology, e.g., in technology $\mathcal{T}_{CRS-TO}$: the existence of free production implies the existence of unlimited production, and vice versa. In a non-cone technology, e.g., in $\mathcal{T}_{VRS-TO}$, the two notions are generally different.

Furthermore, in any convex technology (e.g., in $\mathcal{T}_{CRS-TO}$ and $\mathcal{T}_{VRS-TO}$), the specification of vector $X_o$ in Definition 4 is unimportant: if vector $Y_o$ can be produced in an unlimited quantity $\alpha$ from the input vector $X_o$, then it can be produced in an unlimited quantity from the input vector $X$ of any other unit $(X, Y)$ in the technology.

It is straightforward to verify that, under the nonnegativity assumptions made about the observed DMUs, conventional CRS and VRS production technologies do not allow free or unlimited production of output vectors, but the incorporation of weight restrictions (production trade-offs) may create it. The following two examples demonstrate this effect.

**Example 4** Suppose we made a mistake in the assessment of production trade-offs (5.8) and (5.10), and stated them as follows:

$$\tilde{P}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tilde{Q}_1 = \begin{pmatrix} 4 \\ -1 \\ 0 \end{pmatrix}, \tag{5.25}$$

$$\tilde{P}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tilde{Q}_2 = \begin{pmatrix} -3 \\ 2 \\ 0 \end{pmatrix}. \tag{5.26}$$

It is easy to see that the above trade-offs induce unlimited production of the two outputs (undergraduate and master students) in the VRS and CRS technology. Figure 5.4 is a modification of Fig. 5.1 to this case.

Starting from unit $A$ and applying trade-off (5.25) 100 times, we substitute 100 master students by 400 undergraduate students. This creates point $E_1$ on the graph. Subsequently applying trade-off (5.26) 100 times, we substitute 300 undergraduate students by 200 master students. The resulting unit $A_1$ has 100 more of both types of student compared to the original unit $A$, and "achieves" this without any extra input. We can continue this process and generate a further sequence of units $A_2$, $A_3$, ..., taking the production of outputs to infinity. (The lightly shaded area in Fig. 5.4 shows the region of units dominated by $A_3$. By free disposability of output, this region is also included in the technology. As the sequence of units $A_t$, $t = 1,2,\ldots$, tends to infinity, the corresponding dominated area covers the whole nonnegative orthant.)

**Example 5** Consider the VRS technology as in Fig. 5.2. Assume we replaced the production trade-off (5.21) by the following trade-off:

$$\tilde{P} = (-1), \tilde{Q} = (-10). \tag{5.27}$$

Figure 5.5 shows the effect of trade-off (5.27) on the VRS technology. Starting at unit $A$ and consecutively applying this trade-off, we generate the line $AK$ which, together with the region below it, should be added to the technology. Note that unit $K$ has a zero input and a strictly positive output. This means that the expanded technology allows free production and indicates that trade-off (5.27) should be reconsidered.

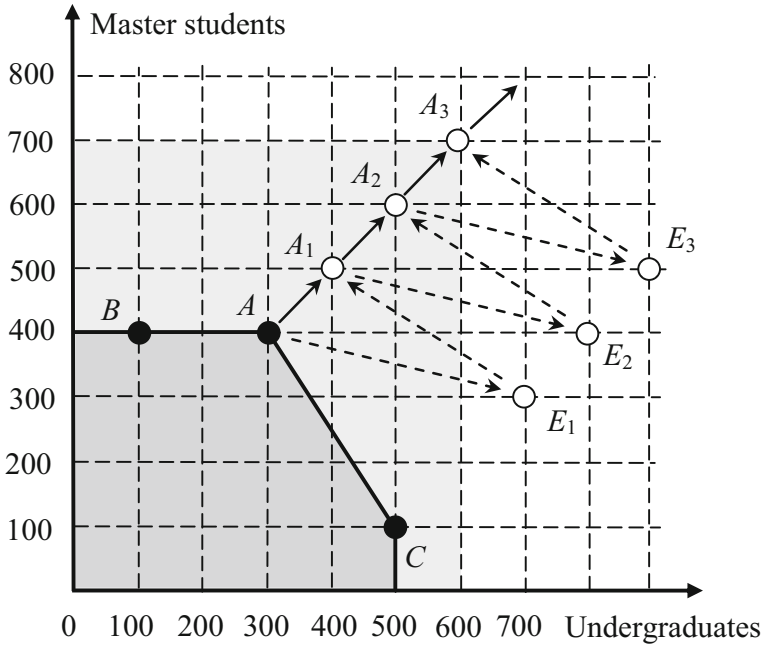



**Fig. 5.4** Free production created by two trade-offs in output dimensions

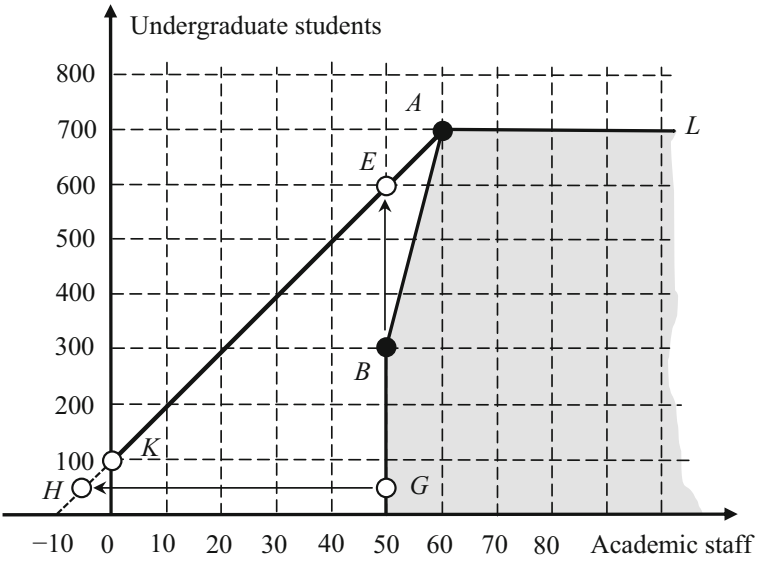



**Fig. 5.5** Free production created by the linked trade-off $(\tilde{P}, \tilde{Q}) = (-1, -10)$ in the VRS technology, and the negative "efficiency" of unit $G$

Note that the above problem cannot be observed by the efficiency calculations: the output radial efficiency of departments $A$ and $B$ in this example is equal to 1 and 0.5, respectively, and is not suspicious. However, because the slope of the efficient boundary $KA$ is incorrect, the calculated efficiencies are also incorrect.

### 5.6.2 Theoretical Results

Below we give a complete characterisation of problematic outcomes in the CRS and VRS DEA models with weight restrictions (production trade-offs) that are caused by free or unlimited production of vector $Y_o$ in the corresponding technology. If any of such outcomes are observed in practical computations, this implies that an error has occurred in the assessment of weight restrictions (or, equivalently, production trade-offs), and these need to be reconsidered.

The first theorem deals with the case of CRS.

**Theorem 5 (Podinovski and Bouzdine-Chameeva 2013)** Let $(X_o, Y_o) \in \mathcal{T}_{CRS-TO}$ and let $X_o \neq 0$ and $Y_o \neq 0$. (For example, $(X_o, Y_o)$ may be an observed unit.) Then the following three statements are equivalent:

a. There exists free and unlimited production of output vector $Y_o$ in technology $\mathcal{T}_{CRS-TO}$.
b. The CRS input-minimisation envelopment model $\mathbb{E}^1_{CRS}$ is unbounded or has a finite optimal value $\theta^* = 0$. Its dual multiplier model $\mathbb{M}^1_{CRS}$ is infeasible or has an optimal value $\theta^* = 0$, respectively.
c. The CRS output-maximisation envelopment model $\mathbb{E}^2_{CRS}$ is unbounded. Its dual multiplier model $\mathbb{M}^2_{CRS}$ is infeasible.

The next result deals with the case of VRS. Because in this technology the notions of free and unlimited production are generally not equivalent, these are considered separately.

**Theorem 6 (Podinovski and Bouzdine-Chameeva 2013)** Let $(X_o, Y_o) \in \mathcal{T}_{VRS-TO}$ and let $X_o \neq 0$ and $Y_o \neq 0$. (For example, $(X_o, Y_o)$ may be an observed unit.) Then the following statements are true:

a. There exists free production of output vector $Y_o$ in technology $\mathcal{T}_{VRS-TO}$ if and only if the VRS input-minimisation envelopment model $\mathbb{E}^1_{VRS}$ is either unbounded or has a finite optimal value $\theta^* \leq 0$. Its dual multiplier model $\mathbb{M}^1_{VRS}$ is, respectively, infeasible or has a finite optimal value $\theta^* \leq 0$.
b. There exists unlimited production of output vector $Y_o$ in technology $\mathcal{T}_{VRS-TO}$ if and only if the VRS output-maximisation multiplier model $\mathbb{E}^2_{VRS}$ is unbounded. Its dual multiplier model $\mathbb{M}^2_{VRS}$ is infeasible.

One of the differences between the cases of CRS and VRS highlighted by Theorems 5 and 6 is that free production in the VRS technology may result in a finite negative value of the input efficiency $\theta^*$. For example, consider unit $G = (50, 50)$ in the VRS

technology in Fig. 5.5. The input radial projection of $G$ is $H = (-5, 50)$. Solving the envelopment model $\mathbb{E}^1_{VRS}$ produces the finite value $\theta^* = -5/50 = -0.1$ and illustrates part (a) of Theorem 6.

The above two theorems do not solve the problem of identifying problematic weight restrictions (trade-offs) completely: even if no problematic outcomes occur with the assessment of all observed units $(X_j, Y_j)$, this guarantees only that there is no free or unlimited production of the output vectors $Y_j$ of observed units. This does not however guarantee that there is no free or unlimited production of other output vectors in the technology. For example, in the case of VRS technology in Fig. 5.5, Theorem 6 would not identify any problem when the input or output radial efficiency of both observed units $A$ and $B$ is assessed.

Podinovski and Bouzdine-Chameeva (2013) suggest two approaches, analytical and computational, that allow us to examine if the incorporation of production trade-offs (weight restrictions) has induced free or unlimited production in the technology. This task is simplified by the following statement.

**Theorem 7 (Podinovski and Bouzdine-Chameeva 2013)** The existence of free (and therefore unlimited production) of the output vector $Y_o$ in technology $\mathcal{T}_{CRS-TO}$ is equivalent to the existence of either free or unlimited production of vector $Y_o$ (but not necessarily both) in technology $\mathcal{T}_{VRS-TO}$.

According to Theorem 7, if there is a problem with free or unlimited production in either CRS or VRS technology, then there is a similar problem in the other. Because the notions of free and unlimited production are equivalent in the CRS technology, and also because the choice of vector $X$ is unimportant for the latter notion, it suffices to test for the existence of unlimited production with the input vector $X$ of an arbitrary unit $(X, Y)$ in the CRS technology $\mathcal{T}_{CRS-TO}$.

Podinovski and Bouzdine-Chameeva (2013) consider two cases. The simpler case arises if weight restrictions (5.1) are not linked. In this case the testing is reduced to verifying a simple algebraic condition. If weight restrictions (5.1) include linked restrictions, the testing is performed by solving specially constructed linear programs. Below we outline the two cases.

### 5.6.3 *Free Production with Not Linked Trade-offs*

The most straightforward case arises if the weight restrictions are not linked. Then (5.1) can be restated as follows:

$$u^\top Q_t \le 0, \quad t = 1, \ldots, K_1, \tag{5.28}$$

$$-v^\top P_t \le 0, \quad t = 1, \ldots, K_2. \tag{5.29}$$

**Theorem 8 (Podinovski and Bouzdine-Chameeva 2013)** Technology $\mathcal{T}_{CRS-TO}$ does not allow free (and unlimited) production if and only if both of the following two conditions are satisfied:

a. there exists a strictly positive vector $u^* > 0$ that satisfies (5.28);
b. there exists a nonnegative vector $v^* \geq 0$ that satisfies (5.29) such that $(v^*)^\top X_j > 0$ holds for all observed units $j = 1,\ldots, n$.

(If either group of weight restrictions (5.28) or (5.29) is missing, then the corresponding condition (a) or (b) is removed from the above statement.)

Note that the vectors $u^*$ and $v^*$ do not need to satisfy the conditions of models (5.2) or (5.3)—all that is required is that such vectors satisfy (5.28) and (5.29).

In practical applications all inputs of all observed DMUs $j = 1,\ldots, n$ are usually strictly positive. In this case condition (b) of Theorem 8 is equivalent to the simpler condition: there exists a nonzero vector $v^* \geq 0$ that satisfies (5.29).

If some of the inputs of observed DMUs are equal to zero, the above simplified condition does not apply. However, to prove that there is no free production, a simpler *sufficient condition* may be used. (Obviously, if it is not satisfied, this does not mean that there is free production—we need to use Theorem 8 for a definitive answer.)

**Corollary 1** If there exist strictly positive vectors $u^* > 0$ and $v^* > 0$ that satisfy (5.28) and (5.29), then technology $\mathcal{T}_{CRS-TO}$ does not allow free (and unlimited) production.

As an illustration, refer to Example 1 in which we used the trade-offs between undergraduate and master students as stated in (5.8) and (5.10). The resulting technology was illustrated in Fig. 5.1. The corresponding weight restrictions (5.9) and (5.11) are simultaneously satisfied, for example, by strictly positive weights $u_1 = u_2 = 1$. This means that condition (a) of Theorem 8 is true. Because there are no weight restrictions involving input weights, condition (b) of Theorem 8 should be ignored. By Theorem 8 or its Corollary 1, the two trade-offs (5.8) and (5.10) do not cause free or unlimited production in either CRS or VRS technology, which is consistent with Fig. 5.1.

Let us illustrate how Theorem 8 can be used to detect free production when it exists, even if all efficiency scores appear unproblematic.

**Example 6** In Example 4 we showed how the use of trade-offs (5.25) and (5.26) resulted in the unlimited production of two outputs (undergraduate and master students). If we use the same two trade-offs with the data set in Table 5.1, they induce unlimited production of the two outputs in the same way but the problem *is not observed* from the efficiency calculations and becomes hidden.

Table 5.6 shows the efficiency scores (in %) in the CRS and VRS DEA models for the departments as in Table 5.1. Both the CRS and VRS models incorporate only two production trade-offs (5.25) and (5.26). (These models are obtained from the models CRS 2 and VRS 2 discussed above in which the "good" trade-offs (5.8) and (5.10) are replaced by the problematic trade-offs (5.25) and (5.26).)

Note that the results of computations in Table 5.6 do not appear problematic—the only exception may be the unusually low "efficiency" of department D2 in both models. In such cases it is easy to miss the underlying problem. To see if there is a problem we use Theorem 8 and restate production trade-offs (5.25) and (5.26) as the weight restrictions

$$4u_1 - u_2 \leq 0, \tag{5.30}$$

**Table 5.6** Output radial efficiency (%) of departments in the CRS and VRS models with trade-offs (5.25) and (5.26) causing free production

| Department | CRS | VRS |
|---|---|---|
| D1 | 75.64 | 91.09 |
| D2 | 23.14 | 23.55 |
| D3 | 65.62 | 66.67 |
| D4 | 100.00 | 100.00 |
| D5 | 100.00 | 100.00 |
| D6 | 86.00 | 100.00 |
| D7 | 100.00 | 100.00 |

$$-3u_1 + 2u_2 \leq 0. \tag{5.31}$$

It is straightforward to show that the above inequalities cannot be satisfied by strictly positive weights $u_1$ and $u_2$. Indeed, adding the two inequalities (5.30) and (5.31), we obtain $u_1 + u_2 \leq 0$, which does not allow a strictly positive solution vector. By Theorem 8, production trade-offs (5.25) and (5.26) induce free (and unlimited) production in the CRS technology, and the CRS efficiency scores are, although plausible, obviously meaningless. By Theorem 7, the efficiency scores in the VRS model are also incorrect.

### *5.6.4 Free Production with Linked Trade-offs*

In the general case of linked weight restrictions (5.1) Podinovski and Bouzdine-Chameeva (2013) develop two computational procedures to test if there is free (and unlimited) production in the CRS technology. Below we describe one of them.

The idea of this method is simple and based on the following fact: technology $\mathcal{T}_{CRS-TO}$ allows an unlimited production of a vector $Y_o$ *if and only if* it allows an unlimited production of each of its individual *positive* outputs, provided all the other individual outputs are taken equal to zero. (The "only if" part of this statement is obvious. The "if" part follows from the following. Suppose the technology allows the production of each individual output $(Y_o)_r$, $r = 1, \ldots, s$, in any proportion $\alpha \geq 0$, from the input vector $X_o$. Then the simple average of all $s$ such units, each producing the single output $\alpha(Y_o)_r$, is the unit $(X_o, (\alpha/s)Y_o) \in \mathcal{T}_{CRS-TO}$. Because $s$ is constant and $\alpha$ is arbitrarily large, technology $\mathcal{T}_{CRS-TO}$ allows an unlimited production of vector $Y_o$.)

The above suggests that we can test for unlimited production as follows. First, we select any (e.g., observed) unit $(X_o, Y_o) \in \mathcal{T}_{CRS-TO}$ such that all components of vector $Y_o$ are strictly positive: $(Y_o)_r > 0$, for all $r = 1, \ldots, s$. If no such observed unit exists, we can always take the simple average of all observed units. Because

each output $r$ is strictly positive for at least one observed unit $j$, the average of all observed units will have a strictly positive output vector.

Define $s$ artificial output vectors $U_r, r = 1, \ldots, s$, as follows. Each of these vectors has only one positive component:

$$U_1 = ((Y_o)_1, 0, \ldots, 0), \ \ldots \ , U_s = (0, \ldots, 0, (Y_o)_s)^\top.$$

Consider $s$ DMUs in the form $(X_o, U_r)$, where $r = 1, \ldots, s$. Each of such units is dominated by the original unit $(X_o, Y_o)$ and therefore $(X_o, U_r) \in \mathcal{T}_{CRS-TO}$. We can now expand the set of observed DMUs $J$ by incorporating the above $s$ artificial units. Because the latter units are dominated, the technology $\mathcal{T}_{CRS-TO}$ remains unchanged.

We now solve $s$ output-maximisation multiplier models, one for each unit $(X_o, U_\rho)$, $\rho = 1, \ldots, s$. (We use index $\rho$ to differentiate from $r$ in the same formulation.)

$$\eta^* = \min \quad v^\top X_o, \tag{5.32}$$

subject to

$$\begin{aligned} & u^\top U_\rho = 1, \\ & u^\top Y_j - v^\top X_j \le 0, \quad j = 1, \ldots, n, \\ & u^\top U_r - v^\top X_o \le 0, \quad r = 1, \ldots, s, \\ & u^\top Q_t - v^\top P_t \le 0, \quad t = 1, \ldots, K, \\ & u, v \ge 0. \end{aligned}$$

**Theorem 9 (Podinovski and Bouzdine-Chameeva 2013)** Technology $\mathcal{T}_{CRS-TO}$ allows free (and unlimited) production if and only if there exists a $\rho = 1, \ldots, s$ such that the multiplier model (5.32) is infeasible.

Obviously, instead of model (5.32), we can solve its dual envelopment model. In this case the infeasibility of model (5.32) is equivalent to the unboundness of the envelopment model. Also note that the constraints $u^\top U_r - v^\top X_o \le 0$ in model (5.32) are redundant and can in principle be removed because, as discussed, units $(X_o, U_r)$ are dominated. From the practical point of view, however, it may be beneficial to keep model (5.32) as stated, because in this case it can be solved by standard DEA solvers.

**Example 7** As an illustration, consider the university departments in Table 5.1 and the eight trade-offs discussed in Sect. 5.3. Because some of these trade-offs are linked, we use the method based on program (5.32) to verify that the combination *of this particular data set* and trade-offs does not induce free or unlimited production.

As the starting point, let us choose department D1 as the unit $(X_o, Y_o)$. (Alternatively, we can choose any department from D1 to D6 for this purpose, but not D7 because its second output is zero.) Following the above procedure, define three artificial units with the vector of inputs $X_o = (92,15)^\top$ as in department D1, and the following different output vectors:

$$U_1 = (800, 0, 0)^\top, U_2 = (0, 200, 0)^\top, U_3 = (0, 0, 90)^\top.$$

We now add the three units $(X_o, U_1)$, $(X_o, U_2)$ and $(X_o, U_3)$ to the set of departments D1–D7. Because all three additional departments are dominated by D1, the technology does not change. Finally, we assess the output radial efficiency of the three additional departments in the CRS multiplier model (5.32). The corresponding three optimal values of program (5.32) are finite and equal, respectively, to 3.632, 6.041 and 2.037. (The output radial efficiency of the three artificial units is, respectively, 0.2753, 0.1655 and 0.4909. The output radial efficiency of departments D1–D7, if calculated simultaneously by the software, is the same as without the additional three units.) By Theorem 9, the CRS (and consequently VRS) technology based on the data set in Table 5.1 and the eight trade-offs does not allow free or unlimited production.

## 5.7 Solving DEA Models with Production Trade-offs

Conventional CRS and VRS DEA models (without weight restrictions) are usually solved using either a two-stage computational procedure or an analogous single-stage method utilizing a non-Archimedean $\varepsilon$ (in practice taken equal to a very small positive number). These methods are summarized in Thanassoulis et al. (2008) and Cooper et al. (2011b).

In many applications of DEA only the radial efficiency of the DMUs is of interest, and the first stage of the two-stage method suffices for this purpose. It identifies the radial projection of the assessed DMU on the boundary of the VRS or CRS technology and produces the DMU's radial input or output efficiency. Because the radial projection of an inefficient DMU may be only *weakly* efficient, the identification of its efficient target (in the Pareto sense) requires the second optimisation stage in which the sum of input and output slacks is maximised. Performing the second stage identifies the efficient target of the DMU and the reference set of its efficient peers. The latter are the observed DMUs $j$ that have a corresponding multiplier $\lambda_j > 0$ in the optimal solution to the second-stage linear program.

Podinovski (2007b) shows that the application of the standard second stage to DEA models with weight restrictions (or production trade-offs) may result in a target unit with meaningless negative values of some inputs. (This is unrelated to the issue of inconsistent weight restrictions discussed in the previous section.) In the suggested corrected procedure, the conventional second stage is split into two new stages, and the complete solution method becomes a three-stage procedure. Depending on the purpose of a DEA study, only the first, two first or all three computational stages may need to be performed.

Below we outline these three stages. We assume that the weight restrictions (production trade-offs) have already been checked using the methods described in the

previous section, and that the underlying VRS or CRS technology does not allow free or unlimited production of non-zero output vectors.

*Stage 1 (Assessing the radial efficiency)* This task is straightforward and requires the solution of the appropriate CRS or VRS envelopment model, or their dual multiplier forms, as stated in Sect. 5.2.

*Stage 2 (Identifying efficient targets)* An efficient target of DMU $(X_o, Y_o)$ is obtained by solving the specially constructed additive DEA model formulated in Sect. 5.7.2 below.

*Stage 3 (Identifying reference sets of efficient peer units)* This stage is required because, even if the multiplier $\lambda_j$ is strictly positive in an optimal solution to the model used at Stage 2, the corresponding observed DMU $j$ may be inefficient. An example of this is given in Podinovski (2007b). The linear program solved at Stage 3 is presented in Sect. 5.7.3.

### 5.7.1 Stage 1: Assessing the Radial Efficiency

Most applications of DEA are concerned only with the input or output radial efficiency of the units. In such applications this stage is the only one that needs performing. Depending on the assumption of CRS or VRS and the orientation of the model (input minimisation or output maximisation), the radial efficiency of DMU $(X_o, Y_o)$ is assessed by solving the corresponding envelopment (or multiplier) model stated in Sect. 5.2.

This stage also identifies the radial projection (target) unit $(X^*, Y^*)$ of the DMU $(X_o, Y_o)$. In the case of input minimisation, $(X^*, Y^*) = (\theta^* X_o, Y_o)$, where $\theta^*$ is the input radial efficiency of DMU $(X_o, Y_o)$. In the case of output maximisation, $(X^*, Y^*) = (X_o, \eta^* Y_o)$, where $\eta^*$ is the inverse output radial efficiency of DMU $(X_o, Y_o)$. (The value $\eta^*$ is the optimal value in the corresponding envelopment and multiplier models that is inverse to the output efficiency measure.)

### 5.7.2 Stage 2: Identifying Efficient Targets

As in the case of conventional CRS and VRS DEA models, this stage should be performed only if we need to identify efficient targets of individual DMUs. In particular, the computations at this stage do not alter the radial efficiency assessed at Stage 1.

The need of the second stage arises because the radial target $(X^*, Y^*)$ assessed at Stage 1 may be a weakly efficient unit and not efficient in the Pareto sense. The conventional second optimisation stage aims at maximising the sum of input and output slacks that improve the unit $(X^*, Y^*)$. The same idea is applicable to DEA models with weight restrictions (production trade-offs), but an additional care has

to be taken of the nonnegativity of inputs in the resulting efficient unit (which is automatically maintained in the standard models without weight restrictions).

The following program identifies possible individual improvements to the inputs and outputs of the unit $(X^*, Y^*)$:

$$\sigma^* = \max \quad \sum_{r=1}^{s} \varepsilon_r + \sum_{i=1}^{m} \delta_i, \text{ subject to } (X^* - \delta, Y^* + \varepsilon) \in \mathcal{T}, \tag{5.33}$$

where $\varepsilon \in \mathbb{R}^s_+$, $\delta \in \mathbb{R}^m_+$, and technology $\mathcal{T}$ is either $\mathcal{T}_{CRS-TO}$ or $\mathcal{T}_{VRS-TO}$.

To be specific, consider the case of CRS. Based on Theorem 1, program (5.33) takes on the form

$$\sigma^* = \max \quad \sum_{r=1}^{s} \varepsilon_r + \sum_{i=1}^{m} \delta_i, \tag{5.34a}$$

$$\text{subject to} \quad \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t - e = Y^* + \varepsilon, \tag{5.34b}$$

$$\sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t + d = X^* - \delta, \tag{5.34c}$$

$$Y^* + \varepsilon \geq 0, \tag{5.34d}$$

$$X^* - \delta \geq 0, \tag{5.34e}$$

$$\lambda, \pi, e, d, \varepsilon, \delta \geq 0. \tag{5.34f}$$

Note that program (5.34) can be simplified. First, at any of its optimal solutions the vector $e$ must be a zero vector. Indeed, if we assume the converse ( $e \geq 0$ and $e \neq 0$) then redefining $\tilde{e} = 0$ and $\tilde{\varepsilon} = \varepsilon + e$ keeps (5.34b) true and improves the objective function (5.34a), which is impossible due to the assumed optimality of the current solution. Therefore, vector $e$ in program (5.34) can be assumed zero and removed from the formulation. Second, condition (5.34d) is redundant because both vectors $Y^*$ and $\varepsilon$ are nonnegative.

The resulting model is as follows:

$$\sigma^* = \max \quad \sum_{r=1}^{s} \varepsilon_r + \sum_{i=1}^{m} \delta_i, \tag{5.35a}$$

$$\text{subject to} \quad \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t = Y^* + \varepsilon \tag{5.35b}$$

$$\sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t + d = X^* - \delta, \tag{5.35c}$$

$$X^* - \delta \geq 0, \tag{5.35d}$$

$$\lambda, \pi, d, \varepsilon, \delta \geq 0. \tag{5.35e}$$

Model (5.3) is the same as model (6) stated in Podinovski (2007b). In the latter model the above condition (5.35d) is replaced by an equivalent requirement that the expression on the left-hand side of equality (5.35c) is nonnegative.

As already stated, we assume that technology $\mathcal{T}_{CRS-TO}$ does not allow free and unlimited production. Therefore the objective function (5.35a) is bounded above, and there exists an optimal solution to program (5.35) that we denote

$$\lambda', \pi', d', \varepsilon', \delta'. \tag{5.36}$$

This defines the efficient target of DMU $(X_o, Y_o)$ as

$$(X', Y') = (X^* - \delta', Y^* + \varepsilon'). \tag{5.37}$$

By the conditions of model (5.35), $(X', Y') \in \mathcal{T}_{CRS-TO}$.

**Theorem 10 (Podinovski 2007b)** DMU $(X', Y')$ in (5.37) is efficient in technology $\mathcal{T}_{CRS-TO}$.

Obviously, if all optimal slacks in (5.35), and hence the optimal value $\sigma^*$, are equal to zero, the efficient target $(X', Y')$ coincides with the radial target $(X^*, Y^*)$. In particular, DMU $(X_o, Y_o)$ is efficient if and only if $(X_o, Y_o) = (X', Y')$.

In the case of VRS, model (5.35) requires an additional normalising condition (5.7). The same formula (5.37) defines the efficient target $(X', Y')$ in this case.

Note that the inequality (5.35d) in model (5.35) guarantees that the maximisation of the sum of component slacks (5.35a) is performed within the technology by requiring that inputs remain nonnegative. As shown by example in Podinovski (2007b), the simple maximisation of the sum of slacks without condition (5.35d) (in this case $d$ could be assumed to be a zero vector) may result in negative values of some of the inputs.

**Remark 2** Model (5.35) is an additive CRS DEA model based on technology $\mathcal{T}_{CRS-TO}$. It assesses the efficiency of the unit $(X^*, Y^*)$ by maximising the sum of component slacks $\varepsilon_r$ and $\delta_i$, provided the resulting unit remains within the technology (and, in particular, does not have negative inputs). In the case of VRS, we need to add the normalising condition (5.7) to the constraints of model (5.35).

Model (5.35) and its VRS variant become standard additive DEA models (Charnes et al. 1985) in the absence of trade-offs (5.6). Indeed, in this case the trade-off terms on the left-hand side of conditions (5.35b) and (5.35c) are omitted. Furthermore, the maximisation of the sum of slack variables in (5.35) implies that at optimality $d = 0$, and therefore vector $d$ can be removed from the formulation. Finally, the nonnegativity condition (5.35d) is redundant because, in the absence of trade-offs, it follows from (5.35c).

Like conventional additive DEA models, model (5.35) and its VRS variant can be used independently for the assessment of efficiency of any unit $(X^*, Y^*) \in \mathcal{T}_{CRS-TO}$, without the need to perform the first (radial projection) optimisation stage.

### 5.7.3 Stage 3: Identifying Reference Sets of Efficient Peer Units

In conventional DEA models without weight restrictions (production trade-offs), the reference set of efficient peers consists of the observed DMUs $j$ such that $\lambda_j > 0$ in an optimal solution to the second-stage optimisation model. In a DEA model with weight restrictions, an observed DMU $j$ with a strictly positive value $\lambda'_j$ in the optimal solution (5.36) may be inefficient—an example of this is given in Podinovski (2007b). As proved, in this case there exists an alternative optimal solution to program (5.35) that results in the same efficient target $(X', Y')$ and for which the condition $\lambda_j > 0$ implies that the observed unit $j$ is efficient, for all $j$. Identifying such an optimal solution to (5.35) requires solving another linear program.

As with Stage 2, the computations of Stage 3 should be performed only if needed. These computations do not affect the radial efficiency, radial targets and efficient targets already obtained at Stages 1 and 2.

Following Podinovski (2007b), efficient peers of DMU $(X_o, Y_o)$ corresponding to the efficient target $(X', Y')$ can be obtained by maximising the sum of components of vector $d$ as the secondary goal in program (5.35), while keeping vectors $\varepsilon'$ and $\delta'$ at their optimum level as in (5.36). In this case, by (5.37), the constant vectors $Y^* + \varepsilon'$ and $X^* - \delta'$ on the right-hand side of conditions (5.35b) and (5.35c) can be replaced by $Y'$ and $X'$, respectively. The resulting model takes on the form:

$$D^* = \max \quad \sum_{i=1}^{m} d_i, \tag{5.38a}$$

$$\text{subject to} \quad \sum_{j=1}^{n} \lambda_j Y_j + \sum_{t=1}^{K} \pi_t Q_t = Y', \tag{5.38b}$$

$$\sum_{j=1}^{n} \lambda_j X_j + \sum_{t=1}^{K} \pi_t P_t + d = X', \tag{5.38c}$$

$$\lambda, \pi, d \geq 0. \tag{5.38d}$$

Note that the inequality (5.35d) no longer contains decision variables (because the vector $\delta = \delta'$ is kept constant) and is omitted as redundant in program (5.38).

Because the objective function of program (5.38) is bounded above, there exists an optimal solution $\tilde{\lambda}, \tilde{\pi}, \tilde{d}$ to this program. Taken together with the constant vectors $\varepsilon'$ and $\delta'$, solution

$$\tilde{\lambda}, \tilde{\pi}, \tilde{d}, \varepsilon', \delta' \tag{5.39}$$

is an optimal solution to program (5.35). If the optimal solution (5.36) to program (5.35) is unique, then (5.39) is the same as (5.36). Otherwise, (5.39) is an optimal solution to (5.35) that additionally maximises the sum of components of vector $d$ as in (5.38a).

**Theorem 11 (Podinovski 2007b)** If $\tilde{\lambda}_j > 0$ then DMU $j$ is efficient in technology $\mathcal{T}_{CRS-TO}$ and, consequently, in the smaller standard CRS technology $\mathcal{T}_{CRS} \subset \mathcal{T}_{CRS-TO}$.

An alternative model to (5.38) is obtained in Podinovski (2000). It has the same objective (5.38a) as above maximised over the set of constraints (5.35b–e), with the additional condition

$$\sum_{r=1}^{s} \varepsilon_r + \sum_{i=1}^{m} \delta_i = \sigma^*, \tag{5.40}$$

and keeping vectors $\varepsilon$ and $\delta$ variable.

The difference between model (5.38) and the latter model is that, by solving the former, we identify the reference sets for DMU $(X_o, Y_o)$ that are used in the composition of its specific efficient target $(X', Y')$ which is fixed. In the latter approach, the efficient target is not fixed. The model based on condition (5.40) generally has alternative optima $\lambda'', \pi'', d'', \varepsilon'', \delta''$, each identifying a generally different efficient target $(X'', Y'')$ and the corresponding reference set of efficient peers $j$.

The above results extend to the case of VRS with obvious modifications. As noted, in the case of VRS model (5.35) incorporates the additional normalising equality (5.7). The latter should also be incorporated in (5.38). Let

$$\hat{\lambda}, \hat{\pi}, \hat{d} \tag{5.41}$$

be an optimal solution to program (5.38) with the condition (5.7).

**Theorem 12 (Podinovski 2007b)** If $\hat{\lambda}_j > 0$ then DMU $j$ is efficient in technology $\mathcal{T}_{VRS-TO}$ and, consequently, in the smaller standard VRS technology $\mathcal{T}_{VRS} \subset \mathcal{T}_{VRS-TO}$.

The above theorem implies the existence of at least one efficient DMU $j \in J$ in any technology $\mathcal{T}_{VRS-TO}$ (under the assumption that there is no free or unlimited production, as stated beforehand).

**Corollary 2** In any technology $\mathcal{T}_{VRS-TO}$, there exists at least one efficient observed DMU.

**Proof of Corollary 2** Because of condition (5.7), in solution (5.41) there exists a $j$ such that $\hat{\lambda}_j > 0$. By Theorem 12, DMU $j$ is efficient in technology $\mathcal{T}_{VRS-TO}$. ■

Note that Corollary 2 does not unconditionally extend to the case of CRS. According to Theorem 1 stated in Charnes et al. (1990), there exists at least one efficient observed DMU in the CRS technology $\mathcal{T}_{CRS-TO}$, under the condition that weight restrictions (5.1) are not linked. The following example shows that the same statement is generally not true in the case of linked weight restrictions.

**Example 8** Consider CRS technology $\mathcal{T}_{CRS-TO}$ discussed in Example 3 and illustrated in Fig. 5.3. The only observed unit $A = (2,1)$ is inefficient in the CRS technology induced by itself and the single linked production trade-off $(P, Q) = (1,2)$. (The latter is equivalent to the linked weight restriction $2u - v \leq 0$.) Therefore, there are no efficient observed units in technology $\mathcal{T}_{CRS-TO}$ in Fig. 5.3. Furthermore, the

output radial efficiency of $A$ is equal to 0.25. Its unique efficient target is (2,4)—it is constructed entirely from the above trade-off $(P, Q)$ applied 4 times to the origin, with no contribution from the unit $A$ itself. Therefore, unit $A$ has no efficient peers among observed units, and the efficient target is composed entirely from the production trade-off. Finally note that $A$ is efficient in the VRS technology $\mathcal{T}_{VRS-TO}$ in Fig. 5.3, which is consistent with Corollary 2.

## 5.8 Conclusion

In this chapter we presented the notion of production trade-offs as the dual forms of weight restrictions. We explored various theoretical, methodological and computational issues arising from the application of production trade-offs in DEA models.

Although production trade-offs are mathematically equivalent to weight restrictions, the assessment of the former is conducted in the language of possible changes to the inputs and outputs in the technology. In contrast, the assessment of weight restrictions often involves value judgements that are more managerial in nature and not directly related to the technological possibilities.

Based on the results of this chapter, the following standard workflow can be suggested for the practical implementation of production trade-offs and weight restrictions. This consists of three steps that may need to be repeated iteratively as the model is being modified by the incorporation of additional trade-offs.

1. *Construction of production trade-offs and weight restrictions*. As illustrated in Sect. 5.3, production trade-offs should represent realistic assumptions about the technology. In practice, we should be certain that all observed DMUs would be willing to accept the simultaneous changes stated by the trade-offs.
2. *Verification that the trade-offs (or weight restrictions) do not generate free or unlimited production for the given set of observed DMUs*. As discussed in Sect. 5.6, this stage is important because, if there is free or unlimited production in the technology, the results of the next stage may be inconsistent and puzzling. Alternatively, such results may appear unproblematic but still be erroneous. This stage requires either the checking of simple inequalities or, in the case of linked weight restrictions, the use of standard DEA software with the extended set of observed DMUs.
3. *Computation of efficiency, efficient targets and efficient peers*. There are three stages in the computational procedure described in Sect. 5.7. In many practical applications only the first stage would be needed and may be performed using standard DEA software. The implementation of Stages 2 and 3 would currently require the use of general linear solvers.

The use of production trade-offs in DEA models, or the use of weight restrictions obtained from such trade-offs, is interesting for a number of reasons.

First, production trade-offs allow us to specify additional information about the technology that is not otherwise captured by the observed data and standard production assumptions. This leads to *a meaningful extension* of the conventional CRS or VRS production technology and results in a better-informed model of the production process. Furthermore, this generally improves the efficiency discrimination of the model in a technologically meaningful way.

Second, the use of production trade-offs or weight restrictions based on them does not have the well-known drawback of weight restrictions assessed by other methods. The use of the latter generally leads to an uncontrolled expansion of the model of technology. In particular, the value judgements used in the construction of weight restrictions cannot generally explain the technological meaning of the expanded technology and its new efficient frontier. As a result, the radial and efficient targets of inefficient units may not be producible. The meaning of radial efficiency as the ultimate and technologically feasible improvement factor is no longer preserved. In contrast, the assessment of production trade-offs explicitly takes into account the meaning of the resulting expansion of the technology. The use of such trade-offs or weight restrictions based on them preserves the traditional meaning of efficiency.

Third, because the use of production trade-offs results in a meaningful model of production technology, the well-established notions of productivity analysis such as returns to scale, productivity change, and other can be extended to it in a straightforward fashion. In particular, the former can be explored by the generic method of reference technologies developed by Färe et al. (1985) and further explored by Podinovski (2004c). The Malmquist productivity index in models with production trade-offs was discussed in Alirezaee and Afsharian (2010).

Fourth, the clear technological meaning of production trade-offs allows us to make relatively complex statements involving several inputs and outputs in a single trade-off or weight restriction. Examples of such statements were production trade-offs (5.14) and (5.16) and the corresponding weight restrictions (5.15) and (5.17). An advantage of such complex production trade-offs is that they generally add more points to the model of production technology than simple statements, and therefore contribute to better efficiency discrimination. It is unlikely that weight restrictions (5.15) and (5.17) could be obtained using value judgements.

Fifth, production trade-offs can be used in DEA models that do not have dual multiplier forms. An example of this is the FDH technology.

Sixth, the interpretation of weight restrictions as the dual forms of production trade-offs allows us to clarify and resolve some theoretical, methodological and computational issues arising in the context of weight restrictions. For example, as discussed, the interpretation of weight restrictions in terms of production trade-offs gives a positive answer to the long-standing question of applicability of weight restrictions in the VRS technology. In Sect. 5.6 we showed that the notion of trade-offs is instrumental in understanding the infeasibility and related problems in DEA models with weight restrictions.

## References

Alirezaee MR, Afsharian M (2010) Improving the discrimination of data envelopment analysis in multiple time periods. Int Trans Oper Res 17(5):667–679

Allen R, Athanassopoulos A, Dyson RG, Thanassoulis E (1997) Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. Ann Oper Res 73:13–34

Amado CAF, Dyson RG (2009) Exploring the use of DEA for formative evaluation in primary diabetes care: an application to compare English practices. J Oper Res Soc 60(11):1469–1482

Amado CAF, Santos SP (2009) Challenges for performance assessment and improvement in primary health care: the case of the Portuguese health centres. Health Policy 91(1):43–56

Atici KB (2012) Using data envelopment analysis for the efficiency and elasticity evaluation of agricultural farms. Ph.D. thesis. University of Warwick, UK

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30(9):1078–1092

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Charnes A, Cooper WW, Golany B, Seiford L (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econom 30(1–2):91–107

Charnes A, Cooper WW, Wei QL, Huang ZM (1989) Cone ratio data envelopment analysis and multi-objective programming. Int J Syst Sci 20(7):1099–1118

Charnes A, Cooper WW, Huang ZM, Sun DB (1990) Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econom 46(1–2):73–91

Cook WD, Zhu J (2008) Context-dependent assurance regions in DEA. Oper Res 56(1):69–78

Cooper WW, Ruiz JL, Sirvent I (2011a) Choices and uses of DEA weights. In: Cooper WW, Seiford LM, Zhu J (eds) Handbook on data envelopment analysis, 2nd edn. Springer Science + Business Media, New York, pp 93–126

Cooper WW, Seiford LM, Zhu J (2011b) Data envelopment analysis: history, models, and interpretations. In: Cooper WW, Seiford LM, Zhu J (eds) Handbook on data envelopment analysis, 2nd edn. Springer Science + Business Media, New York, pp 1–39

Deprins D, Simar L, Tulkens H (1984) Measuring labor-efficiency in post offices. In: Marchand M, Pestieau P, Tulkens H (eds) The performance of public enterprises. Elsevier Science, Amsterdam, pp 243–267

Dyson RD, Thanassoulis E (1988) Reducing weight flexibility in data envelopment analysis. J Oper Res Soc 39(6):563–576

Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer Academic, Boston

Førsund FR (2013) Weight restrictions in DEA: misplaced emphasis? J Prod Anal 40(3):271–283

Halme M, Korhonen P (2000) Restricting weights in value efficiency analysis. Eur J Oper Res 126(1):175–188

Jones CN, Kerrigan EC, Maciejowski JM (2008) On polyhedral projection and parametric programming. J Optim Theory Appl 138(2):207–220

Khalili M, Camanho AS, Portela MCAS, Alirezaee MR (2010) The measurement of relative efficiency using data envelopment analysis with assurance regions that link inputs and outputs. Eur J Oper Res 203(3):761–770

Pedraja-Chaparro F, Salinas-Jimenez J, Smith P (1997) On the role of weight restrictions in data envelopment analysis. J Prod Anal 8(2):215–230

Podinovski VV (1999) Side effects of absolute weight bounds in DEA models. Eur J Oper Res 115(3):583–595

Podinovski VV (2000) Weight restrictions and production trade-offs in DEA models. Working paper no. 330. Warwick Business School. http://www.wbs.ac.uk/downloads/working_papers/330.pdf

Podinovski VV (2004a) Production trade-offs and weight restrictions in data envelopment analysis. J Oper Res Soc 55(12):1311–1322

Podinovski VV (2004b) Suitability and redundancy of non-homogeneous weight restrictions for measuring the relative efficiency in DEA. Eur J Oper Res 154(2):380–395
Podinovski VV (2004c) Efficiency and returns to scale on the "no free lunch" assumption only. J Prod Anal 22(3):227–257
Podinovski VV (2005) The explicit role of weight bounds in models of data envelopment analysis. J Oper Res Soc 56(12):1408–1418
Podinovski VV (2007a) Improving data envelopment analysis by the use of production trade-offs. J Oper Res Soc 58(10):1261–1270
Podinovski VV (2007b) Computation of efficient targets in DEA models with production trade-offs and weight restrictions. Eur J Oper Res 181(2):586–591
Podinovski VV, Athanassopoulos AD (1998) Assessing the relative efficiency of decision making units using DEA models with weight restrictions. J Oper Res Soc 49(5):500–508
Podinovski VV, Bouzdine-Chameeva T (2013) Weight restrictions and free production in data envelopment analysis. Oper Res 61(2):426–437
Roll Y, Cook WD, Golany B (1991) Controlling factor weights in data envelopment analysis. IIE Trans 23(1):2–9
Santos SP, Amado CAF, Rosado JR (2011) Formative evaluation of electricity distribution utilities using data envelopment analysis. J Oper Res Soc 62(7):1298–1319
Schaffnit C, Rosen D, Paradi JC (1997) Best practice analysis of bank branches: an application of DEA in a large Canadian bank. Eur J Oper Res 98(2):269–289
Shephard RW (1974) Semi-homogeneous production functions and scaling of production. In: Eichhorn W, Henn R, Opitz O, Shephard RW (eds) Production theory. Springer, New York, pp 253–285
Thanassoulis E, Allen R (1998) Simulating weights restrictions in data envelopment analysis by means of unobserved DMUs. Manage Sci 44(4):586–594
Thanassoulis E, Dyson RG, Foster MJ (1987) Relative efficiency assessments using data envelopment analysis: an application to data on rates departments. J Oper Res Soc 38(5):397–411
Thanassoulis E, Portela MC, Allen R (2004) Incorporating value judgements in DEA. In: Cooper WW, Seiford LM, Zhu J (eds) Handbook on data envelopment analysis. Kluwer Academic, Boston, pp 99–138
Thanassoulis E, Portela MCS, Despić O (2008) Data envelopment analysis: the mathematical programming approach to efficiency analysis. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press, New York, pp 251–420
Thompson RG, Langemeier LN, Lee CT, Lee E, Thrall RM (1990) The role of multiplier bounds in efficiency analysis with application to Kansas farming. J Econom 46(1–2):93–108

# Chapter 6
# Facet Analysis in Data Envelopment Analysis

**Ole B. Olesen and Niels Chr. Petersen**

**Abstract** Data Envelopment Analysis (DEA) employs mathematical programming to measure the relative efficiency of Decision Making Units (DMUs). One of the topics of this chapter is concerned with development of indicators to determine whether or not the specification of the input and output space is supported by data in the sense that the variation in data is sufficient for estimation of a frontier of the same dimension as the input output space. Insufficient variation in data implies that some inputs/outputs can be substituted along the efficient frontier but only in fixed proportions. Data thus locally support variation in a subspace of a lower dimension rather than in the input output space of full dimension. The proposed indicators are related to the existence of so-called Full Dimensional Efficient Facets (FDEFs). To characterize the facet structure of the CCR- or the BCC-estimators, (Charnes et al. Eur J Oper Res 2:429–444, 1978; Banker et al. Manage Sci 30(9):1078–1092, 1984) of the efficient frontier we derive a dual representation of the technologies. This dual representation is derived from polar cones. Relying on the characterization of efficient faces and facets in Steuer (Multiple criteria optimization. Theory, computation and application, 1986), we use the dual representation to define the FDEFs. We provide small examples where no FDEFs exist, both for the CCR- and the BCC estimator. Thrall (Ann Oper Res 66:109–138, 1996) introduces a distinction between interior and exterior facets. In this chapter we discuss the relationship between this classification of facets and the distinction in Olesen and Petersen (Manage Sci 42:205–219, 1996) between non-full dimensional and full dimensional efficient facets. Procedures for identification of all interior and exterior facets are discussed and a specific small example using *Qhull* to generate all facets is presented. In Appendix B we present the details of the input to and the output from *Qhull.* It is shown that the existence of well-defined marginal rates of substitution along the estimated strongly efficient frontier segments requires the existence of FDEFs. A test for the existence of FDEFs is developed, and a technology called EXFA that relies only on FDEFs and the extension of these facets is proposed, both in the context of the CCR-model and the BCC-model. This technology is related to the Cone-Ratio DEA. The EXFA technology is used to define the EXFA efficiency index providing a lower

O. B. Olesen (✉) · N. Chr. Petersen
Department of Business and Economics, The University of Southern Denmark, Odense, Denmark
e-mail: ole@sam.sdu.dk

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_6

bound on the efficiency rating of the DMU under evaluation. An upper bound on the efficiency rating is provided by a technology defined as the (non-convex) union of the input output sets generated from FDEFs only. Finally, we review recent uses of efficient faces and facets in the literature.

**Keywords** Efficiency measurement · Data envelopment analysis · Dual representation of technologies · Virtual multipliers · Model misspecification · Rates of substitutions · Frontier estimation · Convex analysis · Faces · Facets · Test for facets

## 6.1 Introduction

[1]Data Envelopment Analysis (DEA) models like the CCR-model (Charnes et al. 1978), (Charnes et al. 1979) or the BCC-model (Banker et al. 1984) are nonparametric and extremal methods for estimating production frontiers and evaluating the efficiency of Decision Making Units (DMUs). In this chapter we illustrate how to use DEA as a method for estimation of a strongly efficient frontier in line with (Charnes et al. 1985) and (Charnes et al. 1989). It has been suggested in the literature to use the optimal virtual multipliers to estimate local scale and substitution characteristics for the strongly efficient frontier ((Lewin and Morey 1981), (Banker et al. 1986), (Banker et al. 1988), (Charnes et al. 1990)). In this chapter we state a number of reservations on the usefulness of virtual multipliers provided by a CCR- or a BCC estimation. It is argued that a CCR- or a BCC estimation does not in general provide a strongly efficient frontier with well-defined rates of substitution and that the optimal virtual multipliers therefore should be interpreted with care, when DEA is used for estimation of efficiency scores and local substitution characteristics. This chapter provides alternative nonparametric estimations of efficiency scores which guarantee a set of well-defined local scale and substitution characteristics.

Efficiency evaluation and estimation of substitutional rates are important objectives in a large number of production studies. Data are often passively generated by an experimental design proposed by society and collected by agencies for administrative rather than for research purposes, the number of observations are sometimes limited and variables may not vary over a sufficiently wide range. As a result, the sample may not provide enough information for estimation of a frontier with well-defined rates of substitution, i.e. insufficient variation in data may imply estimation of a frontier actually located in subspaces of lower dimension than the specified input output space. Lack of variation in data leads to the collinearity problem in a parametric efficiency analysis as reflected by the existence of general interrelationships

[1] Parts of the results presented in this chapter are based upon (Olesen and Petersen 1996) with permission from the Institute for Operations Research and the Management Sciences, and (Olesen and Petersen 2003) with permission from Springer.

among the set of explanatory variables and hence instability of parameter estimates.[2] The potential presence of collinearity thus constitutes a limit on the level of disaggregation in a parametric estimation. More important, perhaps, a number of practical warning signals for detection of multicollinearity are available within the parametric approach, e.g. condition numbers, condition indices, and regression coefficient variance decomposition. DEA provides no practical warning signals in case of insufficient variation in data. The point to be made is that if the analysis is concerned with estimation of local scale and substitution characteristics, then the variation in data must be consistent with the a priori specification of the input-output space, i.e. the estimated frontier must provide piecewise constant rates of substitution in efficient production between all pairs of inputs and outputs. The dimensions of the efficient faces of the estimated strongly efficient frontier is an important determinant for data consistent specifications of the input-output space. The efficient faces are described by Koopmans (1957) in the following description of the linear activity model:

> Because of the finiteness of the technological basis the graph of the production function is put together from "flat" linear pieces known as facets.. . . Substitution of inputs and outputs within one facet therefore takes the form of simultaneous changes in the activity level of two or more processes, so coordinated as not to affect the net output of commodities not involved in the substitution. If the facet has the number n-1 of dimensions [3] required for it to make up a flat piece of the production function, constant rates of substitution (in efficient production) between all pairs of commodities (inputs, or outputs, or one of each) are uniquely determined within the facet.. .. However points on the "edges" of a facet, which are points common to more than one facet, possess infinitely many associated price systems, including those specific to the facets they help join. (Koopmans 1957, p. 94)

Consider the output isoquant in Fig. 6.1a with DMUs A, B and C producing three outputs using the same amount of only one input in a technology characterized by constant returns to scale (CRS). Table 6.1a summarizes the three input output vectors. The triangle ABC indicates the boundary of a "flat" segment denoted an efficient facet[4] of dimension two in the three dimensional output space. The vector normal to an efficient facet (scaled to unity length) can be given an interpretation

---

[2] In the case of parametric estimation, extreme collinearity is defined as the existence of an exact linear relation among the explanatory variables. However, the distinction between independent and explanatory variables is not straightforward in the context of DEA. Hence, we will refrain from giving any precise definition of extreme collinearity in relation to DEA; it suffices to note that extreme collinearity in relation to DEA includes the case, where there exists a vector a $\alpha \in \mathbb{R}^m, \alpha \neq 0$, such that $\alpha^T X = 0$, or $\beta \in \mathbb{R}^s, \beta \neq 0$, such that $\beta^T Y = 0$, where $X$ is the $m \times N$ matrix of $N$ input vectors and $Y$ is the $s \times N$ matrix of $N$ output vectors from $N$ DMUs.

[3] n corresponds to the number of inputs and outputs in a DEA analysis.

[4] In this paper we take the term *facet* of a convex polyhedral set $P \in \mathbb{R}^n$ to mean a maximal face of $P$ distinct from $P$ (maximal under inclusion), (see Klee 1953; Schrijver 1986). Furthermore, we follow this definition in our specification of the term *efficient facet* in the sense that an efficient facet is a maximal efficient face, i.e. an efficient face not included in any other efficient face. Using this definition of the term *efficient facet* we may encounter efficient facets of any dimension $d, 1 \leq d \leq n-1$; efficient facets of dimension $n-1$ is termed full dimensional efficient facets (FDEF). However, as pointed out by one of the referees of (Olesen and Petersen 1996), some authors define a facet of a convex polyhedral set in $\mathbb{R}^n$ as a face of dimension $n-1$; consequently, all facets are of dimension $n-1$ and a non full dimensional facet does not exist. Following that tradition a FDEF is simply an efficient facet and a non-FDEF is a maximal efficient face of dimension less than $n-1$.

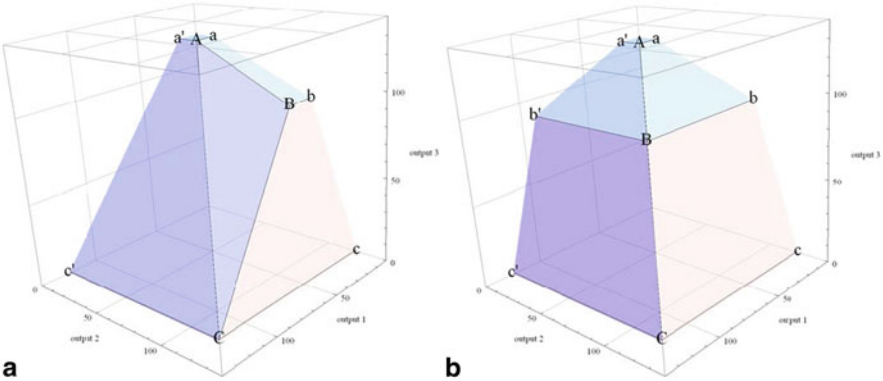



**Fig. 6.1** **a** The CCR frontier in output space ABC as an efficient facet of dimension two **b** The CCR frontier ABC in output space as the union of two efficient facets each of dimension one

**Table 6.1** Input output data for the output isoquants in Fig. 6.1a and b

| DMU: | A | B | C | DMU: | A | B | C |
|---|---|---|---|---|---|---|---|
| Input | 1 | 1 | 1 | Input | 1 | 1 | 1 |
| Output 1 | 15 | 20 | 120 | Output 1 | 15 | 90 | 120 |
| Output 2 | 15 | 90 | 120 | Output 2 | 15 | 90 | 120 |
| Output 3 | 120 | 90 | 1 | Output 3 | 120 | 90 | 1 |
| Table 6a: Input output data for Fig. 6.1a | | | | Table 6b: Input output data for Fig. 6.1b | | | |

in terms of a vector of relative prices. The scaled price vector associated with a point in the relative interior of an efficient facet is unique if the efficient facet is of full dimension. The components of a unique price vector associated with an efficient facet of full dimension define the marginal rates of substitution between inputs and/or outputs on the efficient facet, i.e. from one efficient point to another in the relative interior of the efficient facet. The choice between different modes of production thus opens the same alternatives as would trading at constant prices defined by the ratios in pairs between components of the price vector within the limit of any given efficient facet. Hence, we have constant rates of substitution between the three outputs along the relative interior of the efficient facet, i.e. the relative interior of the strongly efficient frontier. The specification of the input-output space is thus data consistent; the estimated strongly efficient frontier is of full dimension and a set of well-defined rates of substitution exists.

Points in the output possibility set outside ABC correspond either to inefficient or weakly efficient activities. Consider for instance the weakly efficient segment of the

frontier ACc′a′. This segment is also a facet of full dimension[5] and hence associated with a unique scaled normal vector with the component corresponding to output dimension 2 equal to zero. This vector should not be given an interpretation in terms of a price vector representing efficient rates of substitution because points in the relative interior of the facet are not (strongly) efficient. Analogous observations hold true for the remaining weakly efficient segments of the frontier.

DEA is based on a linear programming approach. The virtual multipliers provided by the CCR model thus correspond to optimal basic solutions. It is well known that an optimal solution for an extreme efficient[6] DMU in the input-output space is highly degenerate which implies the existence of alternative optimal solutions in the multiplier space. An optimal solution in the multiplier space with all virtual multipliers as (strict positive) basics defines a normal vector to the efficient facet ABC. Such solutions render DMUs A, B, and C efficient and provide estimates of substitutional rates along the strongly efficient frontier. The remaining alternative optimal basic solutions in the example in Fig. 6.1a correspond to some projections and provide estimates of substitutional rates in that projection and not along the efficient frontier.

A firm distinction between the alternative optimal bases is not required if the achievement of conservative efficiency measures is the main purpose of the analysis. But the solution corresponding to the efficient facet ABC is the only solution which provides estimates of the marginal rates of substitution along the efficient frontier. It is also the only solution defined by observed data solely; the estimated multipliers corresponding to the remaining set of alternative optimal solutions are easily seen to be affected by the assumption concerning disposability. The basic solution corresponding to the efficient facet should therefore be identified if the analysis is concerned with estimation of substitutional rates along the efficient frontier.

A data set may well yield a strongly efficient CCR frontier without well-defined rates of substitution[7]. Consider the situation in Fig. 6.1b again with DMUs A, B and

---

[5] Thrall (1996) distinguishes between interior and exterior facets. In this notation ABC is an interior facet and ACc'a' is an exterior facet. We will return to this classification of facets in Sect. 6.4.

[6] An efficient point $(x_o, y_o)$ in the production possibility set $T$ is *weakly efficient* if $(\theta x_o, y_o) \notin T, \forall \theta \in (0,1)$ and $(\{(x,y) : x \leq x_o, y \geq y_o\} \cap T) \setminus \{(x_o, y_o)\} \neq \varnothing$. $(x_o, y_o)$ is *strongly efficient* if $\{(x,y) : x \leq x_o, y \geq y_o\} \cap T = \{(x_o, y_o)\}$. Charnes et al. (1991, p. 205) further classify a strongly efficient DMU$j$ in relation to the CCR-model as being *extreme efficient* if the dimension of the cone of feasible multipliers

$$\mathcal{F}^{CCR}(\{j\}) \equiv \left\{(u,-v) \in \mathcal{P}_{CCR}, u^T Y_j - v^T X_j = 0\right\}$$

in (6.5) below is equal to $s + m$. Otherwise, the strongly efficient DMU is denoted non-extreme efficient.

[7] It is a peculiar phenomenon that any CCR frontier with at least two extreme efficient observations in a three dimensional input output space will have no non full dimensional efficient facets. To illustrate geometrically the concept of a non full dimensional efficient facet we need at least a four dimensional input output space. Hence, for a geometric illustration of this concept (Olesen and Petersen 1996) use an output isoquant in a three dimensional output space. Notice however, that it is very easy to illustrate a non full dimensional efficient facet in the BCC model in a three dimensional input output space, see below.

C producing three outputs using the same amount of one input. As summarized in Table 6.1b, the two line segments AB and BC define the efficient frontier. The marginal rates of substitution along the efficient frontier are consequently not well-defined as each segment possesses infinitely many associated price systems, including those specific to the (weakly efficient) facets they help join. The variation in data does not support rates of substitution between the three outputs as outputs 1 and 2 can only be substituted efficiently for output 3 in fixed proportions along the efficient frontier. Data thus locally support variation in a two dimensional output subspace rather than the full three dimensional space[8]. Each segment of the efficient frontier is in this sense subject to local collinearity.

A CCR efficiency evaluation on data as in Fig.6.1b is performed relative to a frontier technology without well-defined marginal rates of substitution along the efficient frontier; variations in inputs and outputs are thus (locally) constrained to occur in fixed proportions. The optimal bases in the multiplier space reflect substitutional possibilities along the weakly efficient segments of the estimated frontier, e.g. the "flat" segments aABb or a′ABb′. Hence, estimates are affected by the hypothesis concerning disposability and not determined by observed data solely. The estimated substitutional rates relate to either one of the two projections shown in the figure, i.e. substitution along ab, bc, a′b′, or b′c′. In effect, no optimal set of multipliers provided by the CCR model can be given a firm interpretation in terms of marginal rates of substitution.

More data is no help in generating well-defined rates of substitution if it is simply more of the same. Additional data from CCR extreme efficient observations generating efficient facets of a suitable dimension is needed. However, there is often no easy way to get better data. Data are generated by the functioning of the technology, and the non-existence of well-defined rates of substitutions may reflect the nature of the underlying production process. Well-defined rates of substitution thus constitute a limit on the level of disaggregation in a DEA estimation. The problem can be ignored if the achievement of conservative efficiency measures is the main purpose of the analysis, but not if the analysis is concerned with estimation of substitutional rates along the efficient frontier.

For the case of the BCC model introduced in Banker et al. (1984) let us consider the production possibility set in Fig. 6.2a with DMUs A, B and C producing one output using two inputs in a technology characterized by varying returns to scale. Table 6.2a summarizes the three input output vectors.

The triangle ABC indicates the boundary of a "flat" segment of an efficient facet of dimension two in this three dimensional input-output space. Notice, that with varying returns to scale we observe an efficient facet of dimension two being spanned by (at least) three extreme efficient DMUs (the origin does not necessarily belong to the facet). This is in contrast to the situation with CRS, where a facet of dimension two is spanned by typically only two observations and the origin. The facet ABC determines a unique normal vector and *a unique intercept term.* The vector normal to an efficient

[8] The problem is thus of the same nature as collinearity in the parametric approach.

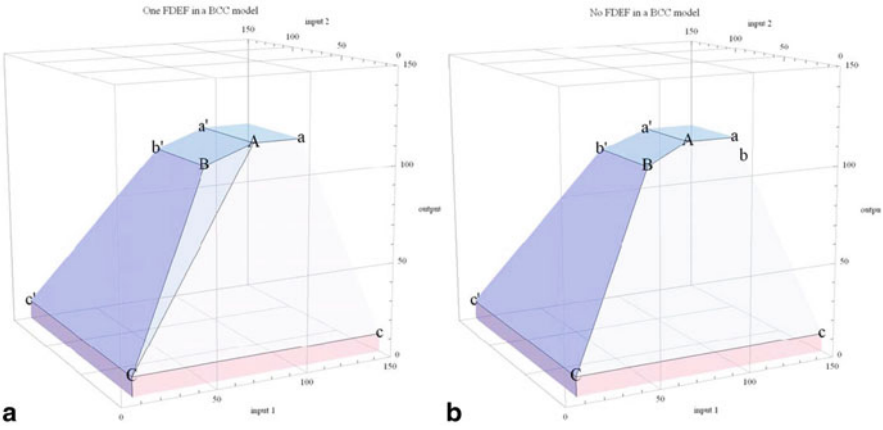



**Fig. 6.2** **a** The frontier ABC as an efficient facet of dimension two. **b** The frontier ABC as the union of two efficient facets each of dimension one

**Table 6.2** Input output data for Fig. 6.2a and 6.2b

| DMU: | A | B | C |
|---|---|---|---|
| Input 1 | 120 | 90 | 10 |
| Input 2 | 90 | 90 | 10 |
| Output | 100 | 90 | 10 |

Table 6.2a: Input output data for Fig. 6.2a

| DMU: | A | B | C |
|---|---|---|---|
| Input 1 | 120 | 90 | 10 |
| Input 2 | 100 | 90 | 10 |
| Output | 100 | 90 | 10 |

Table 6.2b: Input output data for Fig. 6.2b

facet (scaled to unity length) can be given the same interpretation as described in the CRS case. The scaled price vector associated with a point in the relative interior of an efficient facet is unique if the efficient facet is of full dimension. The unique intercept term can be interpreted as a measure related to the local scale elasticity characteristic, see (Banker et al. 1984).

As in the CRS case, a data set may well yield a strongly efficient BCC frontier without well-defined rates of substitution. Consider the situation in Fig. 6.2b again with DMUs A, B and C producing one output using the two inputs. As summarized in Table 6.2b, the two line segments AB and BC define the efficient frontier. The marginal rates of substitution along the efficient frontier are consequently not well-defined as each segment possesses infinitely many associated price systems, including those specific to the (weakly efficient) facets they help join.

The chapter is organized as follows. In Sect. 6.2 we are concerned with an analysis of the facet structure of a CCR- estimator of the production possibility set (PPS). We focus on the non-existence of full dimensional efficient facets (FDEFs) as a theoretical bound for the level of disaggregation in the CCR-model. We develop a dual characterization of the empirical production possibility set using the theory of polar cones. This dual description allows us to trace the impact of restrictions on feasible multipliers on the estimated PPS. The results from Sect. 6.2 is generalized

to the BCC-model in Sect. 6.3. In Sect. 6.4 we discuss the relationship between the notion of FDEFs and the notion of interior and exterior facet proposed by Thrall (1996). A presentation of procedures for an identification of all interior and exterior facets and hence all FDEFs using specialized algorithms is at focus in Sect. 6.5. A detailed "user guide" for one particular software *Qhull* is included, and detailed input to and output from *Qhull* is discussed in Appendix B. In Sect. 6.6 we derive two alternative technologies for the CCR and the BCC technology. In the first technology we only allow multipliers in the cone spanned by normal vectors from FDEFs. In the second technology we create a possibly non-convex set as the union of the input output sets generated from FDEFs only. Mixed Integer Linear Programs (MILPs) for test of the existence of FDEFs are developed. Procedures for estimation of efficiency indices relative to these two technologies are presented and it is argued that these indices provide a lower and an upper bound on the technical efficiency index (with the CCR or the BCC index in between). The procedures involve estimation of an extended facet efficiency index along with a restricted CCR-index or BCC-index based upon the concept of a frontier technology spanned by only FDEFs. Section 6.7 finally reviews some of the recent uses of efficient faces and facets in the literature.

## 6.2 Primal and Dual Description of the Production Possibility Set $T^{CCR}$

Following the approach outlined in Olesen and Petersen (1996) we will derive a polar or dual description of a piecewise linear enveloped production possibility set given as the conical hull (the convex hull is derived in the next section) of the observed input output combinations set added to $\mathbb{R}^s_- \times \mathbb{R}^m_+$, where $s$ ($m$) is the number of outputs (inputs). However, contrary to Olesen and Petersen (1996) we will not focus entirely on the envelopment by Full Dimensional Efficient Facets (FDEFs) only, but include both *interior facets* and *exterior facets* (see Sect. 6.4 for a formal definition of interior and exterior facets). The discussion in this section refers to a maintained hypothesis of Constant Returns to Scale (CRS) used in the CCR model; the BCC model with Varying Returns to Scale (VRS) is developed in the next section.

Let $\mathbb{E}$ be an index set for the CCR strongly efficient[9] DMUs in a sample of observations where the $j$'th DMU produces an $s$-dimensional output vector $Y_j$ while consuming an $m$-dimensional input vector $X_j$, i.e. $Y \equiv [Y_1, \ldots, Y_n]$, $X \equiv [X_1, \ldots, X_n]$. We assume that $Y_j > 0, X_j > 0, \forall\, j$. Let $T^{CCR}$ be the estimated polyhedral empirical production possibility set in the CCR model:

$$T^{CCR} \equiv \left\{ (y,x) \in \mathbb{R}^{s+m}_+ \mid \sum_{j\in\mathbb{E}} \lambda_j Y_j \geq y, \sum_{j\in\mathbb{E}} \lambda_j X_j \leq x, \lambda_j \geq 0, j \in \mathbb{E} \right\} \tag{6.1}$$

[9] See note 6.

Let $\mathcal{P}_{CCR}$ represent the polyhedral cone of virtual output input multipliers $(u, -v)$ determined from the halfspace constraints in a CCR-model in multiplier form and from $(u, v) \neq (0, 0)$:

$$\mathcal{P}_{CCR} \equiv \left\{(u, -v) \in \mathbb{R}^s_+ \times \mathbb{R}^m_- \,|\, u^T Y_j - v^T X_j \leq 0, j \in \mathbb{E}, (u, v) \neq (0, 0)\right\} \quad (6.2)$$

The representation of the non-parametric production possibility set $T^{CCR}$ in terms of $\mathcal{P}_{CCR}$ follows from Theorem 1:

**Theorem 1** *Let the polyhedral cone of feasible virtual multipliers $\mathcal{P}_{CCR}$ be given by (6.2). The corresponding production possibility set T in (6.1) is the intersection of the polar cone $\mathcal{P}^\circ_{CCR}$ and the non-negative orthant $\mathbb{R}^{s+m}_+$:*

$$T^{CCR} = \mathcal{P}^\circ_{CCR} \cap \mathbb{R}^{s+m}_+ = \left\{(y, x) \in \mathbb{R}^{s+m}_+ \,|\, \forall\, (u, -v) \in \mathcal{P}_{CCR} : u^T y - v^T x \leq 0\right\} \quad (6.3)$$

*Proof* (The following proof is based on Rockefellar (1970), Sec. 14) Consider the production possibility set

$$T^{CCR} = \left\{(y, x) \in \mathbb{R}^{s+m}_+ \,|\, \sum_{j\in\mathbb{E}} \lambda_j Y_j \geq y, \sum_{j\in\mathbb{E}} \lambda_j X_j \leq x, \lambda_j \geq 0, j \in \mathbb{E}\right\}$$

$T^{CCR} = K \cap \mathbb{R}^{s+m}_+$, where $K$ is the convex cone:

$$K \equiv conv\left(cone\left(\left(Y_j, X_j\right),\ j \in \mathbb{E}\right)\right) + \mathbb{R}^s_- \times \mathbb{R}^m_+$$

where $conv\{\bullet\}$ is the convex hull operator[10] and $cone\{z_i, i \in I\} \equiv \{z \,|\, z = \lambda_i z_i, \lambda_i \geq 0, i \in I\}$. $K$ is generated by the following set of $N + s + m$ vectors:

$$\left\{\left(Y_j, X_j\right),\ j \in \mathbb{E}^{BCC}, -e_k, k = 1, \ldots, s, e_{s+i}, i = 1, \ldots, m\right\}$$

where $e_l$ is the $l$'th unit vector and[11] $N = |\mathbb{E}|$. The polar cone $K^\circ$ is

$$K^\circ = \left\{(u, -v) \in \mathbb{R}^s_+ \times \mathbb{R}^m_- \,|\, u^T Y_j - v^T X_j \leq 0, j = 1, \ldots, N\right\}$$

Hence $K^\circ = \mathcal{P}_{CCR} \cup \{0\}$ and since $K$ is a closed cone, $K = K^{\circ\circ} = \mathcal{P}^\circ_{CCR}$

□

The definition of the efficient frontier *Eff* $T^{CCR}$ is

$$Eff T^{CCR} \equiv \left\{(y, x) \in T^{CCR} \,|\, \exists\ (u, -v) \in \mathcal{P}_{CCR}, u \geq \epsilon e, v \geq \epsilon e : u^T y - v^T x = 0\right\} \quad (6.4)$$

---

[10] $conv\left(\{z_1, \ldots, z_n\}\right) \equiv \left\{z \,|\, z = \sum_{j=1}^n \lambda_j z_j, \sum_{j=1}^n \lambda_j = 1, \lambda_j \in [0, 1], \forall j\right\}$.

[11] $|\mathbb{E}|$ denotes the number of elements in the index set $\mathbb{E}$.

where $e^T$ is the vector $(1, \ldots 1)$ of an appropriate dimension, and $\epsilon$ is a non-Archimedian. Let $\mathcal{F}^{CCR}(\{j\})$ be the set of virtual multipliers in $\mathcal{P}_{CCR}$ which render $\mathrm{DMU}_j$, $j \in \mathbb{E}$, efficient:

$$\mathcal{F}^{CCR}(\{j\}) \equiv \left\{(u, -v) \in \mathcal{P}_{CCR}, u^T Y_j - v^T X_j = 0\right\} \tag{6.5}$$

and let for some $J \subseteq \mathbb{E}$, $\mathcal{F}^{CCR}(J) = \bigcap_{j \in J} \mathcal{F}^{CCR}\{j\}$. The elements in $\mathcal{F}^{CCR}\{J\}$ correspond to the set of virtual multipliers in $\mathcal{P}_{CCR}$ which render all DMUs in the index set $J$ efficient (weakly or strongly efficient).

$\Phi \subseteq T^{CCR}$ is a face of $T^{CCR}$ if $\Phi = T^{CCR}$ or if $\Phi$ is the intersection of $T^{CCR}$ with a supporting hyperplane of $T^{CCR}$. Let us consider the following face $F(J)$:

$$\begin{aligned} F(J) = &\left\{(\widehat{y}, \widehat{x}) \in T^{CCR} \mid \forall (\widehat{u}, -\widehat{v}) \right. \\ &\left. \in \mathcal{F}^{CCR}(J), \forall (y, x) \in T^{CCR}, \widehat{u}^T \widehat{y} - \widehat{v}^T \widehat{x} \geq \widehat{u}^T y - \widehat{v}^T x\right\} \\ = &\left\{(\widehat{y}, \widehat{x}) \in T^{CCR} \mid \forall (\widehat{u}, -\widehat{v}) \in \mathcal{F}^{CCR}(J), \widehat{u}^T \widehat{y} - \widehat{v}^T \widehat{x} = 0\right\} \end{aligned} \tag{6.6}$$

and define $EF(J) = F(J) \cap \mathit{Eff}T^{CCR}$.

$\mathcal{F}^{CCR}(J) = \varnothing$ if the set of the DMUs in the index set $J$ is located at different supporting hyperplanes of $\mathit{Eff}T^{CCR}$. $\mathcal{F}^{CCR}(J) \neq \varnothing$ implies the existence of at least one $(\widehat{u}, \widehat{v}) \neq 0$ which renders all DMUs in the index set $J$ efficient and hence is a generating normal vector for a supporting hyperplane of $\mathit{Eff}T^{CCR}$ with each and every member of $J$ located on it, see (Banker et al. 1984). $EF(J)$ is hence an efficient face of $T^{CCR}$ by construction.

**Definition 1** An efficient face $EF(J)$ with a normal vector $(\widehat{u}, -\widehat{v}) \neq 0$ is denoted an efficient facet if the dimension $d$ of $EF(J)$ is maximal, i.e. no efficient face $EF(J')$ exists where $\mathcal{F}^{CCR}\left(J'\right) \neq \varnothing$, with $dim(F(J')) = d'$, $EF(J) \subset EF(J')$ with $d' > d$ (See e.g. (Steuer 1986, p. 182)).

**Definition 2** An efficient facet $F(J)$ is denoted a Full Dimensional Efficient Facet (FDEF) if the dimension is equal to $s + m - 1$. An efficient facet $F(J)$ is denoted a Nonfull Dimensional Efficient Facet (NFDEF) if the dimension is less than $s+m-1$.

From Definitions 1 and 2 follows that DMU $j_o$ contributes to the spanning of an FDEF iff:

$$\exists (\widehat{u}, -\widehat{v}) \in \mathcal{F}^{CCR}(\{j_o\}) : (\widehat{u}, -\widehat{v})^T \begin{bmatrix} Y_j \\ X_j \end{bmatrix} = 0, \text{ for } j \in \widehat{J} \equiv \left\{j \in \mathbb{E} : \widehat{u}^T Y_j - \widehat{v}^T X_j = 0\right\} \tag{6.7}$$

and $rank(D) = s + m - 1$, where $D$ is the $(s+m) \times \left|\widehat{J}\right|$ matrix[12] $\left[\begin{bmatrix} Y_j \\ X_j \end{bmatrix}, j \in \widehat{J}\right]$.

[12] $|J|$ denotes the number of elements in the index set $J$.

Clearly,

$$EF\left(\widehat{J}\right) = \left\{(y,x) : (y,x) = \sum_{j\in\widehat{J}} \lambda_j \left(Y_j, X_j\right), \lambda_j \geq 0, j \in \widehat{J}\right\} \tag{6.8}$$

and $dim(EF(\hat{J})) = rank(D)$.

Non-existence of so-called FDEFs indicates a misspecification of the input-output space, see (Olesen and Petersen 1996), in the sense that the estimated rates of substitution by construction must refer to a projection and not necessarily the same one, i.e. substitutions take place in various subspaces of the input-output space; the data set is ill-conditioned in the sense that data do not support the specification of the input-output space, see (Olesen and Petersen 1996) for details.

## 6.3 Primal and Dual Description of the Production Possibility Set $T^{BCC}$

Following the approach in the previous section we now extent the polar or dual description to a piecewise linear enveloped production possibility set given as the convex hull of the observed input output combinations set added to $\mathbb{R}^s_- \times \mathbb{R}^m_+$. Let $\mathbb{E}^{BCC} = \{1,\ldots,n\}$ be an index set for the BCC strongly efficient DMUs in the sample. Let $T^{BCC}$ be the estimated polyhedral empirical production possibility set in the BCC model. To allow for an implicit cone representation of $T^{BCC}$ we extend the input output vector from each DMU in $\mathbb{E}^{BCC}$ with an additional element equal to one for all DMUs, $Z_j = 1, \forall j$. We can now express $T^{BCC}$ as the intersection of the cone

$$CT^{BCC} \equiv \left\{(y,x,z) \in \mathbb{R}^{s+m}_+ \times \mathbb{R}_+ \mid \sum_{j=1}^{n} \lambda_j Y_j \geq y, \sum_{j=1}^{n} \lambda_j X_j \leq x, \sum_{j=1}^{n} \lambda_j Z_j = z, \lambda_j \geq 0, j \in \mathbb{E}^{BCC}\right\} \tag{6.9}$$

with the hyperplane $\left\{(y,x) \mid (y,x,z) \in CT^{BCC}, z = 1\right\}$, i.e.

$$T^{BCC} = \left\{(y,x) \in \mathbb{R}^{s+m}_+ \mid (y,x,1) \in CT^{BCC}\right\} \tag{6.10}$$

Let $\mathcal{P}_{BCC}$ represent the polyhedral cone of virtual output input multipliers and intercept multiplier $(u,-v,v_o)$ determined from the halfspace constraints in a BCC-model in multiplier form and from $(u,v) \neq (0,0)$:

$$\mathcal{P}_{BCC} \equiv \left\{(u,-v,v_o) \in \mathbb{R}^s_+ \times \mathbb{R}^m_- \times \mathbb{R} \mid u^T Y_j - v^T X_j + v_o \leq 0, j \in \mathbb{E}^{BCC}, (u,v) \neq (0,0)\right\} \tag{6.11}$$

Let

$$\widetilde{\mathcal{P}}_{BCC} \equiv \left\{(u,-v,v_o) \in \mathbb{R}^s_+ \times \mathbb{R}^m_- \times \mathbb{R} \mid u^T Y_j - v^T X_j + v_o Z_j \leq 0, j \in \mathbb{E}^{BCC}, (u,v) \neq (0,0)\right\} \tag{6.12}$$

Clearly, $\mathcal{P}_{BCC} = \widetilde{\mathcal{P}}_{BCC}$. The representation of the non-parametric production possibility set $T^{BCC}$ in terms of $\mathcal{P}_{BCC}$ follows from Theorem 2:

**Theorem 2** *Let the polyhedral cone of feasible virtual multipliers $\mathcal{P}_{BCC}$ be given by (6.11). The corresponding production possibility set $T^{BCC}$ in (6.10) is a section of the intersection of the polar cone $\mathcal{P}^{\circ}_{BCC}$ and the non-negative orthant $\mathbb{R}^{s+m+1}_{+}$:*

$$\begin{aligned} T^{BCC} &= \left\{(y,x) \in \mathbb{R}^{s+m}_{+} \mid (y,x,1) \in \mathcal{P}^{\circ}_{BCC} \cap \mathbb{R}^{s+m+1}_{+}\right\} \\ &= \left\{(y,x) \in \mathbb{R}^{s+m}_{+} | \forall\, (u,-v,v_o) \in \mathcal{P}_{BCC} : u^T y - v^T x + v_o \le 0\right\} \end{aligned} \tag{6.13}$$

*Proof* Consider the production possibility set

$$CT^{BCC} = \left\{ (y,x,z) \in \mathbb{R}^{s+m}_{+} \times \mathbb{R}_{+} \mid \sum_{j=1}^{n} \lambda_j Y_j \ge y, \sum_{j=1}^{n} \lambda_j X_j \le x, \right.$$
$$\left. \sum_{j=1}^{n} \lambda_j Z_j = z, \lambda_j \ge 0, j \in \mathbb{E}^{BCC} \right\}$$

$CT^{BCC} = K \cap \left(\mathbb{R}^{s+m}_{+} \times \mathbb{R}_{+}\right)$, where $K$ is the convex cone:

$$K \equiv conv\left(cone\left(\left(Y_j, X_j, Z_j\right), j \in \mathbb{E}^{BCC}\right)\right) + \mathbb{R}^{s}_{-} \times \mathbb{R}^{m}_{+} \times \{0\}$$

$K$ is generated by the following set of $N + s + m$ vectors:

$$\left\{\left(Y_j, X_j, Z_j\right), j \in \mathbb{E}^{BCC}, -e_k, k = 1, \ldots, s, e_{s+i}, i = 1, \ldots, m\right\}$$

where $e_l$ is the $l$'th unit vector and $N = \left|\mathbb{E}^{BCC}\right|$. The polar cone $K^{\circ}$ is

$$K^{\circ} = \left\{(u,-v,v_o) \in \mathbb{R}^{s}_{+} \times \mathbb{R}^{m}_{-} \times \mathbb{R} \mid u^T Y_j - v^T X_j + v_o Z_j \le 0, j \in \mathbb{E}^{BCC}\right\}$$

$K^{\circ} = \widetilde{\mathcal{P}}_{BCC} \cup \{0\} = \mathcal{P}_{BCC} \cup \{0\}$ and since $K$ is a closed cone, $K = K^{\circ\circ} = \mathcal{P}^{\circ}_{BCC}$. Hence, to summarize we know that

$$CT^{BCC} = K \cap \mathbb{R}^{s+m+1}_{+} \tag{*}$$

$$K = \mathcal{P}^{\circ}_{BCC} \tag{**}$$

and from (*) and (**) follows that

$$CT^{BCC} = \mathcal{P}^{\circ}_{BCC} \cap \mathbb{R}^{s+m+1}_{+}$$

and

$$\begin{aligned} T^{BCC} &= \left\{(y,x) \in \mathbb{R}_+^{s+m} \mid (y,x,1) \in CT^{BCC}\right\} \\ &= \left\{(y,x) \in \mathbb{R}_+^{s+m} \mid (y,x,1) \in \mathcal{P}_{BCC}^{\circ} \cap \mathbb{R}_+^{s+m+1}\right\} \\ &= \left\{(y,x) \in \mathbb{R}_+^{s+m} \mid \forall\, (u,-v,v_o) \in \mathcal{P}_{BCC} : u^T y - v^T x + v_o \le 0\right\} \end{aligned}$$

□

The definition of the efficient frontier $Eff\ T^{BCC}$ is

$$Eff T^{BCC} \equiv \left\{(y,x) \in T^{BCC} \mid \exists\, (u,-v,v_o) \in \mathcal{P}_{BCC}, u \ge \varepsilon e, v \ge \varepsilon e : u^T y - v^T x + v_o = 0\right\} \tag{6.14}$$

where $e$ is the vector $(1,\ldots 1)$ of an appropriate dimension, and $\epsilon$ is a non-Archimedian. Let $\mathcal{F}^{BCC}(\{j\})$ be the set of virtual multipliers in $\mathcal{P}_{BCC}$ which render $\mathrm{DMU}_j$, $j \in \mathbb{E}$, efficient:

$$\mathcal{F}^{BCC}(\{j\}) \equiv \left\{(u,-v,v_o) \in \mathcal{P}_{BCC} \mid u^T Y_j - v^T X_j + v_o = 0\right\} \tag{6.15}$$

and let for some $J \subseteq \mathbb{E}^{BCC}$, $\mathcal{F}^{BCC}(J) = \bigcap_{j \in J} \mathcal{F}^{BCC}\{j\}$. The elements in $\mathcal{F}^{BCC}(J)$ correspond to the set of virtual multipliers in $\mathcal{P}_{BCC}$ which render all the DMUs in the index set $J$ efficient. Let us consider the following face $F^{BCC}(J)$:

$$\begin{aligned} F^{BCC}(J) = \Big\{(\widehat{y},\widehat{x}) \in T^{BCC} \mid \forall\, (\widehat{u},-\widehat{v},v_o) \in \mathcal{F}^{BCC}(J), \forall\, (y,x) \in T^{BCC}, \\ \widehat{u}^T\widehat{y} - \widehat{v}^T\widehat{x} + \widehat{v}_o \ge \widehat{u}^T y - \widehat{v}^T x + \widehat{v}_o\Big\} \\ = \left\{(\widehat{y},\widehat{x}) \in T^{BCC} \mid \forall\, (\widehat{u},-\widehat{v},v_o) \in \mathcal{F}^{BCC}(J), \widehat{u}^T\widehat{y} - \widehat{v}^T\widehat{x} + \widehat{v}_o = 0\right\} \end{aligned} \tag{6.16}$$

and define $EF^{BCC}(J) = F^{BCC}(J) \cap Eff T^{BCC}$.

$\mathcal{F}^{BCC}(J) = \varnothing$ if the set of DMUs in the index set $J$ is located at different supporting hyperplanes of $Eff T^{BCC}$. $\mathcal{F}^{BCC}(J) \neq \varnothing$ implies the existence of at least one $(\widehat{u},-\widehat{v},\widehat{v}_o) \neq 0$ which renders all DMUs in the index set $J$ efficient and hence is a generating normal vector for a supporting hyperplane of $Eff T^{BCC}$ with each and every member of $J$ located on it, see (Banker et al. 1984). $EF^{BCC}(J)$ is hence an efficient face of $T^{CCR}$ by construction.

**Definition 3** An efficient face $EF^{BCC}(J)$ with a normal vector $(\widehat{u},-\widehat{v}) \neq 0$ and an intercept term $v_o$ is denoted an efficient facet if the dimension $d$ of $EF^{BCC}(J)$ is maximal, i.e. no efficient face $EF^{BCC}(J')$ exists where $\mathcal{F}^{BCC}(J') \neq \varnothing$, with $dim(F(J')) = d'$, $EF^{BCC}(J) \subset EF^{BCC}(J')$ with $d' > d$ (See e.g. Steuer 1986, p. 182).

**Definition 4** An efficient facet $F^{BCC}(J)$ is denoted a Full Dimensional Efficient Facet (FDEF) if the dimension is equal to $s + m - 1$. An efficient facet $F^{BCC}(J)$ is denoted a Nonfull Dimensional Efficient Facet (NFDEF) if the dimension is less than $s + m - 1$.

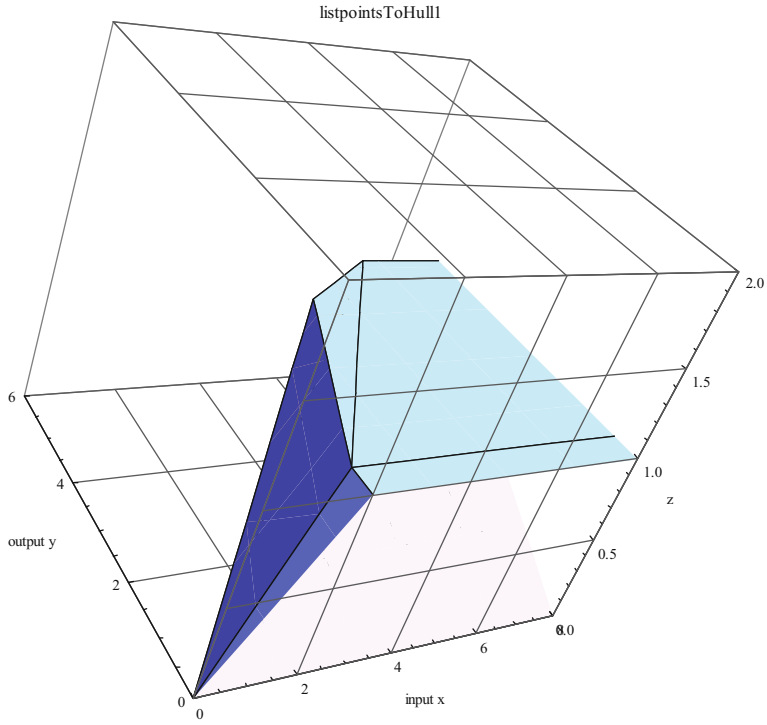



**Fig. 6.3** The cone $K$ generated from observations $\begin{bmatrix} 2 \\ 0.5 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \\ 1 \end{bmatrix}, \begin{bmatrix} 6 \\ 5 \\ 1 \end{bmatrix}$

From Definitions 3 and 4 follows that DMU $j_o$ contributes to the spanning of an FDEF iff:

$$\exists\,(\widehat{u}, -\widehat{v}, \widehat{v}_o) \in \mathcal{F}^{BCC}\left(\{j_o\}\right) \mid (\widehat{u}, -\widehat{v})^T \begin{bmatrix} Y_j \\ X_j \end{bmatrix} + \widehat{v}_o = 0, \text{ for } j \in \widehat{J}$$

$$\equiv \left\{ j \in \mathbb{E}^{BCC} \mid \widehat{u}^T Y_j - \widehat{v}^T X_j + \widehat{v}_o = 0 \right\} \tag{6.17}$$

and $rank(D) = s + m$, where $D$ is the $(s+m) \times \left|\widehat{J}\right|$ matrix $\left[ \begin{bmatrix} Y_j \\ X_j \end{bmatrix}, j \in \widehat{J} \right]$.

Clearly,

$$EF\left(\widehat{J}\right) = \left\{ (y,x) \mid (y,x,1) \in \left\{ (y,x,z) = \sum_{j \in \widehat{J}} \lambda_j \left(Y_j, X_j, 1\right), \lambda_j \geq 0, j \in \widehat{J} \right\} \right\} \tag{6.18}$$

and $dim(EF(\hat{J})) = rank(D) - 1$.

## 6.4 Interior and Exterior Facets

A distinction between *interior facets* and *exterior facets* is suggested in Thrall (1996). Focussing on the CRS CCR-model, this distinction is related to the following three sets, the production possibility set $T^{CCR}$, the efficient frontier $FR^{CCR}$, and the extended efficient frontier $EFR^{CCR}$:

$$T^{CCR} \equiv \left\{ (y,x) \in \mathbb{R}_+^{s+m} \mid \sum_{j=1}^{n} \lambda_j Y_j \geq y, \sum_{j=1}^{n} \lambda_j X_j \leq x, \lambda_j \geq 0, j \in \mathbb{K}, \lambda_k = 0, k \notin \mathbb{K} \right\}$$

$$FR^{CCR} = \left\{ (y,x) \in T^{CCR} \mid \left(y', -x'\right) \geq (y,-x), \left(y', x'\right) \in T^{CCR} \text{ requires } \left(y', -x'\right) = (y,-x) \right\}$$

$$EFR^{CCR} = \left\{ (y,x) \in T^{CCR} \mid \left(y', -x'\right) > (y,-x) \text{ holds for no } \left(y', x'\right) \in T^{CCR} \right\}$$

where $\mathbb{K}$ is the index set corresponding to weakly and strongly CCR-efficient DMUs. Let us define the following cone $C = \mathbb{R}_+^s \times \mathbb{R}_-^m$. $(y,x) \in T^{CCR}$ belongs to the efficient frontier $FR^{CCR}$ if no other point $\left(y', x'\right) \in T^{CCR}$ dominates $(y,x)$ on a Pareto criterion, i.e. $((y,x) + C) \cap T^{CCR} = (y,x)$. $FR^{CCR}$ is identical to $EffT^{CCR}$ in (6.4). The extended efficient frontier $EFR^{CCR}$ extends the efficient frontier $FR^{CCR}$ with the weakly efficient input output combinations generated from input and output disposability. $(y,x) \in T^{CCR}$ belongs to the extended efficient frontier $EFR^{CCR}$ if $((y,x) + int\,(C)) \cap T^{CCR} = (y,x)$. All three sets are cones in $\mathbb{R}_+^{s+m}$. $T^{CCR}$ is a convex cone but neither $FR$ or $EFR$ are in general convex sets. Thrall defines this distinction as follows: "A facet is called *interior* if it is a subset of $FR$, otherwise it is called *exterior*. It is quite possible that a domain $D$ may have no interior facets", (Thrall 1996, p. 133).

Any FDEF is an interior facet. An NFDEF is an efficient face that is i) contained in an exterior facet, and ii) never part of an interior facet. Thrall (1996) illustrates using a numerical example a situation where a data set may generate a $T^{CCR}$ without any interior facets. In this example no FDEF exists. All efficient facets are NFDEF.

These two different approaches to classification of facets reflect two different definitions of an efficient facet. Thrall (1996) takes the term *facet* of a convex polyhedral set $P \subset \mathbb{R}^n$ to mean a maximal face of $P$ distinct from $P$ (maximal under inclusion), (see Klee 1953; Schrijver 1986). Hence, with $T^{CCR} \subset \mathbb{R}_+^{s+m}$ and assuming that all observed input output vectors are strictly positive we have that any facet, interior or exterior is of a dimension $s + m - 1$. Following this tradition an FDEF is simply an efficient facet and a non-FDEF is a maximal efficient face of dimension less than $s + m - 1$.

The approach suggested in (Olesen and Petersen 1996) follows this definition in the specification of the term *efficient facet* in the sense that an efficient facet is a

maximal efficient face, i.e. is an efficient face not included in any other efficient face. Using this definition of the term efficient facet we may encounter efficient facets of any dimension $d, 1 \leq d \leq n-1$. Efficient facets of dimension $s+m-1$ are termed full dimensional efficient facets.

## 6.5 Procedures for Identification of the Total Set of FDEFs

This section is concerned with a presentation of a procedure for an identification of all interior and exterior facets and hence all FDEFs in real life data sets. The procedure identifies all facets of a polyhedron with either known extreme points or a known collection of halfspaces [13] by specialized convex hull algorithms.

Generating all facets from a piecewise linear envelopment of a number of data points is a time consuming task, especially in higher dimensions, i.e. with many inputs and/or outputs, and with many extreme efficient DMUs. However, as more efficient convex hull algorithms are designed and better implementation taking advantage of fast memory and fast processors/parallel processors is provided we will see an increase in the size of data sets that can be processed with reasonable running times. The number of facets in a polyhedral set in an input output space as defined by the application at hand may be large, since each extreme efficient DMU defines an extreme ray in the cone of feasible input output combinations under conditions of CRS, and since the production possibility set is obtained by an expansion of this cone due to the assumption of strong disposability, which in turn implies an introduction of additional extreme rays.

The identification of all FDEFs/interior facets and all exterior facets is for this reason highly facilitated when carried out in a sequence of local segments of the polyhedral possibility set so that the identification of FDEFs in one segment is independent compared to other segments. The decomposition of the search for interior and exterior facets into a number of local and mutually independent segments is an important device.

The underlying idea for the decomposition can be described as follows. Bear in mind that $\mathbb{E}$ is an index set for the set of extreme efficient DMUs, and let $\mathbb{N}$ denote an index set for all DMUs. It is obvious that only extreme efficient DMUs may contribute to the spanning of interior and exterior facets, which implies that $\text{DMU}_j$, $j \in \mathbb{N} \setminus \mathbb{E}$, can be eliminated from the sample in an identification of all FDEFs. Moreover, the only candidates for spanning an interior or exterior facet including any given $\text{DMU}_{j_0}$, $j_0 \in \mathbb{E}$, are those that can be termed efficient along with $\text{DMU}_{j_0}$ itself. Let $\mathbb{E}_{j_0}$ denote an index set for this set of extreme efficient DMUs. It is now possible to identify all interior and exterior facets in the local frontier segment affected by $\text{DMU}_{j_0}$ by an identification of all facets spanned by DMUs in the candidate set $\mathbb{E}_{j_0}$ only and with $\text{DMU}_{j_0}$ being one of the spanning units.

[13] In other words we know the convex set in either sum form or in intersection form.

### 6.5.1 Convex Hull Generation

Several codes for convex hulls generation/vertex enumeration are available. The following selection of implementations are some of the codes mentioned in directory of Computational Geometry Software on the website for The Geometry Center, University of Minnesota:[14]

1. *Qhull*, by Brad Barber, David Dobkin and Hannu Huhdanpaa, The Geometry Center.
2. *chD*, by Ioannis Emiris, U.C. Berkeley.
3. *Hull*, by Ken Clarkson, Bell Labs.
4. *Porta*, by Thomas Christof, Heidelburg University and Andreas Loebel, Konrad-Zuse-Zentrum fur Informatik (ZIB).
5. *cdd*, by Komei Fukuda, ETH Zurich, Switzerland and University of Tsukuba, Japan.
6. *lrs*, by David Avis, McGill University

We have experimented with the use of *Qhull*, *cdd* and *lrs* both for the generation of all interior facets (FDEFs) as well as all interior and exterior facets (FDEFs and NFDEFs). *Qhull* is very fast on small and moderate sized problems compared to *cdd* and *lrs.* Both *cdd* and *lrs* are developed to handle a degenerated hull generation and both use rational (exact) arithmetic. However, *cdd* also exists in a floating point version. *lrs* uses very little memory compared to *Qhull* and apparently it can solve very large convex hull problems. However, it is a slow code compared to *Qhull,*and *Qhull* can solve large problems, if sufficient fast memory is available (see below).

In this section we will concentrate on how to use *Qhull* for the generation of a convex hull. *Qhull* picks in a first phase a subset of $s+m+1$ data points. These points are chosen such that the convex hull equals a simplex in $R^{s+m}$. Using a simplified beneath-beyond algorithm (Grünbaum 1961) each of the remaining points is added one by one and the convex hull is extended successively. Output from *Qhull* is the collection of facets, the extreme points for each facet and a neighborhood relationship between the facets. The three programs have different input format but each one can to some extent use the format of the others. An input file for *Qhull* consists of two integers specifying the dimension $(s+m)$ of each vector and the number of vectors $n$ to be enveloped. Then follows the coordinates of these $n$ vectors, i.e. $n \times (s+m)$ real numbers. An $(s+m)$ -dimensional zero-vector must be included if zero is regarded a feasible input output vector as e.g. in the CCR-model. *Qhull* has a number of options but the following basic call to *Qhull* produces a listing of all facets enveloping the convex hull of the points from data in an input file $ifile$ listed to an output file *ofile*

$$Qhull.\text{exe s FF} < ifile > ofile$$

The convex hull generated has of course a number of facets which are not part of the strongly efficient frontier (in fact, if no FDEFs exists, then none of the generated

[14] http://www.geom.uiuc.edu/software/cglist/ch.html.

**Table 6.3** Example of an input file for *Qhull* for both interior and exterior facet generation. The case of two inputs, two outputs, $n$ DMUs, and a CCR model

| | | | |
|---|---|---|---|
| 4 | | | |
| $n+1+4$ | | | |
| $x_{11}$ | $x_{21}$ | $y_{11}$ | $y_{21}$ |
| $x_{12}$ | $x_{22}$ | $y_{12}$ | $y_{22}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{1n}$ | $x_{2n}$ | $y_{1n}$ | $y_{2n}$ |
| 0 | 0 | 0 | 0 |
| $M$ | 0 | 0 | 0 |
| 0 | $M$ | 0 | 0 |
| 0 | 0 | $-M$ | 0 |
| 0 | 0 | 0 | $-M$ |

facets will do). However, the inclusion of all facets with positive output components and negative input components in the normal vector only will provide all interior facets. Furthermore, facets with zero offset are the only ones of interest in the CCR-model.

The three convex hull programs can also be used to generate all interior and exterior facets. Both *cdd* and *lrs* allow directly for envelopment of both extreme points and extreme rays. Hence, using these two codes one simply adds $(s+m)$ additional rays to the input file, namely the $s$ negative ($m$ positive) unit vectors corresponding to the $s$ outputs (the $m$ inputs). The format of the input file for *Qhull* does not presently allow for specification of both extreme points and extreme rays. However, the following procedure remedies this problem. Simply add $(s+m)$ additional vectors to the input file, namely the $s$ negative ($m$ positive) unit vectors corresponding to the $s$ outputs (the $m$ inputs) multiplied by some large (but not too large) number. Consider a situation with two inputs, two outputs and $n$ DMUs. Table 6.3 illustrates the input file[15] for a CCR model, and consists of $2+(n+1)\times 4+4\times 4$ real numbers ($M$ is a large number[16]).

The Ifile in Table 6.4 will generate all interior and exterior facets corresponding to Fig. 6.2a (two inputs and one output).

Part of the information from the output file from *Qhull* is listed in Appendix B together with some explanation of the structure of this output.

Exterior facets for the CCR model can now be identified from the generated list of facets from *Qhull* as the facets which either contain one or more of the vectors $(M,0,0,0)$, $(0,M,0,0)$, $(0,0,-M,0)$, $(0,0,0,-M)$ in combination with one

[15] In Olesen and Petersen (2003), Table 6.3, p. 353 there are typos/errors. In the first two columns of this table inputs are wrongly stated with minus signs. Table 6.3 in this chapter corrects these typos/errors.

[16] Actually, it is recommended only to put one number on each line in the input file to *Qhull*.

**Table 6.4** An Ifile to *Qhull* to generate all interior and exterior facets corresponding to Fig. 6.2a

| | | |
|---|---|---|
| 3 | | |
| 6 | | |
| 120 | 90 | 100 |
| 90 | 90 | 90 |
| 10 | 10 | 10 |
| 100000 | 0 | 0 |
| 0 | 100000 | 0 |
| 0 | 0 | -100000 |

or more of the vertices $\left(x_{1j}, x_{2j}, y_{1j}, y_{2j}\right)$, $j = 1, \ldots, n$ and $(0, 0, 0, 0)$ or equivalently have a normal vector with at least one component equal/close to zero and an offset equal/close to zero.

Notice, that the theoretical correct value of $M$ is $\infty$ and that any finite value of $M$ will only produce exterior facets up to a certain precision. Hence, the normal component equal to zero for exterior facets will be estimated as close to zero as possible for any finite $M$ and as approaching[17] zero for $M \to \infty$.

To summarize, *Qhull* can be used for an identification of all interior and exterior facets from DEA models with constant, decreasing and non increasing returns to scale as follows:

1. Interior facets (FDEFs) in a BCC-DEA production possibility set are determined as all facets with strictly positive output components and strictly negative input components of the normals in a convex hull of all observed input output combinations. Add $s$ (and $m$) additional unit vectors multiplied by $M$ (and $-M$) to the input file to get exterior facets as well. In a BCC-DEA production model we accept exterior facets with a normal vector with all input or all output components equal/close to zero.
2. Interior facets (FDEFs) in a CCR-DEA production possibility set (i.e. the conical hull of all observations set added to $R_+^m \times R_-^s$) is determined as all facets that includes the origin and with strictly positive output components and strictly negative input components of the normals in a convex hull of all observed input output combinations and the origin. Add $m$ (and $s$) additional unit vectors multiplied by $M$ (and $-M$) to the input file to get exterior facets as well. In a CCR-DEA production model we do not accept exterior facets with a normal vector with all

[17] How to specify $M$ relatively to the size of the numbers expressing the inputs and outputs is left for further research. Clearly, there is a trade-off between precision and numerical stability of the results obtained from *Qhull*. If high precision of the normal vectors and the offsets is of importance then one probably should "reestimate" these vectors as soon as the vertices on each facet are identified. Consider e.g. a facet containing the vertices $(0, 0, 0, 0)$, $(x_{11}, x_{21}, y_{11}, y_{21})$, $(M, 0, 0, 0)$, $(0, 0, -M, 0)$. A second stage reestimation of the normal vector would be an estimation of the hyperplane containing the following four points: $(0, 0, 0, 0)$, $(x_{11}, x_{21}, y_{11}, y_{21})$, $(x_{11} + M, x_{21}, y_{11}, y_{21})$, $(x_{11}, x_{21}, 0, y_{21})$.

input or all output components equal/close to zero. At least one component of each has to be positive.

The *Qhull* code was tested in 2003 on a data set from Danish farmers (9 inputs and 4 outputs). A DEA analysis on 399 Danish farms estimated 136 CCR-efficient farms and 190 BCC-efficient farms. The following results are from using *Qhull* to generate all FDEFs:

1. CCR-DEA with 136 CCR efficient DMUs: *Qhull* finds approximately 3600 FDEFs. Generated in 1 h (pentium 2 pro, 0.5 Gb RAM).
2. BCC-DEA with 190 BCC efficient DMUs: *Qhull* finds approximately 80000 FDEFs. Generated in 24 h (HP-9 Unix, 3 Gb RAM).

Using *Qhull* to generate the convex hull of all these efficient observations in this 13 dimensional input output space requires a large amount of memory (>1.5 Gb). We experimented with various ways to decompose the total facet generation into a number of separate generations, each one based upon a subset of the efficient vertices. We tried to separate the computer estimation based on the following idea. For any fixed data point $\left(Y_{j_o}, X_{j_o}\right)$, only include other data points $\left(Y_{j_1}, X_{j_1}\right)$ if a feasible vector of multipliers exists such that both data points are efficient at this common set of multipliers. In relation to the CCR model this means that $\left(Y_{j_1}, X_{j_1}\right)$ is only included in the analysis if $\mathcal{F}^{CCR}\left(\{j_o, j_1\}\right) \neq \varnothing$. However, it turned out that some farms were efficient in a convex combination with each of more than 185 other farms (with a total of 190 BCC efficient farms!!). Anyway, focusing on a fixed data point $\left(Y_{j_o}, X_{j_o}\right)$ allowed us to generate the convex hull based on $\left(Y_{j_1}, X_{j_1}\right), j_1 \in J_1$, where $\mathcal{F}^{CCR}\left(\{j_o, j_1\}\right) \neq \varnothing, \forall j_1 \in J_1$. After generating the convex hull of the subset of the DMUs in $J_1 \cup \{j_o\}$ we deleted all facets that had non-strictly positive normals or defined halfspaces that did not contain all data points from DMUs in $\mathbb{E} \setminus \left(J_1 \cup \{j_o\}\right)$.

We are convinced that the rapid progress in the development of the computer technology will allow us to process larger and larger data sets in larger and larger input output spaces for an identification of all interior and exterior facets. Currently, it may not be possible to generate all FDEFs in very large data sets with say more than 25 inputs and outputs and more than 1000 BCC-efficient DMUs by either one of the codes above. But this limit will be pushed outwards during the years to come.

## 6.6 An Efficiency Evaluation Relative to a Technology Spanned by FDEFs

This section provides an operational procedure for a test of the existence of FDEFs in a data set. In addition, two operational radial input-oriented measures of the distance from a feasible input-output combination to a frontier spanned by FDEFs are suggested, an extended facet measure and a facet constrained DEA-measure. The extended facet index provides a lower bound and the facet constrained DEA-measure an upper bound on an input-oriented efficiency index with the CCR-index in between. In the first subsection we focus on the CRS case and leave the variable returns to scale (VRS) case to the second subsection.

### 6.6.1 The CRS Case: Extending the CCR-Model with Facet Extensions

Let again $\mathbb{E}$ be an index set for the CCR strongly efficient DMUs in a sample of observations. In order to ease the exposition of the main results of this paper we employ the following regularity condition[18] (a test for this condition can be found in Appendix A):

**Condition 1** REGULARITY CONDITION (RC1). Every subset of $s + m - 1$ columns for $\text{DMU}_j$, $j \in \mathbb{E}$ of the data matrix $\begin{bmatrix} Y \\ -X \end{bmatrix}$ is linear independent, where $Y$ is $(s \times N)$, $X$ is $(m \times N)$ and $N = |\mathbb{E}|$.

The following mixed integer linear program provides a test for the existence of at least one FDEF (given RC1):

$$\begin{array}{lll} \min & \sum_{j\in\mathbb{E}} b_j & \\ \text{s.t.} & u^t Y_j - v^t X_j + s_j & = 0 \quad j \in \mathbb{E} \\ & e^t v & = 1 \\ & s_j - b_j M & \leq 0 \quad j \in \mathbb{E} \\ & b_j \text{ binary}, s_j \geq 0, \forall j \in \mathbb{E}, u \geq \varepsilon e, v \geq \varepsilon e, u \in \mathbb{R}^s, v \in \mathbb{R}^m & \end{array} \tag{6.19}$$

where $\varepsilon$ is a non-Archimedian, $M = 1/\varepsilon$, and $e$ is a vector $(1, \cdots, 1)^T$ of an appropriate dimension. The program (6.19) minimizes the sum of binary variables $b_j$, $j \in \mathbb{E}$, i.e., maximizes the number of $b_j$'s equal to zero. It is easily seen that

$$b_j = 0 \Leftrightarrow u^t Y_j - v^t X_j = 0. \tag{6.20}$$

Bearing in mind Definition 1 and RC1, an FDEF thus exists if

$$\sum_{j\in\mathbb{E}} b_j^* = |\mathbb{E}| - (s + m - 1), \tag{6.21}$$

where $b^*$ is an optimal vector from (6.19). Let $\mathcal{J}$ be the family of all subsets of $\mathbb{E}$, and $\mathcal{J} \supseteq \mathcal{J}_{FDEF} \equiv \{J_1, \ldots, J_F\}$ be the subset of subsets such that for $k = 1, \ldots, F$ we have $|J_k| = (s + m - 1)$ and (6.19) has a feasible solution with $s_j = 0$, $j \in J_k$.

[18] The necessity of this regularity condition was pointed out by Professor R. M. Thrall. If a subset of $s + m - 1$ columns is linear dependent, and these columns span an efficient face of the frontier, then the dimension of this efficient face will be strictly less than $s + m - 1$. Hence, this regularity condition allows us to determine the dimension of a particular efficient face directly from the number of DMUs located on this efficient face. A test of an eventual violation of RC1 can be performed by a MILP program included in Appendix A.

Hence, $\dim\left(\mathcal{F}^{CCR}\left(J_k\right)\right) = 1$.[19] Let the $F$ corresponding vectors of multipliers be given by $\left(u_k, -v_k\right), k = 1, \ldots, F$, i.e.,

$$\mathcal{F}^{CCR}\left(J_k\right) = \left\{(u,-v) \mid (u,-v) = \varsigma\left(u_k, -v_k\right), \varsigma > 0\right\}, k = 1, \ldots, F \tag{6.22}$$

Let

$$\mathcal{P}^{CCR}_{FDEF} = \left\{(u,-v) \in \mathbb{R}^s_+ \times \mathbb{R}^m_- \mid (u,-v) = \sum_{k=1}^{F} \lambda_k \left(u_k, -v_k\right), \lambda_k \geq 0, k = 1, \ldots, F, \lambda \neq 0\right\} \tag{6.23}$$

The DMUs in the index set $J_k$ thus span an FDEF. $(u_k, -v_k)$ is the scaled normal vector to the efficient facet generating supporting hyperplane and $\mathcal{P}^{CCR}_{FDEF}$ is the polyhedral cone spanned by the collection of scaled normals to efficient facet generating hyperplanes.

We define the $k$'th extended efficient facet, $EEF^{CCR}(J_k)$, as the intersection of an FDEF-generating supporting hyperplane and the nonnegative orthant:

$$EEF^{CCR}\left(J_k\right) = \left\{(y,x) \in \mathbb{R}^{s+m}_+ \mid u^T y - v^T x = 0, (u,-v) \in \mathcal{F}\left(J_k\right)\right\} \tag{6.24}$$

The corresponding empirical production possibility set, $T^{CCR}_{EEF(J_k)}$ is the intersection between the halfspace generated by the supporting hyperplane and the nonnegative orthant:

$$T^{CCR}_{EEF(J_k)} = \left\{(y,x) \in \mathbb{R}^{s+m}_+ \mid u^T y - v^T x \leq 0, (u,-v) \in \mathcal{F}\left(J_k\right)\right\} \tag{6.25}$$

The extended facet production possibility set generated from the sample $\left(Y_j, X_j\right), j \in \mathbb{E}$, is the intersection of halfspaces defined by FDEF-generating supporting hyperplanes and the nonnegative orthant:

$$T^{CCR}_{EXFA} = \left\{(y,x) \in \mathbb{R}^{s+m}_+ \mid u^T y - v^T x \leq 0, \forall\, (u,-v) \in \mathcal{P}^{CCR}_{FDEF}\right\} = \cap_{k=1}^{F} T^{CCR}_{EFF(J_k)} \tag{6.26}$$

$T^{CCR}_{EXFA}$ is hence by construction a convex piecewise linear envelopment of observed data subject to the condition that substitutional rates along the efficient frontier are well-defined, and determined by data solely. Obviously, $T^{CCR}_{EXFA}$ is a polyhedral set which includes $T^{CCR}$, i.e. $T^{CCR} \subseteq T^{CCR}_{EXFA}$.

The following Extended Facet Efficiency Index for $\mathrm{DMU}_{j_o}$, $j_o = 1, \ldots, N$, measures the radial distance from observation $\left(Y_{j_o}, X_{j_o}\right)$ to the efficient frontier for the

[19] Recall the definition in (6.5):

$$\mathcal{F}^{CCR}\left(\{j\}\right) \equiv \left\{(u,-v) \in \mathcal{P}_{CCR}, u^T Y_j - v^T X_j = 0\right\}.$$

extended facet production possibility set:[20]

$$\begin{array}{lll} \max & u^T Y_{j_o} & \\ \text{s.t.} & u^t Y_j - v^t X_j + s_j & = 0 \quad j \in \mathbb{E} \\ & v^T X_{j_o} & = 1 \\ & s_j - b_j M & \leq 0 \quad j \in \mathbb{E} \\ & \sum_{j \in \mathbb{E}} b_j - (|\mathbb{E}| - (s + m - 1)) & \leq 0 \\ & b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}, u \geq \varepsilon e, u \in \mathbb{R}^s_+, v \geq \varepsilon e, v \in \mathbb{R}^m_+ & \end{array} \tag{6.27}$$

The program (6.27) differs from the DEA-model developed by (Charnes et al. 1978, 1979) by the constraints including the binary $b_j$-variables. In combination these constraints imply, assuming RC1, that any feasible dual price vector must render $(s + m - 1)$ DMUs efficient, i.e. the reference point in the evaluation of the $j_o$'th unit must be positioned on an FDEF. The model (6.27) is related to the Polyhedral Cone-Ratio DEA Model presented in Charnes et al. (1990)[21].

By construction, the extended facet input-oriented efficiency index provides a lower bound on the efficiency rating of the DMU under evaluation.

Next, we introduce The Full Dimensional Efficient Facet Efficiency Index, which provides an upper bound on the efficiency rating. The production possibility set

[20] An Extended Facet Approach has been proposed by Bessent et al. (1988) by the name "Constrained Facet Analysis." One problem with the procedure proposed in that paper is that a given subset of DMUs may span a non FDEF. Hence, it may be impossible to reach a FDEF starting from this given subset of DMUs. The existence of non FDEFs could be the reason why their procedure fails to identify FDEFs in 41.4 percents of the reported runs.

[21] The so-called Polyhedral Cone-Ratio DEA Models are presented in Charnes et al. (1990) along with an application of one of these models to a set of 48 commercial banks. Using the CCR model in the evaluation of these 48 U.S. commercial banks the authors conclude that "the results were not satisfactory so recourse was made to a polyhedral cone-ratio DEA model with results that passed muster in subsequent reviews with wide experience in banking" (p. 86). The transformational matrix used by the authors to form the cones used in the cone-ratio model consists of optimal virtual multiplier vectors of three CCR extreme efficient "model banks."

One problem related to this application is the following: there is in general no unique dual optimal multiplier for a CCR extreme efficient DMU. Hence, the three CCR extreme efficient model banks will probably contribute to the spanning of many different efficient facets. Three of these efficient facets of full dimension are indicated by the three strict positive normal vectors exhibited in Table 6.3, p. 87 in Charnes et al. (1990). However, if the model banks contribute to the spanning of say six different FDEFs then a set of all six strict positive scaled normal vectors to these FDEFs could/should have been used in the specification of the transformation matrix; or if only a subset is wanted then any subset of these six scaled normal vectors could equally well have been used.

In relation to the present paper it is of interest to notice that the extended facet model and the polyhedral cone-ratio model have some common characteristics. In fact, the extended facet MILP program (6.27) can be used to solve the problem related to the application of the cone-ratio model to commercial banks. A cone ratio efficiency analysis based on a transformation matrix consisting of all strict positive optimal virtual multiplier vectors of the three model banks is the result of solving

generated from the $k$'th FDEF is spanned by the subset of DMUs on this efficient facet:

$$T_{FDEF}^{CCR}(J_k) = \left\{ (y,x) \in \mathbb{R}_+^{s+m} \mid \sum_{j \in J_k} \lambda_j Y_j \geq y, \sum_{j \in J_k} \lambda_j X_j \leq x, \lambda_j \geq 0, j \in J_k \right\} \quad (6.28)$$

The Full Dimensional Efficient Facet production possibility set generated from the sample $\left(Y_j, X_j\right), j \in \mathbb{E}$, is the union of production possibility sets spanned by FDEFs:

$$T_{FDEF}^{CCR} = \cup_{k=1}^{F} T_{FDEF}^{CCR}(J_k). \quad (6.29)$$

$T_{FDEF}^{CCR}$ is a possibly nonconvex piecewise linear envelopment of observed data subject to the condition that substitutional rates along the efficient frontier are well defined and determined by data solely. Obviously, $T_{FDEF}^{CCR}$ is a set included in $T^{CCR}$, i.e. $T_{FDEF}^{CCR} \subseteq T^{CCR} \subseteq T_{EXFA}^{CCR}$. The Full Dimensional Efficient Facet Efficiency Index for $\mathrm{DMU}_j, j = 1, \ldots, N$, measures the radial distance from observation $\left(Y_{j_o}, X_{j_o}\right)$ to the efficient frontier for the Full Dimensional Efficient Facet production possibility set:

$$\begin{array}{lll}
\min & \theta^{FDEF} - \varepsilon \left(e^T \sigma^+ + e^T \sigma^-\right) & \\
\text{s.t.} & \sum_{j=1}^{|\mathbb{E}|} \lambda_j X_j - \theta^{FDEF} X_{j_o} + \sigma^- & = 0 \\
& \sum_{j=1}^{|\mathbb{E}|} \lambda_j Y_j - \sigma^+ & = Y_{j_o} \\
& u^t Y_j - v^t X_j + s_j & = 0 \quad j \in \mathbb{E} \\
& v^T X_{j_o} & = 1 \\
& s_j - b_j M & \leq 0 \quad j \in \mathbb{E} \\
& \lambda_j - \left(1 - b_j\right) M & \leq 0 \quad j \in \mathbb{E} \\
& \sum_{j \in \mathbb{E}} b_j - (|\mathbb{E}| - (s + m - 1)) & \leq 0 \\
& b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}, u \geq \varepsilon e, u, \ \sigma^+ \in \mathbb{R}_+^s, v, \sigma^- \geq \varepsilon e, v \in \mathbb{R}_+^m \lambda \in \mathbb{R}_+^{|\mathbb{E}|} &
\end{array} \quad (6.30)$$

---

the following MILP program:

$$\begin{array}{lll}
\max & u^T Y_{j_o} & \\
\text{s.t.} & u^t Y_j - v^t X_j + s_j & = 0 \quad j \in \mathbb{E} \\
& v^T X_{j_o} & = 1 \\
& s_j - b_j M & \leq 0 \quad j \in \mathbb{E} \\
& \sum_{j \in \mathbb{E}} b_j - |\mathbb{E}| - (s + m - 1) & \leq 0 \\
& b_j = 0, j \in \{j_1, j_2, j_3\} & \\
& b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}, u \geq \varepsilon e, u \in \ \mathbb{R}_+^s, v \geq \varepsilon e, v \in \mathbb{R}_+^m &
\end{array}$$

if we assume the regularity condition RC1. M is a large scalar, $j_o \in \{1, \ldots, 48\}$ is the index of the bank being evaluated, $\mathbb{E}$ is the index set of all the CCR extreme efficient banks among the 48 banks, and $j_i \in \mathbb{E}, i = 1, 2, 3$ is the index of the three model banks.

where $\varepsilon$ is a non-Archimedian and $M = 1/\varepsilon$. The program (6.30) differs from the Extended Facet model in (6.27) by the linked constraints including the $s_j$ and the $\lambda_j$ variables, both constrained by the same binary $b_j$ -variables. In combination these constraints imply that the efficiency estimation is performed as a DEA analysis with reference to a production possibility set spanned by some subset of $(s + m - 1)$ CCR extreme efficient DMUs, all located on the same FDEF. The MILP program identifies the particular subset of $(s + m - 1)$ CCR extreme efficient DMUs, which maximizes the potential radial contraction of inputs from $\text{DMU}_{j_o}$.

The production possibility set $T^{CCR}_{FDEF}$ includes by construction all observed input output combinations located on FDEFs while observations (CCR-inefficient or CCR-efficient) outside every FDEF may or may not belong to this set.

For DMUs located outside $T^{CCR}_{FDEF}$ we have $\theta^{FDEF} > 1$ which can be interpreted as the minimum proportional increase in the input vector that is required in order to move the observation into $T^{CCR}_{FDEF}$. An upper bound above one for a CCR-efficient or CCR-inefficient DMU indicates that this DMU cannot be dominated on the Pareto criterion by a reference point at an FDEF. Hence, this DMU is located in an area of the input output space, where we refrain from estimating the frontier of $T^{CCR}_{FDEF}$ because of lack of comparable data. We don't know whether or not this DMU is located on the frontier of $T^{CCR}_{FDEF}$, because data does not support an estimation of the frontier in a neighborhood around this observation. Since the location of the frontier is unknown around this observation we assign an upper bound above one.

Table 6.5 summarizes the six different relations between the efficiency indices from the three models: EXFA, CCR and FDEF. Since $T^{CCR}_{EXFA} \supseteq T^{CCR} \supseteq T^{CCR}_{FDEF}$ we have $\theta^{EXFA} \leq \theta^{CCR} \leq \theta^{FDEF}$. If the $j$'th DMU is located on an FDEF, then $\theta^{EXFA} = \theta^{CCR} = \theta^{FDEF} = 1$. If the j'th DMU is CCR-inefficient and dominated by a reference unit located on an FDEF, i.e. the $\theta^{CCR}$-projection of the input output combination: $\left(\theta^{CCR} X_j, Y_j\right)$ is located on an FDEF, then $\theta^{EXFA} = \theta^{CCR} = \theta^{FDEF} < 1$. If the j'th DMU is CCR inefficient and the $\theta^{CCR}$-projection of the input output combination: $\left(\theta^{CCR} X_j, Y_j\right)$ is not located on an FDEF but is dominated by DMUs all located on an FDEF, then we will typically have $\theta^{CCR} = \theta^{FDEF} < 1$. If the j'th DMU is CCR inefficient and dominated by a reference point located on a non FDEF, then $\theta^{EXFA} < \theta^{CCR}$ and $\theta^{CCR} < \theta^{FDEF} < 1$.

In Appendix C we have included a simple example which illustrates the three technologies EXFA, CCR and FDEF and the six different relations in Table 6.5

**Table 6.5** Combinations of efficiency scores from the three models CCR extreme efficient DMUs

| | |
|---|---|
| CCR extreme efficient DMUs $\theta^{CCR} = 1$ | $\theta^{EXFA} = \theta^{CCR} = \theta^{FDEF} (= 1)$ |
| | $\theta^{EXFA} < \theta^{CCR} < \theta^{FDEF} (< 1)$ |
| | |
| CCR extreme efficient DMUs $\theta^{CCR} < 1$ | $\theta^{EXFA} = \theta^{CCR} = \theta^{FDEF}$ |
| | $\theta^{EXFA} < \theta^{CCR} = \theta^{FDEF} (< 1)$ |
| | $\theta^{EXFA} < \theta^{CCR} < \theta^{FDEF} (< 1)^1$ |
| | $\theta^{EXFA} < \theta^{CCR} < \theta^{FDEF} (> 1)^2$ |

between the efficiency indices from the three models. Furthermore, the example illustrates the geometry behind the case where the frontier consists of both FDEFs and NFDEFs.

Notes: 1) + 2) These combinations can e.g. emerge if the $\theta^{CCR}$ -projection of the input output combination: $\left(\theta^{CCR}X_j, Y_j\right)$ is dominated by a reference set which includes a CCR extreme efficient DMU not located on any FDEFs. Case 1) occurs if the output input combination $\left(Y_j, -X_j\right) \leq (y, -x)$ for some $(y, -x)$ on an FDEF. Case 2) occurs if that is not the case. For details, see Appendix C.

### 6.6.2 *The VRS Case: Extending the BCC-Model with Facet Extensions*

Let $\mathbb{E}^{BCC}$ be an index set for the BCC strongly efficient DMUs in the sample. Extending the BCC-model with facet extensions requires a similar regularity condition as in Sect.6.6.1:

**Condition 2** REGULARITY CONDITION (RC2). Every subset of $s+m$ columns for DMU$_j$, $j \in \mathbb{E}^{BCC}$ of the data matrix $\begin{bmatrix} -X \\ Y \end{bmatrix}$ is linear independent, where $Y$ is $(s \times N)$, $X$ is $(m \times N)$ and N=$\left|\mathbb{E}^{BCC}\right|$

The following mixed integer linear program provides a test for the existence of at least one FDEF (given RC2):

$$\begin{array}{lll} \min & \sum_{j\in\mathbb{E}} b_j & \\ \text{s.t.} & u^t Y_j - v^t X_j + v_o + s_j & = 0 \quad j \in \mathbb{E}^{BCC} \\ & e^t v & = 1 \\ & s_j - b_j M & \leq 0 \quad j \in \mathbb{E}^{BCC} \\ & b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}^{BCC} & , u \geq \varepsilon e, v \geq \varepsilon e, u \in \mathbb{R}^s, v \in \mathbb{R}^m \quad v_o \in \mathbb{R} \end{array} \tag{6.31}$$

where $\varepsilon$ is a non-Archimedian, $M = 1/\varepsilon$, and $e$ is a vector $(1, \cdots, 1)^T$ of an appropriate length. The program (6.31) minimizes the sum of binary variables $b_j$, $j \in \mathbb{E}^{BCC}$, i.e., maximizes the number of $b_j$'s equal to zero. It is easily seen that

$$b_j = 0 \Leftrightarrow u^t Y_j - v^t X_j + v_o = 0. \tag{6.32}$$

Bearing in mind Definition 3 and RC2, an FDEF thus exists if

$$\sum_{j\in\mathbb{E}} b_j^* = \left|\mathbb{E}^{BCC}\right| - (s+m), \tag{6.33}$$

where $b^*$ is an optimal vector from (6.31). Let $\mathcal{J}$ be the family of all subsets of $\mathbb{E}^{BCC}$, and $\mathcal{J} \supseteq \mathcal{J}_{FDEF}^{BCC} \equiv \{J_1, \ldots, J_F\}$ be the subset of subsets such that for $k = 1, \ldots, F$

we have $|J_k| = (s+m)$ and (6.31) has a feasible solution with $s_j = 0, j \in J_k$. Hence, $\dim(\mathcal{F}(J_k)) = 1$. Let the $F$ corresponding vectors of multipliers and intercept term be given by $(u_k, -v_k, v_{ok}), k = 1, \ldots, F$, i.e.,

$$\mathcal{F}(J_k) = \{(u, -v, v_o) \mid (u, -v, v_o) = \varsigma(u_k, -v_k, v_{ok}), \varsigma > 0\}, k = 1, \ldots, F \tag{6.34}$$

Let

$$\mathcal{P}_{FDEF}^{BCC} = \Big\{ (u, -v, v_o) \in \mathbb{R}_+^s \times \mathbb{R}_-^m \times \mathbb{R} \mid (u, -v, v_o) = \sum_{k=1}^{F} \lambda_k (u_k, -v_k, v_{ok}), \lambda_k \geq 0, k = 1, \ldots, F, \lambda \neq 0 \Big\} \tag{6.35}$$

The DMUs in the index set $J_k$ thus span an FDEF, $(u_k, -v_k)$ is the scaled normal vector to the efficient facet generating supporting hyperplane, $v_{ok}$ is the scaled intercept term and $\mathcal{P}_{FDEF}^{BCC}$ is the polyhedral cone spanned by the collection of scaled normals and scaled intercepts to efficient facet generating hyperplanes.

We define the $k$'th extended efficient facet, $EEF^{BCC}(J_k)$, as the intersection of an FDEF-generating supporting hyperplane and the nonnegative orthant:

$$EEF^{BCC}(J_k) = \left\{(y, x) \in \mathbb{R}_+^{s+m} \mid u^T y - v^T x + v_o = 0, (u, -v, v_o) \in \mathcal{F}(J_k)\right\} \tag{6.36}$$

We now follow the same approach as in the previous section on CRS using $EEF^{BCC}(J_k)$ to define the corresponding empirical production possibility set, $T_{EEF(J_k)}^{BCC}$ as the intersection between the halfspace generated by the supporting hyperplane and the nonnegative orthant:

$$T_{EEF(J_k)}^{BCC} = \left\{(y, x) \in \mathbb{R}_+^{s+m} \mid u^T y - v^T x + v_o \leq 0, (u, -v, v_o) \in \mathcal{F}(J_k)\right\} \tag{6.37}$$

However, with VRS we need to make sure that no output vector $y \geq 0, y \neq 0$ exists, such that $(y, 0) \in \cap_{k=1}^{F} T_{EFF(J_k)}^{BCC}$. If such an output vector $y \geq 0, y \neq 0$ exists, then the output input combination $(y, 0)$ violates the axiom of "no free lunch", if we follow the approach used in CRS case in the previous subsection and define the extended facet production possibility set $T_{EXFA}^{BCC}$ as $T_{EXFA}^{BCC} = \cap_{k=1}^{F} T_{EFF(J_k)}^{BCC}$. Two different paths seem to be possible to make sure that "no free lunch" is not violated by $T_{EXFA}^{BCC} = \cap_{k=1}^{F} T_{EFF(J_k)}^{BCC}$. One possibility is to assume the following regularity condition RC3:

**Condition 3** REGULARITY CONDITION (RC3)[22]. There exists $\kappa \in \{1, \ldots F\}$ such that $(0, 0) \notin T_{EFF(J_\kappa)}^{BCC}$.

[22] Notice, that regularity condition RC3 is satisfied if we have at least one FDEF with an intercept term being strictly positive.

Alternatively, we can restrict the validity of $T_{EXFA}^{BCC}$ in input space to the convex hull of the observed input vectors, i.e. we define $T_{EXFA}^{BCC}$ as

$$T_{EXFA}^{BCC} = \left(\cap_{k=1}^{F} T_{EFF(J_k)}^{BCC}\right) \cap \left(\mathbb{R}_+^s \times conv\left\{X_1, \ldots, X_n\right\}\right) \tag{6.38}$$

Assuming the regularity condition RC3 the extended facet production possibility set generated from the sample $\left(Y_j, X_j\right), j \in \mathbb{E}^{BCC}$, can be defined as the intersection of halfspaces defined by FDEF-generating supporting hyperplanes and the non-negative orthant:

$$\begin{aligned} T_{EXFA}^{BCC} &= \left\{(y,x) \in \mathbb{R}_+^{s+m} \mid u^T y - v^T x + v_o \leq 0, \forall\, (u,-v,v_o) \in \mathcal{P}_{FDEF}^{BCC}\right\} \\ &= \cap_{k=1}^{F} T_{EFF(J_k)}^{BCC} \end{aligned} \tag{6.39}$$

$T_{EXFA}^{BCC}$ is hence by construction a convex piecewise linear envelopment of observed data subject to the condition that substitutional rates along the efficient frontier are well-defined, and determined by data solely. Obviously, $T_{EXFA}^{BCC}$ is a polyhedral set which includes $T^{BCC}$, i.e. $T^{BCC} \subseteq T_{EXFA}^{BCC}$. Alternatively, if RC3 is violated we defined $T_{EXFA}^{BCC}$ as indicated in (6.38)

Assuming that RC3 is not violated, the following Extended Facet Efficiency Index for $\text{DMU}_{j_o}$, $j_o = 1, \ldots, N$, measures the radial distance from observation $\left(Y_{j_o}, X_{j_o}\right)$ to the efficient frontier for the extended facet production possibility set

$$\begin{array}{lll} \max & u^T Y_{j_o} + v_o & \\ \text{s.t.} & u^t Y_j - v^t X_j + v_o + s_j = 0 & j \in \mathbb{E}^{BCC} \\ & v^T X_{j_o} = 1 & \\ & s_j - b_j M \leq 0 & j \in \mathbb{E}^{BCC} \\ & \sum_{j \in \mathbb{E}^{BCC}} b_j - \left(\left|\mathbb{E}^{BCC}\right| - (s+m)\right) \leq 0 & \\ & b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}^{BCC}, u \geq \varepsilon e,\, u \in \mathbb{R}_+^s, v \geq \varepsilon e, v \in \mathbb{R}_+^m & \end{array} \tag{6.40}$$

The program (6.40) differs from the DEA-model developed by (Banker et al. 1984) by the constraints including the binary $b_j$-variables. In combination these constraints imply, assuming RC2, that any feasible dual price vector must render $(s+m)$ DMUs efficient, i.e. the reference point in the evaluation of the $j_o$'th unit must be positioned on an FDEF.

By construction, the extended facet input-oriented efficiency index provides a lower bound on the efficiency rating of the DMU under evaluation. The Full Dimensional Efficient Facet Efficiency Index provides an upper bound on the efficiency rating. The production possibility set generated from the $k$'th FDEF is spanned by the subset of DMUs on this efficient facet:

$$T_{FDEF}^{BCC}(J_k) = \left\{(y,x) \in \mathbb{R}_+^{s+m} \mid \sum_{j \in J_k} \lambda_j Y_j \geq y, \sum_{j \in J_k} \lambda_j X_j \leq x, \sum_{j \in J_k} \lambda_j = 1, \lambda_j \geq 0, j \in J_k\right\} \tag{6.41}$$

The Full Dimensional Efficient Facet production possibility set generated from the sample $\left(Y_j, X_j\right), j \in \mathbb{E}$, is the union of production possibility sets spanned by FDEFs:

$$T_{FDEF}^{BCC} = \cup_{k=1}^{F} T_{FDEF}^{BCC}(J_k). \tag{6.42}$$

$T_{FDEF}^{BCC}$ is a possibly nonconvex piecewise linear envelopment of observed data subject to the condition that substitutional rates along the efficient frontier are well-defined and determined by data solely. Obviously, $T_{FDEF}^{BCC}$ is included in $T^{BCC}$, i.e. $T_{FDEF}^{BCC} \subseteq T^{BCC} \subseteq T_{EXFA}^{BCC}$. The Full Dimensional Efficient Facet Efficiency Index for $\mathrm{DMU}_j, j = 1, \ldots, N$, measures the radial distance from observation $(Y_k, X_k)$ to the efficient frontier for the Full Dimensional Efficient Facet production possibility set:

$$\begin{array}{lll}
\min & \theta^{FDEF} - \varepsilon\left(e^T\sigma^+ + e^T\sigma^-\right) & \\
\text{s.t.} & \sum_{j=1}^{\left|\mathbb{E}^{BCC}\right|} \lambda_j X_j - \theta^{FDEF} X_{j_o} + \sigma^- & = 0 \\
& \sum_{j=1}^{\left|\mathbb{E}^{BCC}\right|} \lambda_j Y_j - \sigma^+ & = Y_{j_o} \\
& \sum_{j=1}^{\left|\mathbb{E}^{BCC}\right|} \lambda_j & = 1 \\
& u^t Y_j - v^t X_j + v_o + s_j & = 0 \quad j \in \mathbb{E}^{BCC} \\
& v^T X_{j_o} & = 1 \\
& s_j - b_j M & \leq 0 \quad j \in \mathbb{E}^{BCC} \\
& \lambda_j - \left(1 - b_j\right) M & \leq 0 \quad j \in \mathbb{E}^{BCC} \\
& \sum_{j \in \mathbb{E}^{BCC}} b_j - \left(\left|\mathbb{E}^{BCC}\right| - (s+m)\right) & \leq 0 \\
& b_j \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}, u \geq \varepsilon e, u, \sigma^+ \in \mathbb{R}_+^s, v, \sigma^- \geq \varepsilon e, v \in \mathbb{R}_+^m \lambda \in \mathbb{R}_+^{|\mathbb{E}|} &
\end{array} \tag{6.43}$$

where $\varepsilon$ is a non-Archimedian and $M = 1/\varepsilon$. The program (6.43) differs from the Extended Facet model in (6.40) by the linked constraints including the $s_j$ and the $\lambda_j$ variables, both constrained by the same binary $b_j$ -variables. In combination these constraints imply that the efficiency estimation is performed as a DEA analysis with reference to a production possibility set spanned by some subset of $(s + m)$ BCC extreme efficient DMUs, all located on the same FDEF. The MILP program identifies the particular subset of $(s + m)$ BCC extreme efficient DMUs, which maximizes the potential radial contraction of inputs from $\mathrm{DMU}_{j_o}$.

In Appendix D we have included a simple example which illustrates the three technologies EXFA(BCC), BCC and FDEF(BCC). Furthermore, the example illustrates the geometry behind the case where the frontier consists of both FDEFs and NFDEFs.

## 6.7 Use of Efficient Faces and Facets in DEA

We have argued that a CCR- or a BCC-estimation does not in general provide a strongly efficient frontier with well defined marginal rates of substitution. The optimal virtual multipliers should for that reason be interpreted with care, when DEA is used for estimation of efficiency scores and local substitution or scale characteristics. The estimates may well be affected by the axiom of strong disposability of inputs and outputs, and they are not determined by observed data only as in the extended facet approach. The estimation of lower and upper bounds for efficiency scores relative to a strongly efficient frontier with well defined marginal rates of substitution and determined by observed data only is an important use of the extended facet approach.

Strong disposability means that the polyhedral empirical production possibility set includes weakly efficient or exterior facets characterized by zero components in their normal vectors. Zero multipliers are a source of trouble when estimating marginal rates of substitution or transformation, see (Cooper et al. 2011, Chap. 4). Zero multipliers also mean that some of the inputs and/or outputs are eventually ignored. Moreover, zero multipliers may imply non-zero slacks in the primal envelopment form, which means that the unit under assessment using a radial measure is evaluated with reference to a point that is not strongly efficient. Zero values in an optimal solution for a primal or dual LP program are indicators of degeneracy and multiple optimal solutions. The optimal LP-solution for a CCR- or a BCC-model in input-output space is highly degenerate for any extreme efficient DMU, which implies that alternative optimal solutions prevail in multiplier space. (Cooper et al. 2007) discuss the choice of weights from alternative optimal solutions of dual multiplier models in DEA.

Imposing strictly positive lower bounds on the multipliers in the dual formulation of the model is an easy way to avoid zero weights. (Charnes et al. 1979) introduced non-Archimedian concepts to exclude zero weights based upon the assumption "that a reduction in any input or an expansion of any output has some value". The non-Archimedian approach is perfectly consistent with DEA as a value free and highly conservative procedure, where the unit under assessment is put in the best possible position regarding its distance to the production frontier. The approach guarantees that reference points are located on strongly efficient segments of the frontier and provides a firm identification of the set of strongly efficient DMUs. However, the approach can be argued not to produce efficiency scores that can be readily used, because inefficiencies in terms of strictly positive slacks are accounted for in terms of a non-Archimedian weighting only.

Different approaches for avoiding zero weights have been reported in the literature, see (Cooper et al. 2011, Chap. 4) for a survey. A number of these approaches involve the incorporation of price information reflecting meaningful trade-offs, value information, and managerial goals into the analysis by imposing upper and lower bounds on the (relative) multipliers, which in turn means that values based on economic or other considerations are introduced into DEA. The work on facet models and the extension of FDEFs of the frontier can be seen as a highly different approach

for addressing the problems with zero weights and consequently non-zero slacks. The distinguishing feature of the facet approach compared to the incorporation of bounds in multiplier space is that DEA is maintained value free.

Focus in the remainder of this section is on other uses of facet models reported in the literature.

As observed by Cook and Seiford (2009), significant work has been done relating to facet identification and facet extension. Bessent et al. (1988) were the first to introduce the idea of so-called constrained facet analysis. They observe that a review of reported studies in the literature reveals that DEA-solutions in some cases produce efficiency ratings and marginal rates of substitution that are difficult to interpret and unacceptable to unit managers. The problem is argued to arise when an inefficient unit has a mix of inputs and/or outputs that is different from any frontier point. Such units are referred to as not naturally enveloped inefficient units. By contrast, an inefficient DMU is said to be naturally enveloped by a complete frontier facet, if its frontier reference group has $s + m - 1$ observations.

In the case that a unit under assessment is projected to an exterior facet - so that the reference group includes less than $s+m-1$ DMUs - the suggested procedure attempts to augment this reference group iteratively, until a reference set of an appropriate cardinality is identified. Each DMU added to the reference group results in the elimination of a slack in input-output space accompanied by the elimination of a zero weight in multiplier space.

The algorithm is based upon linear programming procedures only and allows for no backtracking; once a DMU has entered the reference group, it is not allowed to be removed from that group again. It should come as no surprise that the procedure may well terminate with slacks in the optimal basis in input-output space and zero weights in multiplier space, since there is no guarantee that the involved sets of DMUs of cardinality less than $s + m - 1$ contribute to the spanning of one (or more) FDEFs. There is no guarantee either that the procedure will identify an FDEF, if one exists. And if the procedure does terminate with the identification of an FDEF, there is no guarantee that it will be the one with the minimum distance to the DMU under assessment.

Lang et al. (1995) improved on the idea in terms of a two-stage approach. The integer programming procedures suggested by Green et al. (1996) and Olesen and Petersen (1996) guarantee radial projections against FDEFs.

The basic DEA models uses an input or output oriented radial measure of efficiency projecting the unit under assessment to a point on the frontier with the same mix of inputs and outputs as that of the unit under assessment. The conservation of the mix is the characteristic that makes the resulting distance measure radial. The perceived arbitrariness in imposing targets preserving the mix may be considered a weakness of radial measures, since a firm's very reason to change its input or output levels may well be an intention to change that mix, Chambers and Mitchell (2001).

The identification of targets is to be considered one of the key practical outcomes in an efficiency assessment. Radial measures may - unless followed by a slack maximizing procedure - find targets on the weakly efficient segments of the frontier. Non-radial measures designed for an identification of targets at the strongly efficient

segments of the frontier have for that reason been developed. The hyperbolic measure suggested by Färe et al. (1985), the Russell-measure suggested by Färe and Lovell (1978) and the additive model due to Charnes et al. (1985) are early examples of DEA-based non-oriented and non-radial efficiency measures. The directional distance function introduced by Chambers et al. (1996, 1998) is another non-oriented measure.

As observed by Portela et al. (2003), the non-oriented DEA-models share the common feature of maximizing slacks. Consequently, the targets identified by these models are in this sense those furthest away from rather than those closest to the unit under assessment. Takeda and Nishimo (2001), Briec and Leleu (2003) and Aparicio et al. (2007) suggest for this reason efficiency to be measured compared to so-called close targets. Fukuyama and Sekitani (2012) label this kind of approach a minimum distance model.

It is well known that models of this kind face a computational difficulty caused by the requirement that targets must be located on the strongly efficient frontier. It is also well known that a complete description of the strongly efficient frontier in terms of its FDEFs and its strongly efficient facets not positioned at an FDEF provides a remedy to this difficulty. These observations are the starting point for the decomposition of the efficient frontier of the DEA production possibility set into a set of so-called Maximal Efficient Faces (MEF) suggested by (Fukuyama and Sekitani 2012). An MEF is in terms of Definitions 2 and 4 either an FDEF or an NFDEF.

The underlying idea is that an explicit identification of the complete set of exterior facets in terms of its defining hyperplanes is not needed for a complete description of the strongly efficient production frontier. An identification of the set of FDEFs and the set of NFDEFs is sufficient. The strongly efficient frontier is the collection of all MEFs, or equivalently the collection of all FDEFs and all NFDEFs. A mixed integer programming procedure is suggested for the identification of all MEFs. Results based upon three real-world data sets related to international airline companies, Japanese banks, and Japanese soccer players are reported. One of the findings is that about 40% or more of the MEFs were not FDEFs. This result highlights the observation above that the optimal virtual multipliers must be interpreted with care, when DEA is used for estimation of efficiency scores and local substitution or scale characteristics, since zero values are involved.

Portela et al. (2003) is an example of a minimum distance model. Traditional non-oriented DEA-models and those based upon them are criticized either for imposing strong restrictions on the movements towards the efficient frontier or for aiming at maximizing slacks rather than looking for so-called close targets. The following measure developed by (Brockett et al. 1997) (BRWZ) is argued to be more appropriate. The *BRWZ*-measure allows each input and each output to change by different proportions, since $h_{io}, i = 1, ...., m$ and $g_{ro}, r = 1, ...., s$ are measures of the relative change in input $i = 1, ...., m$ and output $r = 1, ...., s$. Observe that changes in inputs

and outputs are coupled multiplicatively:

$$BRWZ_o = \frac{\sum_{i=1}^{m} h_{io} \times \sum_{r=1}^{s} 1/g_{ro}}{m \times s}$$

Minimizing $BRWZ_o$ is the only way to assure that the targets are positioned at the Pareto-efficient frontier and that efficiency is measured. However, one needs to maximize $BRWZ_o$ while at the same time assuring projection on the efficient frontier in order to find the closest targets.

$BRWZ_o$ is maximized over all facets - one at the time - in the empirical production possibility set in order to make sure that the maximum $BRWZ_o$ compares to a projection on the efficient frontier. The procedure can be summarized as follows:

1. Determine the set of Pareto efficient units by solving the additive model.
2. Identify all Pareto-efficient facets (and faces) $F_k, k = 1, ...., K$, using *Qhull*.
3. For each $k = 1, ...., K$, maximize $BRWZ_o$ subject to the requirement that $(h_o X_o, g_o Y_o) \in F_k$

The reference set positioned at the facet giving rise to the maximum $BRWZ_o$ is argued to provide the closest target.

Close targets are believed to be more relevant for the units under assessment to attain and more in line with the way management exercise judgment in general. The concept of close targets or minimum distance models require an a priori identification of all facets. It has been applied by Portela and Thanassoulis (2005) in a study of the profitability of a sample of Portuguese bank branches and its decomposition into technical and allocative components. The sample includes two inputs, four outputs and 60 DMUs. The model is an example of a design, where an a priori identification of all facets—interior as well as exterior—is necessary. *Qhull* provides an appropriate tool for that purpose.

The Russell output measure of technical efficiency involves a maximization of the sum of the relative proportional expansion rate of output $r = 1, ...., s$ divided by number of outputs:

$$\Gamma(X_o, Y_o) = \max \left\{ \frac{1}{s} \sum_{r=1}^{s} \Phi_r : \Phi Y_o \in P(X_o), \Phi_r \geq 1, r = 1, ...., s \right\}$$

This measure can be shown to satisfy the following set of desirable properties, see (Färe et al. 1985):

1. $0 < 1/\Gamma(X_o, Y_o) \leq 1$.
2. $1/\Gamma(X_o, Y_o) = 1$ if and only if $Y_o \in \partial^s(P(X_o))$.
3. $1/\Gamma(X_o, Y_o)$ is units invariant.
4. $1/\Gamma(X_o, Y_o)$ is strongly monotonic in outputs.

Aparicio and Pastor (2013) suggest a minimum distance version of the output oriented Russell measure by replacing the max -operator with a min-operator in order to

determine closest targets instead of furthest targets. It is demonstrated that the new measure lacks the property of strong monotonicity when applied in the context of the BCC-model. The problem is shown to be related to the fact that, in general, not all facets of an empirical production possibility set are FDEFs. Then, resorting to the notion of an extended facet production possibility set, the new version of the Russell index is shown to satisfy the complete set of desirable properties including strong monotonicity in outputs.

The suggested procedure for an estimation of the index is similar to the one suggested by Portela et al. (2003) augmented with a fourth step designed to provide an explicit identification of the closest targets and with Step 2 modified in terms of an identification of all FDEFs using *Qhull*. The paper includes a small empirical example with 19 DMUs, one input and three outputs. The model is an example of a design, where the extended facet model provides a highly appropriate framework, where an a priori identification of all FDEFs is necessary, and where *Qhull* provides an appropriate tool for that purpose.

The cross efficiency score of a given DMU is obtained by computing for that DMU the set of $N$ efficiency scores using the $N$ sets of optimal weights corresponding to the $N$ DMUs in the sample and then averaging those scores. Thus, cross efficiency goes beyond the pure self evaluation inherent in conventional DEA analysis, and combines this with the other $(N - 1)$ scores arising from the optimal peer multipliers.

Doyle and Green (1994a) point out that the non-uniqueness of the optimal DEA multipliers possibly reduces the usefulness of cross efficiency and suggest various secondary goals such as given by the so-called aggressive and benevolent models. Appa et al. (2006) examine the aggressive formulations of four cross evaluation approaches taken from Doyle and Green (1994a), and argue that the availability of all alternative multiplier solutions allows for the construction of a complete cross efficiency matrix of dimension $N \times K$ thus enhancing cross evaluation analyses.

The availability of all possible alternative multiplier solutions translates into a requirement of complete frontier information in terms of the set of normal vectors for all facets defining the possibility set - interior as well as exterior. The procedure is for this reason an example of the possibilities made available by the dual representation of the technology as well as the advantages to be gained by identifying all facets of the production possibility set.

The dual representation of polyhedral possibility sets in intersection form corresponding to an envelopment of observed data by FDEFs as in Olesen and Petersen (1996) or by interior as well as exterior facets as in this chapter provides a highly appropriate framework for studying the characteristics of the resulting production frontier and answering questions relating to the two key features of efficient production, substitution properties and scale properties:

- Which are the trade-offs along isoquants?
- How sensitive is the mix of inputs and outputs to changes in relative prices?
- What is Most Productive Scale Size for any given mix of inputs and outputs?
- What is The Expansion Path for a given set of relative prices?
- How does the elasticity of scale change for any given mix of inputs and outputs?

Olesen and Petersen (2003) demonstrate that complete information on the facial structure of an empirical possibility set in terms of an identification of all (internal) facets can be used for a characterization of the underlying data generation process, an estimation of isoquants and relevant elasticities of substitution, and a specification of appropriate constraints on the virtual multipliers in a Cone Ratio model (Charnes et al. 1990). Accordingly, questions of the following types can also be answered:

- Which are the characteristics of the underlying data generation process?
- Is there a local bias in efficiency scores, and—if yes–does it differ between segments of the input-output space?
- Does the sample include outliers?

Førsund et al. (2007) provide a method that can be used to numerically evaluate scale elasticity at any point on the DEA surface using an idea introduced in Krivonozhko et al. (2004). The starting point is a standard neoclassical production or transformation function representing the efficient input-output configurations assumed to be continuously differentiable, increasing in outputs and decreasing in inputs thus exhibiting free disposability such that

$$F(x,y)=0, \frac{\partial F(x,y)}{\partial y_r}>0, r=1,...,s, \frac{\partial F(x,y)}{\partial x_i}<0, i=1,...,m,$$

where the production possibility set $T$ is given as $T=\{x,y|F(x,y)\leq 0\}$. Scale elasticity, as a function of inputs and outputs, is defined as the marginal change in an output expansion factor caused by a marginal change in an input expansion factor over the average ratio.

The polyhedral DEA-frontier of dimension $s+m-1$ is clearly not differentiable. However, one-sided directional derivatives can be shown to exist at every point and in every direction obtained by cutting through the DEA-frontier with a two-dimensional plane for any fixed input mix and any fixed output mix. This result allows for a calculation of the scale elasticity at any point along the intersection of this plane and the frontier with the relevant one-sided derivatives calculated as the slope of the appropriate curves defining the resulting piecewise linear segment of the frontier.

The procedure does not require an a priori identification of all facets. However, the building of intersections of the boundary and the relevant two-dimensional hyperplanes is easy, if facets are known.

We observe that the path along the frontier emanating in the origin and passing through any given input-output configuration may well include segments of exterior facets; in fact, paths of this type may well traverse no FDEFs at all. Accordingly, the computed scale elasticities may relate to weakly efficient segments of the frontier. This is in some conflict with the assumption that the transformation function is increasing in outputs and decreasing in inputs. One may argue that the incorporation of non-Archimedians provides a solution to the problem. However, a calculation of scale elasticities based upon an extended facet specification of a VRS-technology may be considered a more appropriate remedy, since it is based upon observed data, and since the resulting frontier is made up of strongly efficient segments only. A

calculation of scale elasticities based upon an extended facet technology most certainly requires an a priori specification of all FDEFs corresponding to the underlying VRS-frontier.

We also observe that a similar argument applies regarding an estimation of trade-offs along isoquants.

Asmild et al. (2013) provides the final example of use of the facet structure in DEA to be reported here. The paper introduces a test whether observed data points are over- (or under-) represented in certain zones of the input-output space. The test is based upon a comparison of the number of observations located in a certain zone with the expected number of observations in that zone.

The definition of zones is based upon the facet structure. The empirical facets are argued to provide a natural discretization of the range of input (and output) mixes. Each zone is in turn sliced into smaller convex subsets each of which mimics the properties of the underlying facet on the efficient frontier. The decomposition of the possibility set into zones followed by a slicing of zones provides the foundation for a discrete approximation of the distribution of efficiency scores.

The procedure requires an identification of facets of the possibility set followed by a decomposition of the possibility set into $K$ convex subsets - one for each facet - each of which is finally sliced in accordance with the structural properties of the underlying facet. Facets on the production frontier as well as the volumes of the slices are found using *Qhull*. Accordingly, the procedure is an example that information on the complete facet structure of an empirical possibility set can be used for other purposes than an efficiency estimation or an analysis of the properties of the frontier and that *Qhull* is a powerful tool for the analysis of polyhedral sets.

## Appendix A

Let $Z_j = \left(Y_j^T, X_j^T\right)^T \; j \in \mathbb{E}$. The following mixed integer linear program tests for the regularity condition RC1:

$$
\begin{array}{lllll}
\min & \sum_{j\in\mathbb{E}} b_j^+ + \sum_{j\in\mathbb{E}} b_j^- & & & (9)\\
\text{s.t.} & \sum_{j\in\mathbb{E}} \lambda_j^+ Z_j - \sum_{j\in\mathbb{E}} \lambda_j^- Z_j & = 0 & & (9a)\\
& \sum_{j\in\mathbb{E}} b_j^+ + \sum_{j\in\mathbb{E}} b_j^- - (s+m-1) & \leq 0 & & (9b)\\
& \sum_{j\in\mathbb{E}} b_j^+ + \sum_{j\in\mathbb{E}} b_j^- & \geq 1 & & (9b)\\
& M^{-1} b_j^+ \leq \lambda_j^+ \leq M b_j^+ & & j \in \mathbb{E} & (9c)\\
& M^{-1} b_j^- \leq \lambda_j^- \leq M b_j^- & & j \in \mathbb{E} & (9d)\\
& b_j^+ + b_j^- & \leq 1 & j \in \mathbb{E} & (9e)\\
& b_j^+, b_j^- \text{binary}, s_j \geq 0, \forall j \in \mathbb{E}, \lambda_j^+, \lambda_j^- \in \mathbb{R}_+^n & & &
\end{array}
$$

**Table B.1** *Ifile* to *Qhull* to generate all interior and exterior facets corresponding to Fig. 6.2a

| | | |
|---|---|---|
| 3 | | |
| 6 | | |
| 120 | 90 | 100 |
| 90 | 90 | 90 |
| 10 | 10 | 10 |
| 100000 | 0 | 0 |
| 0 | 100000 | 0 |
| 0 | 0 | -100000 |

where $M \in \mathbb{R}_+$ and $M$ large. Assume that a feasible solution exists. Let the optimal values of the variables be denoted by $\widehat{b}_j^+, \widehat{b}_j^-, \widehat{\lambda}_j^+, \widehat{\lambda}_j^-$, and let

$$\mathbb{E}' = \left\{ j \in \mathbb{E} | \widehat{b}_j^+ + \widehat{b}_j^- = 1 \right\}$$

(9c)–(9d) implies that
$\widehat{\lambda}_j^i > 0 \Leftrightarrow \widehat{b}_j^i = 1, i \in \{+, -\}, j \in \mathbb{E}$.
From (9b) follows that
$\left|\mathbb{E}'\right| \leq (s + m - 1)$. By (9e) $\left(\widehat{b}_j^+, \widehat{b}_j^-\right) \neq (1, 1)$ or $\widehat{\lambda}_j^+ \times \widehat{\lambda}_j^- = 0,, j \in \mathbb{E}$.

Hence, a feasible solution specifies at most $s + m - 1$ output input vectors $Z_j, j \in \mathbb{E}$' such that (9a) is true and thereby $\sum_{j \in \mathbb{E}'} \widehat{\lambda}_j Z_j = 0$, for $\widehat{\lambda}_j^+ - \widehat{\lambda}_j^- \equiv \widehat{\lambda}_j \neq 0, j \in \mathbb{E}$'. Hence, RC1 is violated. Conversely, RC1 is satisfied if the program has no feasible solution.

## Appendix B: How to Use *Qhull*

The Ifile in Table B.1 will generate all interior and exterior facets corresponding to Fig. 6.2a for the situation with two inputs and one output.

We execute *Qhull* by writing: *Qhull.exe s FF < Ifile > Ofile*, where the options *s* and *FF* ask for a summary and detailed information on the facets structure. *Qhull* writes the following messages to the screen:

```
Convex hull of 6 points in 3-d:
Number of vertices: 6
Number of facets: 8
Statistics for: | Qhull s FF
Number of points processed: 6
Number of hyperplanes created: 12
Number of distance tests for Qhull: 14
CPU seconds to compute hull (after input): 0
```

Hence, all 6 data points have been declared vertices in the convex hull [23]. The intersection form of the hull is the intersection of 8 halfspaces generated from 8 facets. Inspecting Fig. 6.2a, 7 of the 8 facets are visible. The facet not visible in Figure 2A is the facet spanned by the three "artificial" vectors: $(M, 0, 0)$ , $(0, M, 0)$ , $(0, 0, -M)$.

**The first section of the Ofile** consists of a listing of vertexes. The 6 vertices listed in the sequence used in Table B.1 are p0,...,p5,where p0 is "A", p1 is "B" and p2 is "C". *Qhull*'s internal vertex names are v0,...,v5.

```
Vertices and facets:
- p5 (v2): 0 0 -1e+05
- p3 (v1): 1e+05 0 0
- p4 (v0): 0 1e+05 0
- p0 (v3): 1.2e+02 90 1e+02 <-------- Point "A"
- p2 (v5): 10 10 10 <---------------------- Point "C"
- p1 (v6): 90 90 90 <---------------------- Point "B"
```

**The second section in the Ofile** consists of a detailed listing of the characteristics of each of the 8 facets. *Qhull*'s internal facet names are in this case f1,f2,f5,f7,f8,f9,f10,f11.

The format for each of the 8 subsections is as follows:

```
- name (here fi,i∈ {1,2,5,7,8,9,10,11})
- flags: top simplicial or bottom simplicial
- normal: n1 n2 n3
- offset: n4
- vertices: list of vertices
- neighboring facets: list of facets
```

The 7 visible facets $f_i, i \in \{2, 5, 7, 8, 9, 10, 11\}$ from Fig. 6.2a are indicated in the following Fig. B1:

To illustrate the information available from *Qhull* let us have a closer look at the subsections with information on the FDEF (interior facet) f11 and the exterior facets f9 and f10.

```
- f11
- flags: bottom simplicial
- normal: -0.2673 -0.5345 0.8018
- offset: 0
- vertices: p1 (v6) p2 (v5) p0 (v3)
- neighboring facets: f8 f9 f10
```

The output related to the FDEF f11 (interior facet) from *Qhull* provides us with the following information. The output component of the normal vector is strictly positive and the two input components are strictly negative. The intercept (the offset) is here zero and p0,p1 and p2 are located on the facet. The facet f11 has three neighboring facets: f8,f9 and f10.

[23] This need of course not to be the case. If the data points included in the Ifile include inefficient data points (points that will turn up as being located in the interior of the convex hull) then these points are declared non-vertex points and are ignored in the generation of the convex hull.

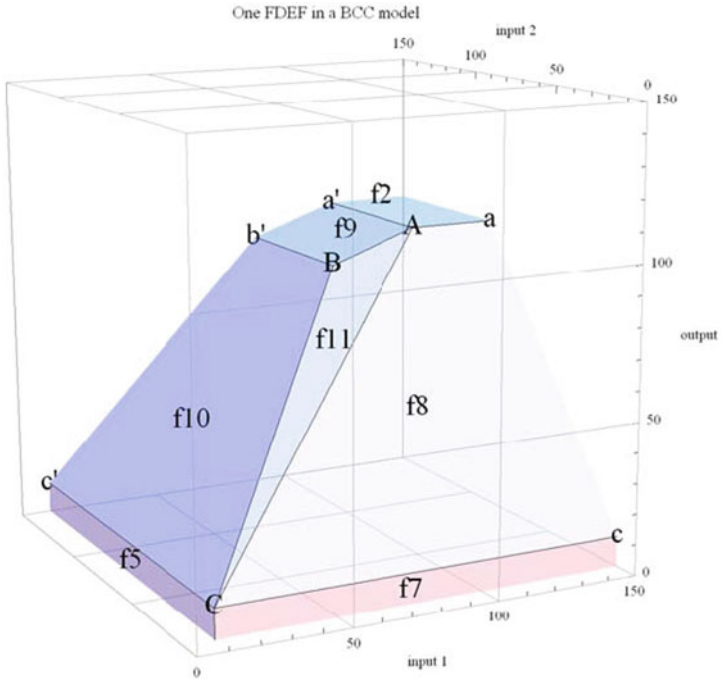



**Fig. B.1** The technology from Fig. 6.2a with indication of the seven visible facets generated from *Qhull*

```
- f9
- flags: bottom simplicial
- normal: -0.3162 0.0005697 0.9487
- offset: -56.97226
- vertices: p1 (v6) p0 (v3) p4 (v0)
- neighboring facets: f2 f10 f11
- f10
- flags: top simplicial
- normal: -0.7071 0 0.7071
- offset: 0
- vertices: p1 (v6) p2 (v5) p4 (v0)
- neighboring facets: f5 f9 f11
```

The output related to the exterior facet f9 and f10 from *Qhull* provides us with the following information. The output components of the normal vectors are both strictly positive but the second input components are zero or close to zero (the first input components are strictly negative). The intercept (the offset) is here zero for facet f10 and negative for facet f9. p1,p2 and p4 are located on the facet f10, where p4 is the artificial vector $(0, M, 0)$. The facet F10 has three neighboring facets: f5,f9 and f11.

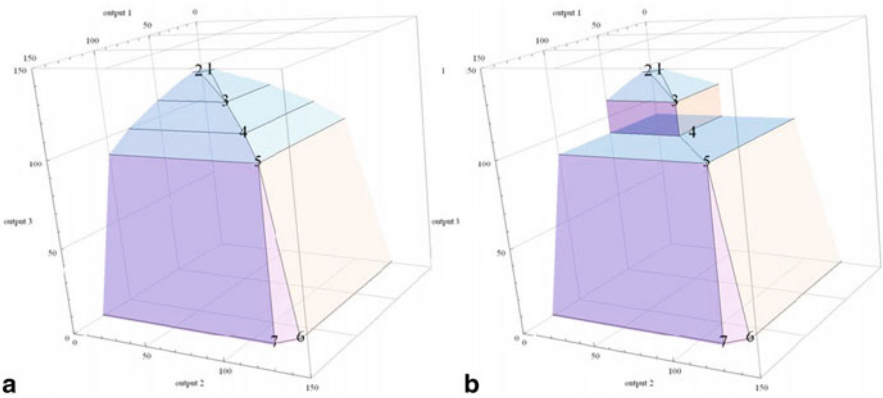



**Fig. C.1** Three Outputs and one input, Seven DMUs, Two FDEFs and Two NFDEFs. **a** The frontier 1-2-3-4-5-6-7 as the collection of four efficient facets. **b** The FDEF output possibility set as the union of two sets spanned from the two FDEFs.

*Qhull* is directly interfaced to or integrated with other programs like Matlab. There exists a nice small interface mPower[24], that allows *Qhull* to be executed from within Mathematica and all output is routed back to Mathematica as lists. This feature allows e.g. for easy graphical illustrations of facets structures in three dimensional space like the ones used in this chapter.

## Appendix C. FDEF and NFDEF in a CCR-Production Possibility Set

As an illustration of a more complex facet structure than the relatively simple one in Fig. 6.1 we have included an illustrative example with data given in Table C1 (seven CCR extreme efficient DMUs, three CCR-inefficient DMUs, one input and three outputs). The seven CCR extreme efficient DMUs span four efficient facets as illustrated in the three dimensional output space in Fig. C1a. The first column in Table C1 contains the index of each DMU, the second column is input consumed, the next three columns are output produced of each of the three types and the three last columns contain the efficiency scores from the three models EXFA, CCR and FDEF.

The last three DMUs in Table C1, i.e. DMU 8, 9, and 10, are CCR inefficient DMUs of the three last types mentioned in Table 6.1. The four efficient facets shown in Fig. C1a as $(1 \rightarrow 2 \rightarrow 3 \rightarrow 1)$, $(5 \rightarrow 6 \rightarrow 7 \rightarrow 5)$, $(3 \rightarrow 4)$, $(4 \rightarrow 5)$, are spanned by $\mathrm{DMU}_j$, $j \in F_i, i = 1, 2, 3, 4$, where $J_1 = \{1, 2, 3\}, J_2 = \{5, 6, 7\}, J_3 = \{3, 4\}, J_4 = \{4, 5\}$. The dimension of the first two facets is two; hence, the two

[24] See e.g. http://xlr8r.info/mPower/install.html.

**Table C.1** Three outputs, one input, ten DMUs, two FDEFs and two NFDEFs

| | Input | Output1 | Output2 | Output3 | $\theta^{EXFA}$ | $\theta^{CCR}$ | $\theta^{FDEF}$ |
|---|---|---|---|---|---|---|---|
| $DMU_1$ | 1.0 | 5 | 10 | 120 | 1.00 | 1.00 | 1.00 |
| $DMU_2$ | 1.0 | 10 | 5 | 120 | 1.00 | 1.00 | 1.00 |
| $DMU_3$ | 1.0 | 50 | 50 | 110 | 1.00 | 1.00 | 1.00 |
| $DMU_4$ | 1.0 | 80 | 80 | 100 | 0.976 | 1.00 | 1.11 |
| $DMU_5$ | 1.0 | 100 | 100 | 90 | 1.00 | 1.00 | 1.00 |
| $DMU_6$ | 1.0 | 115 | 125 | 1 | 1.00 | 1.00 | 1.00 |
| $DMU_7$ | 1.0 | 125 | 115 | 1 | 1.00 | 1.00 | 1.00 |
| $DMU_8$ | 1.0 | 10 | 15 | 55 | 0.476 | 0.480 | 0.480 |
| $DMU_9$ | 1.0 | 42 | 42 | 45 | 0.450 | 0.471 | 0.5000 |
| $DMU_{10}$ | 1.0 | 85 | 85 | 95 | 0.944 | 0.982 | 1.056 |

first facets are FDEFs. The dimension of the two last efficient facets is one; hence, the two last efficient facets are NFDEFs. The CCR extreme efficient DMU4 is the only CCR extreme efficient DMU not located on any FDEF. Let us consider the scores $\theta^{EXFA}$. Clearly, all DMUs located on FDEFs will get a score equal to one. The score for $DMU_4$ is below one because this CCR extreme efficient DMUs is not located on any FDEF. The inefficient $DMU_8$ is carefully placed such that it's $\theta_8^{CCR}$-projected output vector, i.e. $(\theta_8^{CCR}(10, 15, 55))$ is located on the facet spanned by $DMU_2$, $DMU_3$ and these two DMUs projection onto $Y1$-$Y3$ space. Hence, the $\theta_8^{CCR}$-projected input output vector is not on an FDEF and therefore we have $0.476 \approx \theta_8^{EXFA} < \theta_8^{CCR} \approx 0.480$. All three CCR inefficient DMUs share these features and hence $\theta_j^{EXFA} < \theta_j^{CCR}$, $j = 8, 9, 10$.

Next, let us consider the FDEF-technology spanned by these seven CCR extreme efficient DMUs. The geometric picture of this technology is exhibited in Fig. C1b. The first FDEF spanned by $DMU_j$, $j \in J_1$ generates the following subset of the output space (input is equal to one):

$$P_{FDEF}(1, J_1) = \left\{ y \in \mathbb{R}^3_+ | \left[1, y^T\right]^T \in T_{FDEF}(J_1) \right\}$$
$$= \left\{ y \in \mathbb{R}^3_+ | y \leq \lambda_1 Y_1 + \lambda_2 Y_2 + \lambda_3 Y_3, \sum_{j=1}^{3} \lambda_j = 1, \lambda_j \geq 0, j = 1, 2, 3 \right\}$$

Hence, every output vector in and below the triangle $(1 \rightarrow 2 \rightarrow 3 \rightarrow 1)$ belongs to this set. In a similar vein the second facet generates the following subset of the output space (again, input is equal to one):

$$P_{FDEF}(1, J_2) = \left\{ y \in \mathbb{R}^3_+ | \left[1, y^T\right]^T \in T_{FDEF}(J_2) \right\}$$
$$= \left\{ y \in \mathbb{R}^3_+ | y \leq \lambda_1 Y_5 + \lambda_2 Y_6 + \lambda_3 Y_7, \sum_{j=1}^{3} \lambda_j = 1, \lambda_j \geq 0, j = 1, 2, 3 \right\}$$

Hence, every output vectors in and below the triangle ($5 \rightarrow 6 \rightarrow 7 \rightarrow 5$) belongs to this set. Finally, since no more FDEFs are available the total output possibility set $P_{FDEF}(1)$ generated from the FDEF technology is the union of these two sets, i.e.

$$\begin{aligned} P_{FDEF}(1) &= \left\{ y \in \mathbb{R}^3_+ | \left[1, y^T\right]^T \in T_{FDEF} \right\} \\ &= P_{FDEF}(1, J_1) \cup P_{FDEF}(1, J_2) \end{aligned}$$

Let us consider the first CCR inefficient $DMU_8$. The output vector from $DMU_8$ belongs to the interior of $P_{FDEF}(1, J_1)$ and it's $\theta^{CCR}$-projection is onto a facet spanned by $DMU_2$, $DMU_3$ and these two DMUs projection onto $Y_1 - Y_3$ space. This facet is left intact after switching from the CCR-technology in Fig. C1a to the FDEF-technology in Fig. C1b. Hence, the change of model has no consequence for the size of the possible radial expansion of the output vector for $DMU_8$, and $0.480 \approx \theta_8^{CCR} = \theta_8^{FDEF}$.

The picture is different for the two other CCR inefficient DMUs. $DMU_9$ is carefully placed in output space such that its output vector belongs to the interior of $P_{FDEF}(1, J_2)$ and it's $\theta^{CCR}$ -projection is onto a facet spanned by $DMU_4$, $DMU_5$ and these two DMUs projection onto $Y_1 - Y_3$ space. Now, $DMU_4$ is not located on any FDEF. Hence, this facet is no longer valid after switching from the CCR-technology in Fig. C1a to the FDEF-technology in Figure C1b. Hence, the change of model implies an increase of the score: $0.471 \approx \theta_9^{CCR} < \theta_9^{FDEF} = 0.5$. The output vector from the last CCR inefficient $DMU_{10}$ is approximately two times the output vector from $DMU_9$. Hence, $DMU_{10}$'s $\theta^{CCR}$-projection is onto the same facet spanned by $DMU_4$, $DMU_5$ and these two DMUs projection onto $Y_1 - Y_3$ space. However, $DMU_{10}$ is carefully placed in output space such that its output vector does not belong to any of the sets $P_{FDEF}(1, J_i), i = 1, 2$. Hence, we get $0.982 \approx \theta_{10}^{CCR} < \theta_{10}^{FDEF} \approx 1.056$, with $\theta_{10}^{FDEF} > 1$.

## Appendix D. FDEF and NFDEF in a BCC-Production Possibility Set

As an illustration of a more complex facet structure than the relatively simple one in Fig. 6.2 we have included an illustrative example with data given in Table D1 (seven BCC extreme efficient DMUs, three BCC inefficient DMUs, two inputs and one output). The seven BCC extreme efficient DMUs span four efficient facets as illustrated in the three dimensional output space in Fig. D1. The data behind these efficient facets are exhibited in Table D1 where the first column contains the index of each DMU, the second and third columns are inputs consumed, the fourth column is the output produced. The last three columns contain the efficiency scores from the three models EXFA, CCR and FDEF.

The last three DMUs in Table D1, i.e. DMU 8, 9, and 10, are BCC inefficient DMUs of the three last types mentioned in relation to Table 6.1. The four efficient

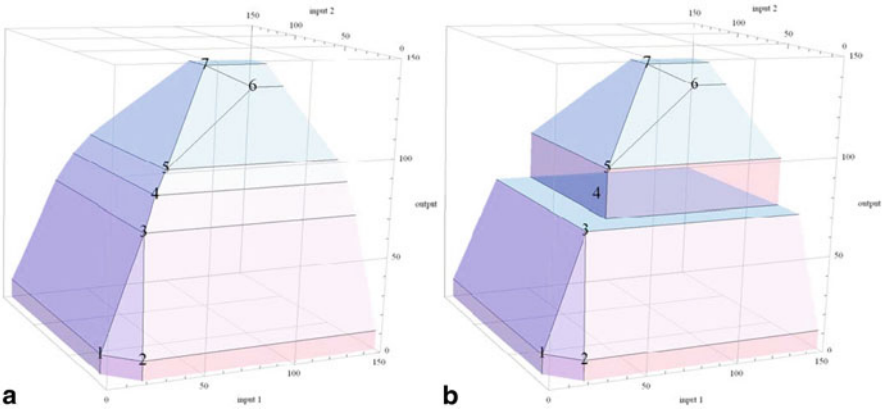



**Fig. D.1** One output, two inputs, seven DMUs, two FDEFs and two NFDEFs. **a** The frontier 1-2-3-4-5-6-7 as the collection of four efficient facets. **b** The FDEF production possibility set as the union of two sets spanned from the two FDEFs

facets shown in Fig. D1a as $(1 \rightarrow 2 \rightarrow 3 \rightarrow 1)$, $(5 \rightarrow 6 \rightarrow 7 \rightarrow 5)$, $(3 \rightarrow 4)$, $(4 \rightarrow 5)$, are spanned by $\text{DMU}_j$, $j \in F_i, i = 1, 2, 3, 4$, where $J_1 = \{1, 2, 3\}$, $J_2 = \{5, 6, 7\}$, $J_3 = \{3, 4\}$, $J_4 = \{4, 5\}$. The dimension of the first two facets is two and the dimension of the last two facets is one. Hence, the first two facets are FDEFs and the last two are NFDEFs. The BCC extreme efficient DMU4 is the only BCC extreme efficient DMU not located on any FDEF. Notice, that the regularity condition is satisfied since the intercept term for the facet $(1 \rightarrow 2 \rightarrow 3 \rightarrow 1)$ clearly is positive.

**Table D.1** One output, two inputs, ten DMUs, two FDEFs and two NFDEFs

| | Input 1 | Input 2 | Output | $\theta^{EXFA}$ | $\theta^{CCR}$ | $\theta^{FDEF}$ |
|---|---|---|---|---|---|---|
| $\text{DMU}_1$ | 5 | 20 | 10 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_2$ | 20 | 5 | 10 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_3$ | 30 | 30 | 65 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_4$ | 40 | 40 | 80 | 0.869 | 1.00 | 1.25 |
| $\text{DMU}_5$ | 50 | 50 | 90 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_6$ | 130 | 110 | 120 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_7$ | 110 | 130 | 130 | 1.00 | 1.00 | 1.00 |
| $\text{DMU}_8$ | 20 | 100 | 15 | 0.235 | 0.364 | 0.364 |
| $\text{DMU}_9$ | 100 | 100 | 80 | 0.348 | 0.400 | 0.500 |
| $\text{DMU}_{10}$ | 40 | 40 | 70 | 0.790 | 0.833 | 1.250 |

As an illustration of the violation of the regularity condition RC3 we can consider the data in Table D2 where we replace the data from $\text{DMU}_1$ and $\text{DMU}_2$ in Table D1 with $(10, 10, 5)$ and $(15, 15, 25)$. The Fig. D2 illustrates the production possibility set, and with only one FDEF spanned by $5 - 6 - 7$ we clearly see, that an extension

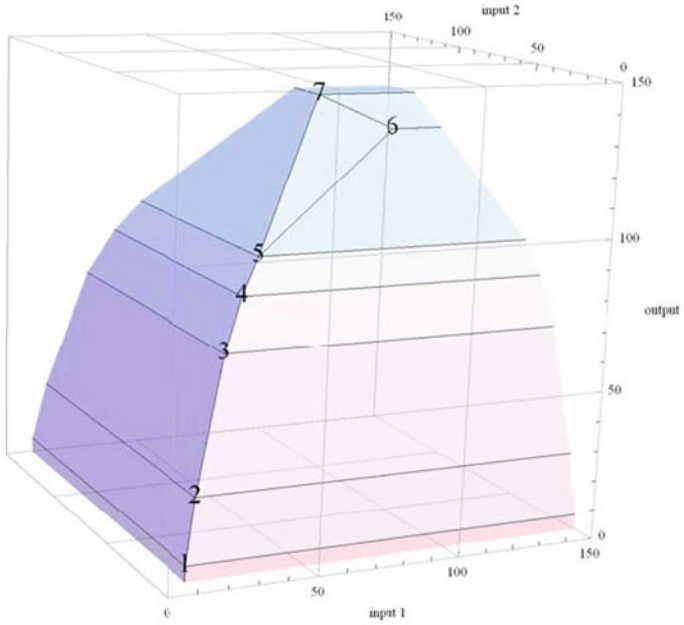



**Fig. D.2** The frontier 1-2-3-4-5-6-7 as the collection of five efficient facets. An example of a convex hull technology where the extended facet will provide a technology that violate "no free lunch'.

of this facet will violate the axiom of no free lunch. The scores estimated based on this technology is:

**Table D.2** One output, two inputs, ten DMUs, two FDEFs and two NFDEFs

| | $\theta^{EXFA}$ | $\theta^{CCR}$ | $\theta^{FDEF}$ |
|---|---|---|---|
| $DMU_1$ | $-12$ | 1.00 | 5 |
| $DMU_2$ | $-5.33$ | 1.00 | 3.333 |
| $DMU_3$ | 0.0 | 1.00 | 1.67 |
| $DMU_4$ | 0.75 | 1.00 | 1.25 |
| $DMU_5$ | 1.00 | 1.00 | 1.00 |
| $DMU_6$ | 1.00 | 1.00 | 1.00 |
| $DMU_7$ | 1.00 | 1.00 | 1.00 |
| $DMU_8$ | $-1$ | 0.625 | 2.5 |
| $DMU_9$ | 0.3 | 0.400 | 0.500 |
| $DMU_{10}$ | 0.25 | 0.833 | 1.250 |

Table D2 illustrates how the estimated possible radial contraction of the input vectors makes no sense, since the contraction to the extension of the facet $(5-6-7-5)$ involves contracted input vectors in the negative othant $\mathbb{R}^2_-$. Notice, that for $DMU_3$ we see a contraction of the input vector $(40, 40)$ to $(0, 0)$, implying

that $(y, x) = (80, (0, 0))$ is a feasible production plan, which clearly illustrates "a free lunch".

## References


1. Aparicio J, Pastor JT (2013) A well-defined efficiency measure for dealing with closest target in DEA. Appl Math Comput 219:9142–9154
2. Aparicio J, Ruiz AJL, Sirvent AI (2007) Closest targets and minimum distance to the Pareto efficient frontier in DEA. J Prod Anal 28:209–218
3. Appa G, Argyris N, Williams HP (2006) A methodology for cross-evaluation in DEA, working paper
4. Asmild M, Hougaard JL, Olesen OB (2013) Testing over-representation of observations in subsets of a DEA technology. Eur J Oper Res 230:88–96
5. Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 309:1078–1092
6. Banker RD, Conrad RF, Strauss RP (1986) A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production. Manage Sci 32(1):230–244
7. Banker RD, Charnes A, Cooper WW, Maindiratta A (1988) A comparison of DEA and translog estimates of production frontiers using simulated observations from a known technology. In: Dogramaci A, Färe R (eds) Applications of modern production theory. Kluwer, Boston, pp 33– 55
8. Bessent A, Bessent W, Elam J, Clark T (1988) Efficiency frontier determination by constrained facet analysis. Oper Res 36(5):785–795
9. Briec W, Leleu H (2003) Dual representations of non-parametric technologies and measurement of technical efficiency. J Prod Anal 20:71–96
10. Brockett PL, Roussau JJ, Wang Y, Zhow L (1997) Implementation of DEA models using GAMS. Research Report 765, University of Texas, Austin
11. Chambers RG, Mitchell T (2001) Homotheticity and non-radial changes. J Prod Anal 15:31–39
12. Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. J Econ Theory 70:407–419
13. Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions and Nerlovian efficiency. J Optim Theory Appl 98(2):351–364
14. Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Oper Res 2:429–444
15. Charnes A, Cooper WW, Rhodes E (1979) Short communication: measuring the efficiency of decision-making units. Eur J Oper Res 3:339
16. Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econom 30:91–107
17. Charnes A, Cooper WW, Huang ZM, Rousseau JJ (1989) Efficient facets and the rate of change: geometry and analysis of some Pareto-efficient empirical production possibility sets. CCS Research Report 622, Center for Cybernetic Studies, University of Texas at Austin
18. Charnes A, Cooper WW, Huang ZM, Sun DB (1990) Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. J Econom 46(1–2):73–91
19. Charnes A, Cooper WW, Thrall RM (1991) A structure for classifying and characterizing efficiencies and inefficiencies in data envelopment analysis. J Prod Anal 2:197–237
20. Cook WD, Seiford L (2009) Data envelopment analysis (DEA)—thirty years on. Eur J Oper Res 192:1–17
21. Cooper WW, Ruiz JL, Sirvent I (2007) Choosing weights from alternative optimal solutions of dual multiplier models in DEA. Eur J Oper Res 180:443–458
22. Cooper WW, Ruiz JL, Sirvent I (2011) Choices and uses of DEA weights. In: Cooper WW, Seiford LW, Zhu J (eds) Handbook of DEA, 2 edn, Chap. 4, p 109

23. Doyle J, Green R (1994a) Efficiency and cross-efficiency in DEA: derivations, meaning and uses. J Oper Res Soc 45(5):567–578
24. Doyle JR, Green RH (1994b) Strategic choice and data envelopment analysis: comparing computeres across many dimensions. J Inform Technol 9:61–69
25. Färe R, Lovell CAK (1978) Measuring the technical efficiency of production. J Econ Theory 19:150–162
26. Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer-Nijhoff Publisher, Boston
27. Fukuyama H, Sekitani K (2012) Decomposing the efficient frontier of the DEA production possibility set into a smallest number of convex polyhedrons by mixed integer programming. Eur J Oper Res 221:165–174
28. Førsund FR, Hjalmarsson L, Krivonozhko VE, Utkin OB (2007) Calculation of scale elsticities in DEA models: direct and indirect approaches. J Prod Anal 28(1/2):45–56
29. Green RH, Doyle JR, Cook WD (1996) Efficiency bounds in data envelopment analysis. Eur J Oper Res 89:482–490
30. Grünbaum B (1961) Measures of symmetry for convex sets, Proceedings of the seventh symposium in pure mathematics of the American Mathematical Society, Symposium on Convexity, pp 233–270
31. Klee VL (1953) Convex sets in linear spaces. Duke Math J 20:33–97
32. Koopmans TC (1957) Three essays on the state of economic science. Mc. Grawhill, New York
33. Krivonozhko V, Volodin AV, Sablin IA, Patrin M (2004) Construction of economic functions and calculating of marginal rates in DEA using parametric optimization methods. J Oper Res Soc 55(10):1049–1058
34. Lang P, Yolalan OR, Kettani O (1995) Controlled envelopment by face extensions in DEA. J Oper Res Soc 46(4):473–491
35. Lewin AY, Morey RC (1981) Measuring the relative efficiency and output potential of public sector organizations: an application of data envelopment analysis. Int J Policy Anal Inform Syst 5(4):267–285
36. Olesen OB, Petersen NC (1996) Indicators of ill-conditioned data sets and model misspecification in data envelopment analysis: an extended facet approach. Manage Sci 42:205–219
37. Olesen OB, Petersen NC (2003) Identification and use of efficient faces and facets in DEA. J Prod Anal 20:323–360
38. Portela MCAS, Thanassoulis E (2005) Profitability of a sample of Portuguese bank branches and its decomposition into technical and allocative components. Eur J Oper Res 162:850–866
39. Portela MCAS, Borges PC, Thanassoulis E (2003) Finding closest targets in non-oriented DEA models: the case of convex and non-convex technologies. J Prod Anal 19:251–269
40. Rockefellar RT (1970) Convex analysis. Princeton University Press, New Jersey
41. Schrijver A (1986) Theory of linear and integer programming. Wiley, New York
42. Steuer RE (1986) Multiple criteria optimization. Theory, computation and application. Wiley, New York
43. Takeda A, Nishimo H (2001) On measuring the inefficiency with the inner-product norm in data envelopment analysis. Eur J Oper Res 133:377–393
44. Thrall RM (1996) Duality, classification and slacks in DEA. Ann Oper Res 66:109–138.

# Chapter 7
# Stochastic Nonparametric Approach to Efficiency Analysis: A Unified Framework

**Timo Kuosmanen, Andrew Johnson and Antti Saastamoinen**

**Abstract** Bridging the gap between axiomatic *Data Envelopment Analysis* (DEA) and econometric *Stochastic Frontier Analysis* (SFA) has been one of the most vexing problems in the field of efficiency analysis. Recent developments in multivariate convex regression, particularly *Convex Nonparametric Least Squares* (CNLS) method, have led to the full integration of DEA and SFA into a unified framework of productivity analysis, referred to as *Stochastic Nonparametric Envelopment of Data* (StoNED). The unified framework of StoNED offers a general and flexible platform for efficiency analysis and related themes such as frontier estimation and production analysis, allowing one to combine existing tools of efficiency analysis in novel ways across the DEA-SFA spectrum, facilitating new opportunities for further methodological development. This chapter provides an updated and elaborated presentation of the CNLS and StoNED methods. This chapter also extends the scope of the StoNED method in several directions. Most notably, this chapter examines quantile estimation using StoNED and an extension of the StoNED method to the general case of multiple inputs and multiple outputs. This chapter also provides a detailed discussion of how to model heteroscedasticity in the inefficiency and noise terms.

**Keywords** Efficiency analysis · Frontier estimation · Multivariate convex regression · Nonparametric least squares · Productivity · Stochastic noise

T. Kuosmanen (✉) · A. Johnson · A. Saastamoinen
School of Business, Aalto University, 00100 Helsinki, Finland
e-mail: timo.kuosmanen@aalto.fi

A. Johnson
Department of Industrial and Systems Engineering, Texas A&M University, 77840, TX, Texas, USA

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_7

## 7.1 Introduction

Efficiency analysis is an essential and extensive research area that provides answers to such important questions as: Who are the best performing firms and can we learn something from their behavior?[1] What are the sources of efficiency differences across firms? Can efficiency be improved by government policy or better managerial practices? Are there benefits to increasing the scale of operations? These are examples of important questions we hope to resolve with efficiency analyses.

Efficiency analysis is an interdisciplinary field that spans such disciplines as economics, econometrics,[2] operations research and management science,[3] and engineering, among others. The methods of efficiency analysis are utilized in several fields of application including agriculture, banking, education, environment, health care, energy, manufacturing, transportation, and utilities, among many others. Efficiency analysis is performed at various different scales. Micro level applications range from individual persons, teams, production plants and facilities to company level and industry level efficiency assessments. Macro level applications range from comparative efficiency assessments of production systems or industries across countries to efficiency assessment of national economies. Indeed, efficiency improvement is one of the key components of productivity growth (e.g., Färe et al. 1994), which in turn is the primary driver of economic welfare. The benefits to understanding the relationship between efficiency and productivity and quantifying efficiency cannot be overstated. In words of Paul Krugman (1992, p. 9), "*Productivity isn't everything, but in the long run it is almost everything. A country's ability to improve its standard of living over time depends almost entirely on its ability to raise its output per worker*." Note that macro-level performance of a country is an aggregate of the individual firms operating within that country. Therefore, sound micro-foundations of efficiency analysis are critical for the integrity of productivity and efficiency analysis at macro level.

Unfortunately, there currently is no commonly accepted methodology of efficiency analysis, but the field is divided between two competing approaches: Data envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA).[4]

---

[1] We will henceforth use the term "firm" referring to any production unit that transforms inputs to output, including both non-profit and for-profit organizations. The firm can refer to an establishment (facility) or sub-division of a company or to an aggregate entity such as an industry, a region, or a country.

[2] Observe that 13 of the 100 most cited articles published in a leading field journal, the *Journal of Econometrics*, are efficiency analysis papers, including Simar and Wilson (2007) that has 436 citations, making it the #32 most cited paper in the journal in just 6 years from its publication (citations data gathered from Scopus, Nov 25, 2013).

[3] In operations research and management science, Charnes et al. (1978) ranks #1 as most cited article published in the *European Journal of Operational Research* (EJOR) and Banker et al. (1984) is the #1 most cited article in *Management Science*, two of the leading journals of this field (the flagship journals of EURO and INFORMS, respectively). In fact, Charnes et al. article has more than five times more citations than the 2nd most cited paper in EJOR (Nov 25, 2013).

[4] Citation statistics of some of the key papers provide undisputable evidence about the significant influence of this field. The four most cited papers are Charnes et al. (1978) with 6152 citations,

*Data envelopment analysis* (DEA, Farrell 1957; Charnes et al. 1978) is an axiomatic, mathematical programming approach to efficiency analysis. DEA's main advantage compared to econometric, regression-based tools is its nonparametric treatment of the frontier, building upon axioms of production theory such as free disposability (monotonicity), convexity (concavity), and constant returns to scale (homogeneity). DEA does not assume any particular functional form for the frontier or the distribution of inefficiency. It's direct, data-driven approach is helpful for communicating the results of efficiency analysis to decision-makers. However, the main shortcoming of DEA is that it attributes all deviations from the frontier to inefficiency. This is often a heroic assumption.

*Stochastic frontier analysis* (SFA, Aigner et al. 1977; Meeusen and van den Broeck 1977) is often, incorrectly, viewed as a direct competitor of DEA. The key strength of SFA is its probabilistic modeling of deviations from the frontier, which are decomposed into a non-negative inefficiency term and an idiosyncratic error term that accounts for omitted factors such as unobserved heterogeneity of firms and their operating environments, random errors of measurement and data processing, specification errors, and other sources of noise. In contrast to DEA, SFA utilizes parametric regression techniques, which require *ex ante* specifications of the functional forms of the frontier and the inefficiency distribution. Since the economic theory rarely justifies a particular functional form, flexible functional forms such as translog are frequently used. However flexible functional forms often violate axioms of production theory, whereas imposing the axioms will reduce flexibility. In summary, the DEA and SFA methods are not direct competitors but rather complements: in the tradeoff between DEA and SFA something is sacrificed for something to be gained. Namely DEA does not model noise, but is able to impose axiomatic properties and estimate the frontier non-parametrically, while SFA cannot impose axiomatic properties, but has the benefit of modeling inefficiency and noise.

Bridging the gap between axiomatic DEA and stochastic SFA was for a long time one of the most vexing problems in the field of efficiency analysis. The recent works on convex nonparametric least squares (CNLS) by Kuosmanen (2008), Kuosmanen and Johnson (2010), and Kuosmanen and Kortelainen (2012) have led to the full integration of DEA and SFA into a unified framework of productivity analysis, which we refer to as *stochastic nonparametric envelopment of data* (StoNED).[5]

We see the development of StoNED as a paradigm shift for efficiency analysis. It is no longer necessary to decide if modeling noise is more important than imposing axioms of production theory: we can do both using StoNED. The unified framework of StoNED offers deeper insights to the foundations of DEA and SFA, but it also provides a more general and flexible platform for efficiency analysis and related

Banker et al. (1984) with 3415 citations, Farrell (1957) with 3296 citations, and Aigner et al. (1977) with 1875 citations (Scopus, Nov 25, 2013).

[5] The term StoNED was coined by Kuosmanen (2006). By request of referees, Kuosmanen and Kortelainen (2012) used the term stochastic "non-smooth" envelopment, as their model specification involves parametric distributional assumptions. In this chapter we show that the distributional assumptions can be relaxed: see Sect. 7.5.2.3 and 7.6.2.

themes such as frontier estimation and production analysis. Further, a number of extensions to the original DEA and SFA methods have been developed over the past decades. The unified StoNED framework allows us to combine the existing tools of efficiency analysis in novel ways across the DEA-SFA spectrum, facilitating new opportunities for further methodological development.

The main objective of this chapter is to provide an updated and elaborated presentation of the CNLS and StoNED methods, the most promising new tools for axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise. Our secondary objective is to extend the scope of the StoNED method in several dimensions. This chapter provides the first extension of the StoNED method to the general case of multiple inputs and multiple outputs. We also consider quantile estimation using StoNED, and present a detailed discussion of how to model heteroscedasticity in the inefficiency and noise terms.

The rest of this chapter is organized as follows. Section 7.2 introduces the unified StoNED framework and its special cases by reviewing alternative sets of assumptions that motivate different estimation methods applied in productivity analysis. Our focus is explicitly on the axiomatic DEA-style approaches. Section 7.3 presents the CNLS regression as a quadratic programming problem. Section 7.4 discusses the intimate connections between CNLS and DEA, and introduces a step-wise $C^2$NLS estimator. Section 7.5 further develops the step-wise estimation approach for the StoNED estimator. Section 7.6 reviews some important extensions to the StoNED, including the multiplicative formulation (Sect. 7.6.1), observations from multiple time periods that make up a panel data (Sect. 7.6.2), directional distance functions (DDF) for modeling multiple output variables (Sect. 7.6.3), and quantile regression formulation (Sect. 7.6.4). The model of contextual variables that represent operational conditions or practices is examined in detail in Sect. 7.7. Testing of heteroscedasticity and modeling heteroscedasticity of inefficiency and noise using a doubly-heteroscedastic model discussed in Sect. 7.8. Finally, Sect. 7.9 concludes with discussion of some promising avenues of future research.

## 7.2 Unified Frontier Model

To maintain direct contact with the SFA literature, we introduce the unified model of frontier production function in the multiple input, single output case. Multiple outputs can be modeled using cost functions (see Kortelainen and Kuosmanen 2012, Sect. 7.4.4; and Kuosmanen 2012) and distance functions. A general multi-input multi-output directional distance function model will be introduced in Sect. 7.6.3.

Production technology is represented by a frontier *production function* $f(\mathbf{x})$, where $\mathbf{x}$ is a $m$-dimensional input vector.[6] Frontier $f(\mathbf{x})$ indicates the maximum output that

[6] For clarity, we denote vectors by bold lower case letters (e.g., $\mathbf{x}$) and matrices by bold capital letters (e.g., $\mathbf{Z}$). All vectors are column vectors, unless otherwise indicated. Note: $\mathbf{x}'$ denotes the transpose of vector $\mathbf{x}$.

can be produced with inputs $\mathbf{x}$, and hence the function $f(\mathbf{x})$ characterizes the boundary of the production possibility set. We assume that function $f$ belongs to the class of continuous, monotonic increasing, and globally concave functions that can be non-differentiable (we denote this class as $F_2$). This is equivalent to stating that the production possibility set satisfies the classic DEA assumptions of free disposability and convexity. In contrast to SFA, no specific functional form for $f$ is assumed.

The observed output $y_i$ of firm $i$ may differ from $f(\mathbf{x}_i)$ due to inefficiency and noise. We follow the SFA literature and introduce a composite error term $\varepsilon_i = v_i - u_i$, which consists of the inefficiency term $u_i > 0$ and the stochastic noise term $v_i$, formally,

$$\begin{aligned} y_i &= f(\mathbf{x}_i) + \varepsilon_i \\ &= f(\mathbf{x}_i) - u_i + v_i,\ i = 1, \ldots, n \end{aligned} \tag{7.1}$$

Variables $u_i$ and $v_i$ ($i = 1, \ldots, n$) are random variables that are assumed to be statistically independent of each other as well as of inputs $\mathbf{x}_i$. We assume that the inefficiency term has a positive mean and a constant finite variance, that is, $E(u_i) = \mu > 0$ and $Var(u_i) = \sigma_u^2 < \infty$. We further assume zero mean noise with a constant finite variance, that is, $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2 < \infty$. Assuming $\sigma_u^2$ and $\sigma_v^2$ are constant across firms is referred to as homoscedasticity; models with heteroskedastic inefficiency and noise will be discussed in Sect. 7.8. For the sake of generality and to maintain the fully nonparametric orientation, we do not introduce any distributional assumptions for $u_i$ or $v_i$ at this point. However, some estimation techniques to be introduced below require additional parametric assumptions.

In model (7.1), the deterministic part (i.e., production function $f$) is defined analogous to the DEA literature, while the stochastic part (i.e., composite error term $\varepsilon_i$) is defined similar to SFA. As a result, model (7.1) encompasses the classic models of the SFA and DEA literature as its constrained special cases. Note that in this chapter we use the term "model" in the sense of the econometric literature to refer to the description of the data generating process (DGP). DEA and SFA are alternative estimators or methods for estimating the production function $f$, the expected inefficiency $\mu$, and the firm-specific realizations of the random inefficiency term $u_i$. We note that in the DEA literature it is common to use the term "model" for the linear programming problem (e.g., LP model) or other mathematical programming formulations for computing the estimator. To avoid confusion, we will follow the econometric terminology and refer to Eq. (7.1) and the related assumptions as the model, whereas DEA, SFA, CNLS, and StoNED are referred to as estimators. In this terminology, "DEA model" or "SFA model" refer to the specific assumptions regarding the variables of model (7.1).

The literature of efficiency analysis has conventionally focused on fully parametric or nonparametric versions of model (7.1). Parametric models postulate a priori a specific functional form for $f$ (e.g., Cobb-Douglas, translog, etc.) and subsequently estimate its unknown parameters. In contrast, axiomatic nonparametric models assume that $f$ satisfies certain regularity axioms (e.g., monotonicity and concavity), but no particular functional form is assumed. At this point, we must emphasize that the

term nonparametric does not necessarily imply that there are no restrictive assumptions. It is not true that the assumptions of a nonparametric model are necessarily less restrictive than those of a parametric model. For example, the fully nonparametric DEA estimator of model (7.1) is based on the assumption of no noise (i.e., $v_i = 0$ for all firms $i$). Assuming away noise does not require any specific parametric specification, but it is nevertheless a restrictive assumption. In fact, it is less restrictive to impose parametric structure and assume $v_i$ are identically and independently distributed according to the normal distribution $N(0, \sigma_v^2)$. Note that this parametric specification contains the fully nonparametric "deterministic" case of no noise as its restricted special case, obtained by imposing the parameter restriction $\sigma_v^2 = 0$.

In addition to the pure parametric and nonparametric alternatives, the intermediate cases of semiparametric and semi-nonparametric models have become increasingly popular in recent years. However, the exact meaning of this terminology is often confused. Chen (2007) provides an intuitive and useful definition that we find worth quoting:

> An econometric model is termed "*parametric*" if all of its parameters are in finite dimensional parameter spaces; a model is "*nonparametric*" if all of its parameters are in infinite-dimensional parameter spaces; a model is "*semiparametric*" if its parameters of interests are in finite-dimensional spaces but its nuisance parameters are in infinite-dimensional spaces; a model is "*semi-nonparametric*" if it contains both finite-dimensional and infinite-dimensional unknown parameters of interests. Chen (2007), p. 5552, footnote 1.

Note that according to the above definition both the semiparametric and semi-nonparametric model contain a nonparametric part and a parametric part. The distinction between the terms semiparametric and semi-nonparametric is subjective, dependent on whether we are interested in the empirical estimates of the nonparametric part or not. The same model can be either semiparametric, if our main interest is in the parameter estimates of the parametric part and the nonparametric part is of no particular interest, or semi-nonparametric, if we are interested in the results of the nonparametric part.

Model (7.1) can be interpreted as a neoclassical or frontier model depending on the interpretation of the disturbance term (cf., Kuosmanen and Fosgerau 2009). The neoclassical model assumes that all firms are efficient and disturbances are random, uncorrelated noise terms. Frontier models typically assume that all or some part of the deviations from the frontier are attributed to systematic inefficiency.

Table 7.1 combines the criteria described above to identify six alternative estimation methods commonly used for estimating the variants of the unified model (7.1), together with some canonical references. On the parametric side, OLS refers to *ordinary least squares*, PP means *parametric programming*, COLS is *corrected ordinary least squares*, and SFA is *stochastic frontier analysis* (see, e.g., Kumbhakar and Lovell 2000, for an introduction to the parametric approach to efficiency analysis). The focus of this chapter is on the axiomatic nonparametric and semi-nonparametric variants of model (7.1): CNLS refers to *convex nonparametric least squares* (Sect. 7.3), DEA is *data envelopment analysis* (Sect. 7.4.1), $C^2$NLS is *corrected convex non-parametric least squares* (Sect. 7.4.2), and StoNED is *stochastic nonparametric envelopment of data* (Sect. 7.5).

**Table 7.1** Classification of methods

| | | Parametric | Nonparametric |
|---|---|---|---|
| **Central tendency** | | *OLS*<br>Cobb and Douglas (1928) | *CNLS* (Sect. 7.3)<br>Hildreth (1954)<br>Hanson and Pledger (1976) |
| **Deterministic frontier** | **Sign constraints** | *PP*<br>Aigner and Chu (1968)<br>Timmer (1971) | *DEA* (Sect. 7.4.1)<br>Farrell (1957)<br>Charnes et al. (1978) |
| | **2-step estimation** | *COLS*<br>Winsten (1957)<br>Greene (1980) | $C^2$*NLS* (Sect. 7.4.2)<br>Kuosmanen and Johnson (2010) |
| **Stochastic frontier** | | *SFA*<br>Aigner et al. (1977)<br>Meeusen and van den Broeck (1977) | *StoNED* (Sect. 7.5)<br>Kuosmanen and Kortelainen (2012) |

## 7.3 Convex Nonparametric Least Squares

In this section we consider the special case of model (7.1) where the composite error term $\varepsilon$ consists exclusively of noise $v$, and there is no inefficiency (i.e., we assume $u = 0$). This special case is relevant for modeling firms that operate in the competitive market environment, which meets (at least by approximation) the conditions of perfect competition considered in microeconomic theory. We will relax this no inefficiency assumption from Sect. 7.4 onwards, but the insights gained in this section will be critical for understanding the developments in the following sections.

In the case of a symmetric zero-mean error term that satisfies $E(\varepsilon_i) = 0$ for all $i$, the expected value of output conditional on inputs equals the value of the production function, that is,

$$E(y_i \,|\mathbf{x}_i) = E(f(\mathbf{x}_i)) + E(\varepsilon_i) = f(\mathbf{x}_i).$$

Therefore, in this setting the production function $f$ can be estimated by nonparametric regression techniques. Note that the term "regression" refers to the conditional mean $E(y_i \,|\mathbf{x}_i)$.

Hildreth (1954) was the first to consider nonparametric regression subject to monotonicity and concavity constraints in the case of a single input variable $x$ (see also Hanson and Pledger 1976). Kuosmanen (2008) extended Hidreth's approach to the multivariate setting with a vector-valued $\mathbf{x}$, and coined the term *convex nonparametric least squares* (CNLS) for this method. CNLS builds upon the assumption that the true but unknown production function $f$ belongs to the set of continuous, monotonic increasing and globally concave functions, $F_2$, imposing exactly the same production axioms as standard DEA.

The CNLS estimator of function $f$ is obtained as the optimal solution to the infinite dimensional least squares problem

$$\min_{f} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$$

subject to

$$f \in F_2 \tag{7.2}$$

The functional form of $f$ is not specified beforehand. Rather, the optimal solution will identify the best-fit function $f$ from the family $F_2$. Note that set $F_2$ includes an infinite number of functions, which makes problem (7.2) impossible to solve through brute force trial and error. Further, problem (7.2) does not generally have a unique solution for any arbitrary input vector $\mathbf{x}$, but a unique solution exists for estimating $f$ for the observed data points $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. Therefore, we will next discuss the estimation of $f$ for the observed data points and extrapolation to unobserved points in sub-section 7.3.2.

### 7.3.1 CNLS Estimator for the Observed Data Points

A unique solution to problem (7.2) for the observed data points $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, can be found by solving the following finite dimensional *quadratic programming* (QP) problem

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2$$

subject to

$$y_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i + \varepsilon_i^{CNLS} \ \forall i$$

$$\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \ \forall h, i$$

$$\boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i \tag{7.3}$$

where $\alpha_i$ and $\boldsymbol{\beta}_i$ define the intercept and slope parameters of tangent hyperplanes that characterize the estimated piece-wise linear frontier (note that $\boldsymbol{\beta}'_i \mathbf{x}_i = \beta_{i1} x_{i1} + \beta_{i2} x_{i2} + \ldots + \beta_{im} x_{im}$). Symbol $\varepsilon_i^{CNLS}$ denotes the CNLS residual, which is an estimator of the true but unobserved $\varepsilon_i = v_i$. Note that in (7.3) the Greek letters are variables and the Latin letters are parameters (i.e., $(\mathbf{x}_i, y_i)$ are observed data).

Kuosmanen (2008) introduced the QP formulation (7.3), and proved its equivalence with the infinite dimensional optimization problem (7.2). Specifically, if we denote the value of the objective function in the optimal solution to the infinite dimensional CNLS formulation (7.2) by $SSE_{CNLS}$ (SSE = the sum of squares of errors), and that of the finite QP problem (7.3) by $SSE_{QP}$, then the equivalence can be stated as follows.

**Theorem 1** $SSE_{CNLS} = SSE_{QP}$.

Proof. See Kuosmanen (2008), Theorem 2.1.

The equivalence result does not restrict to the objective functions, the optimal solution to problem (7.3) also provides us unique estimates of function $f$ for the observed data points. Once the optimal solution is found, we will add "hats" on top of $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$, and refer to them as estimators.[7] In other words, $\alpha_i$, $\boldsymbol{\beta}_i$, and $\varepsilon_i^{CNLS}$ are variables of problem (7.3), whereas estimators $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$, and $\hat{\varepsilon}_i^{CNLS}$ provide the optimal solution to problem (7.3). Given $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ from (7.3), we define

$$\hat{f}^{CNLS}(\mathbf{x}_i) = \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \mathbf{x}_i = y_i - \hat{\varepsilon}_i^{CNLS}. \tag{7.4}$$

This estimator of function $f$ satisfies the following properties:

**Theorem 2** In the case of the neoclassical model with no inefficiency, $\hat{f}^{CNLS}(\mathbf{x}_i)$ is a unique, unbiased and consistent estimator of $f(\mathbf{x}_i)$ for the observed data points $(\mathbf{x}_i, y_i), i = 1, \ldots, n$.

Proof. Uniqueness is proved by Lim and Glynn (2012), Proposition 1. Unbiasedness follows from Seijo and Sen (2011), Lemma 2.4. Consistency is proved under slightly different assumptions in Seijo and Sen (2011), Theorems 3.1 and 3.2, and Lim and Glynn (2012), Theorems 1 and 2.

The constraints of the QP problem (7.3) have the following compelling interpretations.[8] The first constraint of the least squares formulation (7.3) is a linear regression equation. However, the CNLS regression does not assume linear $f$. note that coefficients $\alpha_i$ and $\boldsymbol{\beta}_i$ are specific to each observation $i$. Using the terminology of DEA, $\alpha_i$ and $\boldsymbol{\beta}_i$ are directly analogous to the multiplier coefficients of the dual formulation of DEA. The inequality constraints in (7.3) can be interpreted as a system of *Afriat inequalities* (compare with Afriat 1967, 1972; and Varian 1984). As Kuosmanen (2008) emphasizes, the Afriat inequalities are the key to modeling the concavity axiom in the general multiple regression setting.

Coefficients $\alpha_i$ and $\boldsymbol{\beta}_i$ should not be misinterpreted as parameters of the estimated function $f$, but rather, as parameters characterizing tangent hyperplanes to an unknown production function $f$. These coefficients characterize a convex piece-wise linear function, to be examined in more detail the next sub-section. At this point, we must emphasize that we did not assume or restrict the domain $F_2$ to only include piece-wise linear function. In fact, it turns out that the "optimal" functional form to solving the infinite dimensional least squares problem (7.2) is always a convex piece-wise linear function characterized by coefficients $\alpha_i$ and $\boldsymbol{\beta}_i$. However, this optimal solution is unique only for the observed data points.

[7] In application, when estimators are calculated for a specific data set we will refer to these as estimated parameters.

[8] Note this formulation is written for ease of interpretation. Other formulations might be preferred to improve computational performance.

### 7.3.2 Extrapolating to Unobserved Points

In many applications we are interested in estimating the frontier not only for the observed data points, but also for unobserved input vectors $\mathbf{x}$. Although the CNLS estimator is unique for the observed data points, there is no unique way of extrapolating the CNLS estimator to unobserved points. In general, the optimal solution to the infinite dimensional least squares problem (7.2) is not unique, but there exists a set of functions $f^* \in F_2^*$ that solve the optimization problem (7.2). Formally, we denote the set of alternate optima to (7.2) as

$$F_2^* = \left\{ f^* \,\middle|\, f^* = \arg\min_{f \in F_2} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \right\}.$$

Kuosmanen (2008) characterizes the minimum and maximum bounds for the functions $f^* \in F_2^*$. It turns out that both bounds are piece-wise linear functions. However, only the minimum bound satisfies the postulated monotonicity and concavity properties. To resolve the non-uniqueness issue, Kuosmanen and Kortelainen (2012) appeal to the *minimum extrapolation principle* and propose to use the lower bound

$$\hat{f}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}' x \,\middle|\, \alpha + \boldsymbol{\beta}' \mathbf{x}_i \geq \hat{f}^{CNLS}(\mathbf{x}_i) \; \forall i = 1, \ldots, n \right\} \tag{7.5}$$

Note that the lower bound $\hat{f}_{\min}^{CNLS}$ is simply the DEA estimator (single output, variable returns to scale) applied to the observed inputs $\mathbf{x}_i$ and the fitted outputs $\hat{f}^{CNLS}(\mathbf{x}_i)$ obtained from Eq. (7.4).[9] The lower bound function satisfies the postulated properties of monotonicity and concavity. We can make the following connection between the lower bound (7.5) and the infinite dimensional CNLS problem (7.2).

**Theorem 3** Function $\hat{f}_{\min}^{CNLS}$ stated in Eq. (7.5) is one of the optimal solutions to the infinite dimensional optimization problem (7.2). It is the unique lower bound for the functions that solve problem (7.2), formally

$$\hat{f}_{\min}^{CNLS}(\mathbf{x}) \leq f^*(\mathbf{x}) \text{ for all } \mathbf{x} \in \Re_+^m \text{ and } f^* \in F_2^*.$$

Proof. See Kuosmanen (2008) Theorem 4.1.

Note that while $\hat{f}^{CNLS}$ is unbiased and consistent for the observed points $\mathbf{x}_i$ (Theorem 3), the use of the piece-wise linear minimum function $\hat{f}_{\min}^{CNLS}$ will cause downward bias in finite samples as we apply the minimum extrapolation principle to extrapolate to unobserved points $\mathbf{x}$. Within the observed range of data, the downward bias will diminish as the sample size increases.

It is also worth noting that the optimal solution to the QP problem (7.3) does not necessarily produce unique coefficients $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$. Although $\hat{f}_{\min}^{CNLS}$ is a unique

[9] In addition to the use of DEA to identify the lower bound function, there is a more fundamental connection between CNLS and DEA, to be explored in Sect. 7.4.

lower bound, consistent with the minimum extrapolation principle, the coefficients $\hat{\alpha}_i$ and $\boldsymbol{\beta}_i$ obtained as the optimal solution to (7.5) need not be unique either. It is well-known in the DEA literature that these multiplier coefficients are not unique in the vertices of the piece-wise linear function.

### 7.3.3 Computational Issues

The CNLS problem (7.3) has linear constraints and a quadratic objective function, hence it can be solved by QCP solvers such as CPLEX or MOSEK.[10] Standard solvers work well in relatively small sample sizes (50–200 firms) available in the majority of published applications of efficiency analysis. However, since the number of Afriat inequalities in (7.3) grows at a quadratic rate as a function of the number of observations, the computational burden becomes a significant issue when the sample size increases beyond 300 firms. Note that adding a new firm to the sample increases the number of unknown parameters by $m+2$, and the number of Afriat inequality constraints increases by $2n$. Introducing an additional input variable increases the number of unknown parameters by $n$, but there is no impact on the number of constraints. For these reasons, standard QP algorithms are inadequate for handling large samples with several hundreds or thousands of observations.

As a first step towards improving computational performance in small samples and to allow for larger problems to be solved, Lee et al. (2013) propose to follow the strategy of Dantzig et al. (1954, 1959) to iteratively identify and add violated constraints. The algorithm developed by Lee et al. first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. In computational experiments, this algorithm allowed problems with up to 1000 firms to be solved. Therefore, this algorithm has practical value especially in large sample applications and simulation-based methods such as bootstrapping or Monte Carlo studies. Another recent study by Hannah and Dunson (2013) implements CNLS in Matlab, reporting promising results. However, further algorithm development is needed to make the CNLS problem computable in very large sample sizes containing several thousands or millions of observations.

## 7.4 Deterministic Frontiers

In this section we consider another special case of model (7.1) where the composite error term $\varepsilon$ consists exclusively of inefficiency $u$, and there is no noise (i.e., $v=0$). In the SFA literature, this special case is commonly referred to as the *deterministic model*. This does not imply, however, that probabilistic inferences are impossible.

[10] Examples of computational codes for GAMS are available on the StoNED website: www.nomepre.net/stoned/.

Banker (1993) was the first to show that DEA can be understood as a maximum likelihood estimator of the deterministic model, with a statistical (probabilistic) foundation. However, the known statistical properties and inferences in the DEA literature restrict to the finite sample error that generally diminishes as the sample size increases. Or stated differently, the model specification and input and output data in the deterministic model are assumed to be exact and correct, so the only probabilistic component is the random sample of observations drawn from the production possibility set. This same deterministic model and its associated statistical foundation are used for inference in the bootstrapping methods (e.g., Simar and Wilson 1998, 2000). Thus, statistical inference and confidence intervals estimated using bootstrapping methods only account for uncertainty in sampling and do not account for other sources of random variation or noise. Thus, bootstrap confidence intervals of DEA are not directly comparable to confidence intervals of other models that are genuinely stochastic in their nature (e.g., the SFA confidence intervals).

It is important to recognize that if the no noise assumption ($v = 0$) of the deterministic model does not hold, the statistical foundations of DEA collapse. The bootstrapping methods to adjust for the small sample are not a remedy against noise, rather adjusting for the sampling bias can make the DEA estimator worse if data are perturbed by noise. The stochastic case that includes both inefficiency and noise simultaneously will be considered in Sect. 7.5. The purpose of this section is to establish some useful connections between the 'neoclassical' CNLS and the 'deterministic' DEA to develop a unified framework and pave the way for a stochastic nonparametric StoNED estimator.

### 7.4.1 DEA as Sign-Constrained CNLS

In the single-output case, the variable returns to scale (VRS) DEA estimator of production function $f$ can be stated as

$$\begin{aligned}\hat{f}^{DEA}(\mathbf{x}) &= \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}'\mathrm{x} \,\middle|\, \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq y_i \;\forall i = 1,\ldots,n \right\} \\ &= \max_{\lambda} \left\{ \sum_{h=1}^{n} \lambda_h y_h \,\middle|\, \mathbf{x} \geq \sum_{h=1}^{n} \lambda_h \mathbf{x}_h \,;\, \sum_{h=1}^{n} \lambda_h = 1 \right\} \end{aligned} \tag{7.6}$$

Note the difference between formulations (7.5) and (7.6): the former one uses the estimated output values $\hat{f}^{CNLS}(\mathbf{x}_i)$, whereas in the latter one uses the observed outputs $y_i$. Otherwise the formulations (7.5) and (7.6) are equivalent. The minimization formulation in (7.6) can be interpreted as the DEA multiplier formulation, whereas the maximization formulation of (7.6) is known as the DEA envelopment formulation. The duality theory of linear programming implies that the two formulations are equivalent.

Consider next a version of the CNLS estimator with an additional sign constraint on the residuals

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS-})^2$$

subject to

$$y_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i + \varepsilon_i^{CNLS-} \ \forall i$$
$$\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \ \forall h, i$$
$$\boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i$$
$$\varepsilon_i^{CNLS-} \leq 0 \ \forall i \tag{7.7}$$

Comparing (7.3) and (7.6), we see that the only difference is the last constraint of (7.7), which is not present in the original CNLS formulation. Due to the sign constraint, Kuosmanen and Johnson (2010) interpret (7.6) as an axiomatic, nonparametric counterpart to the classic parametric programming approach of Aigner and Chu (1968).

We now establish the formal connection between CNLS and DEA as follows. Let $\hat{f}_{\min}^{CNLS-}(\mathbf{x})$ denote the piece-wise linear function obtained by applying Eq. (7.5) to the observed inputs $\mathbf{x}_i$ and the fitted values $\hat{y}_i$ of the sign-constrained formulation (7.7).

**Theorem 4** The sign-constrained CNLS estimator is equivalent to the DEA VRS estimator:

$$\hat{f}_{\min}^{CNLS-}(\mathbf{x}) = \hat{f}^{DEA}(\mathbf{x})$$

Proof. Follows directly from Theorem 3.1 in Kuosmanen and Johnson (2010).

Although Theorem 4 was stated in the VRS case, the equivalence of DEA and sign-constrained CNLS does not restrict to the VRS case. Indeed parallel results are available for the other standard specifications of returns to scale by imposing additional constraints on the coefficients $\alpha_i$ in formulations (7.3) or (7.7) as follows:

Constant returns to scale (CRS): impose $\alpha_i = 0 \ \forall i$
Non-increasing returns to scale (NIRS): impose $\alpha_i \geq 0 \ \forall i$
Non-decreasing returns to scale (NDRS): impose $\alpha_i \leq 0 \ \forall i$

Similarly, if the convexity assumption of DEA is relaxed the free disposable hull (FDH), Afriat (1972), estimator provides the minimum envelopment of data subject to free disposability. Keshvari and Kuosmanen (2013) show that the FDH formulation is a sign-constrained special case of isotonic nonparametric least squares (INLS), which in turn is the concavity relaxed version of CNLS.

From a practical point of view, the least squares interpretation of DEA opens up new avenues for applying tools from econometrics to DEA. For example, Kuosmanen and Johnson (2010) propose to measure the goodness-of-fit of DEA estimator by

using the standard *coefficient of determination* from regression analysis, specifically

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}. \tag{7.8}$$

Where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average output in the sample. The $R^2$ statistic measures the proportion of output variation that is explained by the DEA frontier. While this variance decomposition can be applied to any regression model (including DEA), we note that DEA does not maximize the value of $R^2$ and hence negative $R^2$ values are possible for DEA estimators. This variance decomposition assumes a single output, however, one could compute and report separate $R^2$ statistics for each output.

### 7.4.2 Corrected CNLS

DEA builds on the minimum extrapolation principle to estimate the smallest function that envelops all data points. From the statistical point of view, insisting on the minimum extrapolation results in a systematic downward bias (i.e., the small sample error of DEA). For the deterministic model, Kuosmanen and Johnson (2010) show that a consistent and asymptotically unbiased estimator is obtained by applying a nonparametric variant of the classic COLS estimator. The proposed *corrected convex nonparametric least squares* (C$^2$NLS) estimator has always better discriminating power than DEA: the C$^2$NLS frontier envelops the DEA frontier everywhere, and the probability of finding multiple efficient units in randomly generated data approaches zero.

The C$^2$NLS method combines the nonparametric CNLS regression with the stepwise COLS approach first suggested by Winsten (1957), and more formally developed by Gabrielsen (1975) and Greene (1980). In this approach the most efficient firm in the sample is considered to be fully efficient, and the remaining inefficiency terms are normalized accordingly relative to the most efficient firm in the sample. A widely used panel data approach by Schmidt and Sickles (1984) applies a similar two-step approach (see Sect. 7.6.2 for details).

The essential steps of the C$^2$NLS routine can be described as follows:

**Step 1** Apply the CNLS estimator (7.3) to estimate the conditional mean output $E(y_i|\mathbf{x})$.

**Step 2** Identify the most efficient unit in the sample (i.e., $\hat{u}^{C2NLS}_{benchmark} = \max_{h\in\{1,\ldots,n\}} \hat{\varepsilon}^{CNLS}_h$) as the benchmark. Adjust the CNLS residuals according to $\hat{u}^{C2NLS}_i = (\max_{h\in\{1,\ldots,n\}} \hat{\varepsilon}^{CNLS}_h) - \hat{\varepsilon}^{CNLS}_i$.

**Step 3** Apply Eq. (7.5) to estimate the minimum function $\hat{f}^{CNLS}_{\min}(\mathbf{x})$. Adjust the minimum function by adding the residual of the benchmark firm to estimate the

frontier using

$$\hat{f}^{C2NLS}(\mathbf{x}) = \hat{f}^{CNLS}_{\min}(\mathbf{x}) + \hat{u}^{C2NLS}_{benchmark}$$

Thus obtained $\hat{u}_i^{C2NLS}$ can be used as measures of inefficiency in the deterministic setting without noise. The most appealing properties of the C$^2$NLS estimator can be summarized as follows:

**Theorem 5** if $\sigma_v = 0$, then the C$^2$NLS estimator is statistically consistent:

$$\operatorname*{plim}_{n\to\infty} \hat{f}^{\text{C2NLS}}(\mathbf{x}_i) = f(\mathbf{x}_i) \; for \; all \; i = 1, \ldots, n.$$

Proof. Follows from Theorem 4.1 in Kuosmanen and Johnson (2010).

**Theorem 6** the C$^2$NLS frontier envelops the DEA frontier, that is,

$$\hat{f}^{C2NLS}(\mathbf{x}) \geq \hat{f}^{DEA}(\mathbf{x}) \; \forall \mathbf{x} \in \Re^m_+.$$

Proof. Follows from Theorem 4.2 in Kuosmanen and Johnson (2010).

Note that the inefficiency estimates $\hat{u}_i^{C2NLS}$ are non-negative by construction, with the value of zero indicating full efficiency. The inefficiency measures can be converted to Farrell (1957) output efficiency scores ($\hat{\theta}_i^{C2NLS} \in [0,1]$) by using

$$\hat{\theta}_i^{C2NLS} = \frac{y_i}{\hat{f}^{C2NLS}(\mathbf{x}_i)} = \frac{y_i}{y_i + \hat{u}_i^{C2NLS}}. \tag{7.9}$$

## 7.5 Stochastic Nonparametric Envelopment of Data (StoNED)

We are now equipped to consider the general stochastic nonparametric model that does not restrict to any particular functional form of $f$ and includes both inefficiency $u$ and stochastic noise $v$. Before proceeding to estimation, we must emphasize that the shift from the deterministic case to a stochastic model is rather dramatic. For example, measuring the distance from an observed point to the frontier does not provide a measure of inefficiency if the observed point is perturbed by noise. While probabilistic inference in the deterministic case only investigates finite sample error, in the stochastic model the noise term is still relevant even if the sample size approaches infinity. Clearly, when all data points are subject to noise enveloping all observations would overestimate the true frontier production function. The CNLS regression that fits a monotonic increasing and concave curve through the middle of the cloud of data provides a natural starting point for the next generation of DEA

that can deal with noise.[11] Following Kuosmanen (2006), we refer to this approach as *stochastic nonparametric envelopment of data* (StoNED).

Analogous to the parametric COLS and MOLS (*modified OLS*) estimators and the nonparametric $C^2$NLS, the StoNED estimator consists of multiple steps. The main steps can be described as follows (a detailed description of each step follows below):

**Step 1** Apply the CNLS estimator (7.3) to estimate the conditional mean output $E(y_i\,|\mathbf{x}_i)$.

**Step 2** Apply parametric methods (e.g., the method of moments or quasi-likelihood estimation) or nonparametric methods (e.g., kernel deconvolution) to the CNLS residuals $\varepsilon_i^{CNLS}$ to estimate the expected value of inefficiency $\mu$.

**Step 3** Apply Eq. (7.5) to estimate the minimum function $\hat{g}_{\min}^{CNLS}(\mathbf{x})$. Adjust the minimum function by adding the expected inefficiency $\mu$ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \hat{\mu}$$

**Step 4** Apply parametric methods (see e.g., Jondrow et al. 1982, JLMS hereafter) or nonparametric deconvolution (e.g., kernel smoothing, Horrace and Parmeter 2011) to estimate firm-specific inefficiency using the conditional mean $E(u_i\,|\varepsilon_i^{CNLS})$.

We will next describe each step in detail, noting that each step provides alternative modeling choices (depending on the assumptions one is willing to impose), and that it is not necessary to go through all of the steps. We discuss the information available at the end of each step and the possible motivations for proceeding to further steps.

### *7.5.1 Step 1: CNLS Regression*

The CNLS estimator was described in detail in Sect. 7.3 under the assumption of no inefficiency ($u = 0$). If the observed outputs are subject to asymmetric inefficiency, as the general frontier model (7.1) assumes, then the zero-mean assumption $E(\varepsilon_i) = 0$ of regression analysis is violated. Indeed, $E(\varepsilon_i) = E(v_i - u_i) = -E(u_i) < 0$ due to the asymmetric non-negative inefficiency term. Therefore, the CNLS estimator is no longer a consistent estimator of the frontier production function $f$.

[11] Banker and Maindiratta (1992) consider maximum likelihood estimation of the unified frontier model subject to monotonicity and concavity constraints. However, their maximum likelihood problem appears to be computationally prohibitive. We are not aware of any application of this method. Gstach (1998) presents another early attempt to incorporate noise in DEA. However, he needs to make a rather restrictive assumption of truncated noise (see Simar and Wilson 2011, for sharp critique of this assumption).

Recall that CNLS regression estimates the conditional mean. Therefore, define the conditional mean function $g$ as[12]

$$g(\mathbf{x}_i) = E(y_i \,|\mathbf{x}_i) = f(\mathbf{x}_i) - E(u_i). \tag{7.10}$$

If the random inefficiency term $u$ is independent of inputs $\mathbf{x}$, then the CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is an unbiased and consistent estimator of function $g$. The CNLS estimator $\hat{g}^{CNLS}(\mathbf{x}_i)$ is obtained by solving the QP problem (7.3) and applying Eq. (7.4), as already discussed in Sect. 7.3, so we do not reproduce the CNLS formulations again here. Note that function $g$ is simply the frontier production function $f$ less the expected value of the inefficiency term $u$. If the inefficiency term $u$ has a constant variance (i.e., inefficiency term $u$ is homoscedastic), then the expected value of the inefficiency term $u$ is a constant, denoted as $\mu$. In other words, the CNLS provides a consistent estimator of the frontier $f$ minus a constant. The constant $\mu$ can be estimated based on the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, as discussed in more detail in Sect. 7.5.2. The case of heteroscedastic inefficiency where $E(u_i)$ is no longer a constant will be examined in Sect. 7.8.

Even if the data generating process (DGP) involves both inefficiency and noise, the CNLS estimator may be sufficient in some applications, without a need to proceed to the further stages. For example, if one is mainly interested in the relative efficiency rankings, then one could rank the evaluated units in descending order according to the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$. Further, if one is mainly interested in the marginal products of the input factors, the coefficients $\boldsymbol{\beta}_i$ from (7.3), which are analogous to the multiplier coefficients (shadow prices) of DEA, then the CNLS regression provides consistent estimates (Seijo and Sen 2011). The following steps described below do not influence the estimates of marginal products or the relative efficiency ranking of units. If one is interested in the frontier production function, average (in)efficiency in the sample, or cardinal firm-specific (in)efficiency estimates, then it is necessary to proceed further.

In the first step, one can impose some assumptions about returns to scale as described in Sect. 7.4.1. In addition, alternative modeling possibilities concern the multiplicative composite error and contextual variables are discussed as extensions in Sects. 7.6 and 7.7.

### *7.5.2 Step 2: Estimation of the Expected Inefficiency*

Given the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$, it is possible to estimate the expected value of the inefficiency term $\mu = E(u_i)$. Note that if the variance of the inefficiency is constant across firms (the homoscedasticity assumption), then the expectation is taken unconditional and is constant across firms.

[12] Note that we use $g$ to denote the conditional mean function when the composite error term contains inefficiency. This distinction was unnecessary in Sect. 7.3 because $g(\mathrm{x}) = f(\mathrm{x})$ when there is no inefficiency present.

Alternative approaches for estimating $\mu$ are available. We will next briefly review the commonly used parametric approaches based on the method of moments (Aigner et al. 1977), quasi-likelihood estimation (Fan et al. 1996), and the nonparametric kernel deconvolution (Hall and Simar 2002).

#### 7.5.2.1 Method of Moments

The method of moments requires some additional parametric distributional assumptions. The moment conditions are known at least for the commonly used half-normal and exponential inefficiency distributions, but not for all distributions considered in the SFA literature (e.g., the gamma distribution). In the following, we will discuss the commonly assumed case of half-normal inefficiency and normal noise. Stated formally, we assume

$$u_i \sim N^+(0, \sigma_u^2)$$

and

$$v_i \sim N(0, \sigma_v^2)$$

The CNLS residuals are known to sum to zero $\sum_{i=1}^{n} \hat{\varepsilon}_i^{CNLS} = 0$ (Seijo and Sen 2011). Hence, we can calculate the second and the third central moment of the residual distribution as

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^2/(n-1) \tag{7.11}$$

$$\hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^{CNLS})^3/(n-1). \tag{7.12}$$

The second central moment $\hat{M}_2$ is simply the sample variance of the residuals and the third central moment $\hat{M}_3$ is a component of the skewness measure. The hats on top of these statistics indicate these statistics are estimators of the true but unknown values of the central moments. If the parametric assumptions of half-normal inefficiency and normal noise hold, then the second and the third central moments are equal to

$$M_2 = \left[\frac{\pi - 2}{\pi}\right]\sigma_u^2 + \sigma_v^2 \tag{7.13}$$

$$M_3 = \left(\sqrt{\frac{2}{\pi}}\right)\left[1 - \frac{4}{\pi}\right]\sigma_u^3 \tag{7.14}$$

Note that the third moment only depends on the standard deviation of the inefficiency distribution ($\sigma_u$). Thus, given the estimated $\hat{M}_3$ (which should be negative), we can

estimate $\sigma_u$ as

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[1-\frac{4}{\pi}\right]}} \tag{7.15}$$

Subsequently, the standard deviation of the error term $\sigma_v$ is estimated based on (7.12) as

$$\hat{\sigma}_v = \sqrt{\hat{M}_2 - \left[\frac{\pi-2}{\pi}\right]\hat{\sigma}_u^2}. \tag{7.16}$$

There has been considerable discussion in the recent literature regarding the question of how to proceed if $\hat{M}_3$ is positive. Carree (2002), Alminidis et al. (2009), and Alminidis and Sickles (2012) consider alternative inefficiency distributions that allow for positive skewness. Simar and Wilson (2010) maintain the standard distributional assumptions, but suggest instead the use of bootstrapping method.

#### 7.5.2.2 Quasi-likelihood Estimation

Another way to estimate the standard deviations $\sigma_u, \sigma_v$ is to apply the quasi-likelihood method suggested by Fan et al. (1996) (who refer to it as pseudo-likelihood). In this approach we apply the standard maximum likelihood (ML) method to estimate the parameters $\sigma_u, \sigma_v$, taking the shape of the CNLS curve as given (thus the term quasi-likelihood, in contrast to the full information ML which would also parameterize the coefficients of the frontier).

One of the main contributions of Fan et al. (1996) was to show that the quasi-likelihood function can be stated as a function of a single parameter (i.e., the signal-to-noise ratio $\lambda = \sigma_u/\sigma_v$)[13] as,

$$\ln L(\lambda) = -n\ln\hat{\sigma} + \sum_{i=1}^{n}\ln\Phi\left[\frac{-\hat{\varepsilon}_i\lambda}{\hat{\sigma}}\right] - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2, \tag{7.17}$$

where

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \left(\sqrt{2}\lambda\hat{\sigma}\right) / \left[\pi\left(1+\lambda^2\right)\right]^{1/2}, \tag{7.18}$$

$$\hat{\sigma} = \left\{\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\varepsilon}_i^{CNLS}\right)^2 \Big/ \left[1-\frac{2\lambda^2}{\pi(1+\lambda)}\right]\right\}^{1/2}. \tag{7.19}$$

[13] The signal-to-noise ratio $\lambda$ should not be confused with the intensity weights $\lambda_i$ used in the envelopment formulation of DEA.

Symbol $\Phi$ denotes the cumulative distribution function of the standard normal distribution N(0,1). We first use (7.18) and (7.19) to substitute out $\hat{\varepsilon}_i$ and $\hat{\sigma}$ from (7.17). We then maximize the quasi-likelihood function (7.17) by enumerating over $\lambda$ values, using a simple grid search or more sophisticated search algorithms. When the quasi-likelihood estimate $\hat{\lambda}$ that maximizes (7.17) is found, we insert $\hat{\lambda}$ to Eqs (7.18) and (7.19) to obtain estimates of $\varepsilon_i$ and $\sigma$. Subsequently, we can calculate estimates of $\hat{\sigma}_u = \hat{\sigma}\hat{\lambda}/(1+\hat{\lambda})$ and $\hat{\sigma}_v = \hat{\sigma}/(1+\hat{\lambda})$.

A simple practical trick to conduct quasi-likelihood estimation is to use ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). By specifying the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), we can trick the ML algorithm to perform the quasilikelihood estimation. This trick can also be used for estimating models involving contextual variables or heteroscedasticity (to be explored in Sects. 7.7 and 7.8) by applying standard ML techniques as a second step.

#### 7.5.2.3 Nonparametric Kernel Density Estimation for the Convoluted Residual

While both method of moments and quasilikelihood techniques require parametric assumptions, a fully nonparametric alternative is available for estimating the signal-to-noise ratio $\lambda$, as proposed by Hall and Simar (2002). Their strategy is to search for a discontinuity in the residual density. The logic is that if an inefficiency term is left truncated, to represent efficient performance, there must be a discontinuity in distribution. When inefficiency is convoluted with noise, characterized by a continuous and smooth function, the discontinuity will still exist in the convoluted variable's density, the estimated residuals density. Thus, Hall and Simar suggest estimating the density of the residual using kernel methods and use these estimates to identify the largest change in the derivative on the right-side of the distribution (in the case of a production function and left-side in the case of the cost function). Then under the assumption of homoscedastic noise and inefficiency, the location of the largest change in the derivative can be used to estimate the mean inefficiency in the sample.

More formally, note that residuals $\hat{\varepsilon}_i^{CNLS}$ are consistent estimators of $\varepsilon_i^+ = \varepsilon_i + \mu$. Thus, we can apply the kernel density estimator for estimating the density function of $\varepsilon_i^+$. Denote the kernel density estimator by $f_{\varepsilon^+}$. Hall and Simar (2002) show that the first derivative of the density function of the composite error term ($f'_{\varepsilon}$) is proportional to that of the inefficiency term ($f'_u$) in the neighborhood of $\mu$. This is due to the assumption that $f_u$ has a jump discontinuity at zero. Therefore, a robust nonparametric estimator of expected inefficiency $\mu$ is obtained as

$$\hat{\mu} = \arg\max_{z \in \Im} (\hat{f}'_{\varepsilon^+}(z)),$$

where $\Im$ is a closed interval in the right tail of $f_{\varepsilon^+}$.

### 7.5.3 Step 3: Estimating the Frontier Production Function

In the presence of asymmetric inefficiency, the CNLS estimator estimates the conditional mean function $g(\mathbf{x}_i) = f(\mathbf{x}_i) - \mu$. Having estimated the expected inefficiency $\mu$ in Step 2, we can easily adjust the CNLS estimator to obtain an estimator of the frontier $f$. However, recall from Sect. 7.3 that the CNLS estimator of $g$ is unique at the observed points $\mathbf{x}_i$ $(i = 1,\ldots, n)$ but not in unobserved $\mathbf{x}$. Therefore, Kuosmanen and Kortelainen (2012) recommend applying the lower bound of $g$ (analogous to Eq. (7.5)), defined as

$$\hat{g}^{CNLS}_{\min}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{ \alpha + \boldsymbol{\beta}' x \,\middle|\, \alpha + \boldsymbol{\beta}'\mathbf{x}_i \geq \hat{g}^{CNLS}(\mathbf{x}_i)\ \forall i = 1,\ldots,n \right\}. \tag{7.20}$$

We can subsequently add the expected inefficiency $\mu$ to estimate the frontier using

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}^{CNLS}_{\min}(\mathbf{x}) + \hat{\mu}.$$

This equation summarizes the relation between the StoNED frontier and the CNLS estimator as well as the relation between the frontier function $f$ and the conditional mean function $g$. The heteroscedastic case where the shapes of the frontier $f$ and the regression $E(y_i\,|\mathbf{x}_i)$ are different will be discussed in Sect. 7.8 below.

### 7.5.4 Step 4: Estimating Firm-Specific Inefficiencies

Measuring the distance from an observation to frontier is not enough for estimating efficiency in the stochastic setting because all observations are subject to noise. Hence the measured distance to frontier consists of both inefficiency and noise (plus any error in our frontier estimate).

We must emphasize that even though there exist statistically unbiased and consistent methods for the estimation of the frontier $f$, there is no consistent method for estimating firm-specific efficiencies $u$ in the cross-sectional setting subject to noise. In a cross-section, estimating firm-specific realizations of a random variable $u_i$ is impossible because we have only a single observation of each firm and all observations are perturbed by noise. This is not a fault of the methods (let alone their developers), it is just impossible to predict a realization of random variable based on a single observation that is subject to noise.

In the normal—half-normal case, Jondrow et al. (1982) (JLMS) develop a formula for the conditional distribution of inefficiency $u_i$ given $\varepsilon_i$. The commonly used JLMS estimator for inefficiency is the conditional mean $E(u_i\,|\varepsilon_i)$. Given the parameter estimates $\hat{\sigma}_u$ and $\hat{\sigma}_v$, the conditional expected value of inefficiency can be

calculated as [14]

$$E(u_i \mid \hat{\varepsilon}_i) = \frac{\hat{\sigma}_u \hat{\sigma}_v}{\sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \left[ \frac{\phi\left( \frac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)}{1 - \Phi\left( \frac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right)} - \frac{\hat{\varepsilon}_i \hat{\sigma}_u}{\hat{\sigma}_v \sqrt{\hat{\sigma}_u^2 + \hat{\sigma}_v^2}} \right], \tag{7.21}$$

where $\phi$ is the density function of the standard normal distribution $N(0,1)$, $\Phi$ is the corresponding cumulative distribution function, and

$$\hat{\varepsilon}_i = \hat{\varepsilon}_i^{CNLS} - \hat{\sigma}_u \sqrt{2/\pi}$$

is the estimator of the composite error term (compare with (7.18)). It is worth to note that there is nothing "stochastic" in the Eq. (7.21): the JLMS formula is a simply a deterministic transformation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ to a new metric that represents the conditional expected value of the inefficiency term. Indeed, the rank correlation of the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ and the JLMS inefficiency estimates is equal to one (see Ondrich and Ruggiero 2001). For the purposes of relative efficiency rankings, the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ are sufficient.

Horrace and Parmeter (2011) show that the parametric assumption of the inefficiency distribution can be relaxed. Their approach still requires the parametric assumption of normally distributed noise. Rather than assuming a specific parametric distribution for the inefficiency term, the authors assume the density of $u$ belongs to the ordinary smooth family of distributions, which includes exponential, gamma or Laplace (see also Fan 1991). They apply Hall and Simar's (2002) method to estimate the jump discontinuity and thus the signal to noise ratio. Given the mean inefficiency level the authors are then able to construct the full density distribution of the inefficiency term using kernel smoothing and the residuals from a conditional mean estimation.

### 7.5.5 *Statistical Specification Tests of the Frontier Model*

As discussed above, the StoNED estimator consists of four steps. If all firms are efficient and deviations from the frontier are due to noise, the step 1 of estimating the conditional mean function is sufficient, and there is no reason to proceed further to step 2 of estimating the mean inefficiency to step 3 shifting the conditional mean function or step 4 estimating firm specific inefficiencies. To determine whether one should proceed from step 1 further to step 2, the efficiency analyst may want to test the data for evidence of inefficiency. If the results of a statistical specification test indicate that there is significant inefficiency present, this can be a convincing argument even for skeptics who believe that markets function efficiently.

[14] Note that Eq. (7.21) corrects the errors noted in formulations stated by Kuosmanen and Kortelainen (2012) and Keshvari and Kuosmanen (2013).

The residual $\hat{\varepsilon}_i^{CNLS}$ consists of two components, a normally distributed noise term and a left-truncated inefficiency term. Schmidt and Lin (1984) propose a test of the skewness of the residuals as a method to investigate if inefficiency is present. By only looking at the skewness, the method is robust to the common alternative specifications of the inefficiency term in the stochastic frontier model. Thus, the null hypothesis is the residuals are normally distributed and a $\sqrt{b_1}$ test calculated as

$$\sqrt{b_1} = \frac{M_3}{(M_2)^{3/2}} \tag{7.22}$$

Where $M_2$ and $M_3$ are, the second and third moments of the residuals respectively. The distribution of the skewness test statistic, $\sqrt{b_1}$ can be constructed by a simple Monte Carlo simulation as described in D'Agostino and Pearson (1973). The authors also provide tables with critical values of the proposed test statistic for different sample sizes.

Kuosmanen and Fosgerau (2009) consider a fully nonparametric specification test that relaxes the normality assumption of the noise term. They show that the same test statistic $\sqrt{b_1}$ considered by Schmidt and Lin (1984) can be used for testing the null hypothesis of a symmetric $v$ against the alternative hypothesis of skewness. They also recognize the $\sqrt{b_1}$ can wrongly reject the null hypothesis if the distribution is symmetric but has fat tails. Thus, they propose the additional $b_2$ test of the fourth moment

$$b_2 = \frac{M_4}{(M_2)^2} \tag{7.23}$$

Where $M_2$ and $M_4$ are the second and fourth moments of the residuals respectively. The null hypothesis is that the distribution is normally distributed. The alternative hypothesis is that there is non-normal kurtosis. The results of the $\sqrt{b_1}$ and $b_2$ tests can be given the following interpretation:

- If the null hypothesis of normality is rejected in the $\sqrt{b_1}$ test but maintained in the $b_2$ test, there is strong evidence in favor of a frontier model.
- If the null hypothesis of normality is maintained both in the $\sqrt{b_1}$ and $b_2$ tests, this supports the hypothesis of a competitive market with no inefficiency present.
- If the null hypothesis is rejected in the $b_2$ test, there may be data problems or model misspecification. There is no conclusive evidence in favor or against the frontier model.

It is worth noting that the power of the test depends on how specifically the null hypothesis and the alternative hypothesis are stated. For example, the $\sqrt{b_1}$ test of normality is more powerful than the fully nonparametric test of symmetry. If we are willing to impose some distributional assumptions for the inefficiency term, then more powerful specification tests are available. For example, Coelli (1995) proposed a variant of the Wald test to test the null hypothesis that there is no inefficiency, i.e. $\sigma_u^2 = 0$, against the alternative $\sigma_u^2 > 0$. While imposing distributional assumptions can increase the power of the test, it will also increase the risk of misspecification, which would make the statistical test inconsistent.

## 7.6 Extensions

### 7.6.1 Multiplicative Composite Error Term

Most SFA studies use Cobb-Douglas or translog functional forms where inefficiency and noise affect production in a multiplicative fashion. In the present context, it is worth noting that the assumption of constant returns to scale (CRS) would also require multiplicative error structure, as will be discussed in more detail below. Further, a multiplicative error specification implies a specific model of heteroscedasticity in which the variance of the composite error term increases with firm size.

Multiplicative composite error structure is obtained by rephrasing model (7.1) as

$$y_i = f(\mathbf{x}_i) \cdot \exp(\varepsilon_i) = f(\mathbf{x}_i) \cdot \exp(v_i - u_i) \tag{7.24}$$

Applying the log-transformation to Eq. (7.23), we obtain

$$\ln y_i = \ln f(\mathbf{x}_i) + \varepsilon_i. \tag{7.25}$$

Note that the log-transformation cannot be applied directly to inputs $\mathbf{x}$—it must be applied to the production function $f$.

In the multiplicative case, the CNLS formulation (7.3) can be rephrased as

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \varepsilon_i^{CNLS} \ \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \ \forall i \\
&\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \ \forall h, i \\
&\boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i
\end{aligned} \tag{7.26}$$

where $\phi_i + 1$ is the CNLS estimator of $E(y_i | \mathbf{x}_i)$. The value of one is added here to make sure that the computational algorithms do not try to take logarithm of zero. The first equality can be interpreted as the log transformed regression equation (using the natural logarithm function ln(.)). The second through fifth constraints are similar to (7.3) with the exception observed output in (7.3) is replaced with $\phi_i + 1$. The use of $\phi_i$ allows the estimation of a multiplicative relationship between output and input while assuring convexity of the production possibility set in original input-output space.[15]

[15] If we apply the log transformation directly to input data, the resulting frontier would be a piece-wise log-linear frontier, which has been considered in the DEA literature by Charnes et al. (1982) and Banker and Maindiratta (1986). Unfortunately, the piece-wise log-linear frontier does not generally satisfy the concavity of $f$.

Note that the log-transformation of a model variable renders the optimization formulation as a nonlinear programming (NLP) problem. These constraints are shown separately to illustrate the connection to previous formulations, but the first equality constraint can be moved to the objective function by solving and substituting for $\hat{\varepsilon}_i^{CNLS}$. Thus we have a convex solution space and a nonlinear objective function. This formulation can be solved by standard nonlinear programming algorithms and solvers. NLP solvers are available for example in such mathematical programming packages as GAMS, AIMMS, Matlab, and Lindo, among others.

In the multiplicative case, the CNLS estimator (7.25) can be applied, or as the first step of the C$^2$NLS or StoNED estimation routine. The standard method of moment, quasi-likelihood and kernel deconvolution techniques apply, as described in Sect. 7.5. However, note that in step 3 the frontier production function is obtained as $\hat{f}^{StoNED}(\mathbf{x}_i)\hat{g}_{\min}^{CNLS}(\mathbf{x}) \cdot \exp(\hat{\mu})$, where $\hat{g}_{\min}^{CNLS}(\mathbf{x})$ is the minimum function computed using Eq. (19.5) and $\exp(\hat{\mu})$ is the estimated average efficiency. A convenient feature of the multiplicative model is that $\exp(u_i)$ can be interpreted as the Farrell output efficiency measure.

### *7.6.2 Panel Data*

In panel data the sample of firms is observed repeatedly over multiple time periods. Panel data applications are common in the SFA literature and a number of alternative SFA models involving time invariant and time varying inefficiency are available (see, e.g., Greene 2008, Sect. 7.2.7). In contrast, DEA studies ignore the time dimension of the panel data and either pool the panel together as a single cross section or treat each time period as an independent cross section.[16]

The regression interpretation of DEA examined in Sect. 7.4.1 allows us to combine DEA-style axiomatic frontier with the modern panel data methods from econometrics. Kuosmanen and Kortelainen (2012, Sect. 4.1) were the first consider a fixed effects approach to estimating a time invariant inefficiency model. Their fully nonparametric panel data StoNED estimator can be seen as a nonparametric counterpart to the classic SFA approach by Schmidt and Sickles (1984). In the following we consider the random effects approach, building upon Eskelinen and Kuosmanen (2013).

Consider a data set where each firm is observed over time periods $t = 1, \ldots, T$ and define a time invariant frontier model

$$y_{it} = f(\mathbf{x}_{it}) - u_i + v_{it} \;\forall i = 1, \ldots, n \;\forall t = 1, \ldots, T, \tag{7.27}$$

where $y_{it}$ is the observed output of firm $i$ in time period $t$, $\mathbf{x}_{it}$ is a vector of inputs consumed by firm $i$ in time period $t$, and $f$ is a frontier production function that is time invariant and common to all firms. As before, $u_i$ is a firm specific inefficiency term that does not change over time, and $v_{it}$ is a random disturbance term of firm

[16] One notable exception is Ruggiero (2004).

$i$ in period $t$. Similar to the cross-sectional model, we assume that $u_i$ and $v_{it}$ are independent of inputs $\mathbf{x}_{it}$ and of each other.[17]

To estimate the model (7.27), we can adapt the standard CNLS estimator as

$$\min_{\alpha,\boldsymbol{\beta},\varepsilon} \sum_{t=1}^{T}\sum_{i=1}^{n} (\varepsilon_{it}^{CNLS})^2$$

subject to

$$\begin{aligned}
&y_{it} = \alpha_{it} + \boldsymbol{\beta}'_{it}\mathbf{x}_{it} + \varepsilon_{it}^{CNLS} \;\; \forall i = 1,\ldots,n \; \forall t = 1,\ldots,T \\
&\alpha_{it} + \boldsymbol{\beta}'_{it}\mathbf{x}_{it} \leq \alpha_{it} + \boldsymbol{\beta}'_{it}\mathbf{x}_{hs} \; \forall h,i = 1,\ldots,n \; \forall s,t = 1,\ldots,T \\
&\boldsymbol{\beta}_{it} \geq \mathbf{0} \; \forall i = 1,\ldots,n \; \forall t = 1,\ldots,T
\end{aligned} \tag{7.28}$$

where $\hat{\varepsilon}_{it}^{CNLS}$ is the CNLS residual of firm $i$ in period $t$. Note the parameters $\alpha_{it}$ and $\boldsymbol{\beta}_{it}$ that define the tangent hyperplanes of the estimated production function are specific to each firm in each time period. Thus, a piece-wise linear frontier is estimated with as many as $nT$ hyperplanes.

Given the optimal solution to (7.28), we compute the firm-specific effects as

$$\bar{\varepsilon}_i^{CNLS} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_{it}^{CNLS} \tag{7.29}$$

Following Schmidt and Sickles (1984) we measure efficiency relative to the most efficient firm in the sample (analogous to the C$^2$NLS approach considered in Sect. 7.4.2) and define

$$\hat{u}_i^{StoNED} = \left(\max_{h\in\{1,\ldots,n\}} \bar{\varepsilon}_h^{CNLS}\right) - \bar{\varepsilon}_i^{CNLS}. \tag{7.30}$$

To estimate theconditional mean function, we can adapt Eq. (7.20) to panel data as

$$\hat{g}_{\min}^{CNLS}(\mathbf{x}) = \min_{\alpha,\boldsymbol{\beta}} \left\{\alpha + \boldsymbol{\beta}'\mathbf{x} \,\middle|\, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} \geq \hat{g}^{CNLS}(\mathbf{x}_{it}) \; \forall i = 1,\ldots,n; \forall t = 1,\ldots,T\right\}.$$

The StoNED frontier estimator is then obtained as

$$\hat{f}^{StoNED}(\mathbf{x}) = \hat{g}_{\min}^{CNLS}(\mathbf{x}) + \left(\max_{h\in\{1,..,n\}} \bar{\varepsilon}_h^{CNLS}\right).$$

Both the frontier and inefficiency estimators can be shown to be statistically consistent under the assumptions stated above.

Note that the panel data StoNED estimator described above is fully nonparametric in the sense that no parametric functional form or distributional assumptions

[17] The random effects approach to panel data requires that the time invariant inefficiency is uncorrelated with inputs. This is a strong assumption. Marschak and Andrews (1944) were among the first to note that rational firm manager will adjust the inputs to take into account the technical inefficiency, and hence the observed inputs are correlated with inefficiency. In that case, the random effects estimator is biased and inconsistent. The fixed effects estimator considered by Kuosmanen and Kortelainen (2012) does not depend on this assumption.

are required. Still, the model described in Eq. (7.27) relies on two strong assumptions: (i) there is no technical progress, and (ii) inefficiency is constant over time. It is possible to relax these assumptions, but this will require some additional assumptions (typically imposing some parametric structure). Note that random effects estimator considered above may still be useful even if inefficiency changes over time. In that case, the inefficiency estimator can be interpreted as the average efficiency during the time period under study. Eskelinen and Kuosmanen (2013) propose to examine the development trajectories of the normalized CNLS residuals $\hat{\varepsilon}_{it}^{CNLS}/(\max_{h\in\{1,\ldots,n\}}\bar{\varepsilon}_{h}^{CNLS})$ to gain a better understanding how the firm performance has developed during the study period. While the normalized CNLS residuals contain random noise, a growth trend (or decline) provides a clear indication that the performance of the firm has improved (or deteriorated) during the study period.

Based on the previous discussion, two insights are worth noting:

1. Panel data is not a panacea: while we recognize that panel data provides a richer set of information, we must also acknowledge that the intertemporal setting involves complex dynamics such as technological progress and changes in efficiency over time. The random effects approach to panel data considered above would be ideal for modeling experimental data where the researcher can control the input levels and keep the production technology the same across repeated experiments. However, most panel data applications of stochastic frontiers use observational data where both the production function and the level of efficiency will likely change over time.
2. Resorting to a fully nonparametric approach does not imply freedom from restrictive assumptions. In fact, avoidance of parametric assumptions often comes at the cost of very restrictive assumptions of no noise, no technical progress, or time invariant inefficiency. Indeed, insisting on a fully nonparametric approach can be more restrictive than resorting to some parametric assumptions that allow for explicit modeling of noise, technical progress, or time varying inefficiency.

### *7.6.3 Multiple Outputs (DDF Formulation)*

The ability to model multiple inputs and multiple outputs has long been touted as an advantage of DEA over SFA: several DEA papers erroneously state that SFA cannot deal with multiple outputs. Lovell et al. (1994) and Coelli and Perelman (1999, 2000) were the first to consider a stochastic distance function model that characterizes a general multiple inputs and multiple outputs technology using the radial input and output distance functions. The recent paper by Kuosmanen et al. (2013) (henceforth KJP) examines the assumptions of the data generation process that need to be satisfied for econometric identification of the distance function when the data are subject to random noise. Although the econometric estimation of distance functions is feasible, the well-established drawbacks of SFA still apply: a functional form needs to be specified for the distance function and parametric assumptions are typically made to decompose the residual into inefficiency and noise. Further, the commonly used

parametric functional forms have the wrong curvature in output space, which is a serious problem for modeling joint production of multiple outputs.[18]

Up to this point, the CNLS/StoNED framework has been presented in the single output, multiple input setting. In this section we describe the CNLS estimator within the directional distance function (DDF) framework, Chambers et al. (1996, 1998). The CNLS formulation satisfies the axiomatic properties of the DDF by construction, models multiple inputs and multiple outputs, and accounts for stochastic noise explicitly, addressing the key limitations of both DEA and the parametric approaches. In the following we will briefly describe the stochastic data generating process (DGP) and the estimation of the DDF by CNLS. See KJP for a more detailed discussion.

The DDF indicates the distance from a given input-output vector to the boundary of the production possibility set $T$ in some pre-assigned direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$, formally,

$$\overrightarrow{D}_T(\mathbf{x}, \mathbf{y}, \mathbf{g}^x, \mathbf{g}^y) = \sup_{\theta} \left\{ \theta \,|(\mathbf{x} - \theta \mathbf{g}^x, \mathbf{y} + \theta \mathbf{g}^y) \in T \right\}. \tag{7.31}$$

Denote the reference input-output vector of firm $i$ in the direction $(\mathbf{g}^x, \mathbf{g}^y)$ by $(\mathbf{x}_i^*, \mathbf{y}_i^*)$. In this section we do not impose any particular behavioral hypothesis, but it may be illustrative to interpret $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ as the optimal solution to firm $i$'s profit maximization problem. Regardless of the firm manager's objective, we assume $(\mathbf{x}_i^*, \mathbf{y}_i^*)$ lies on the boundary of the production possibility set $T$ and hence the values of the DDF satisfy

$$\overrightarrow{D}_T(\mathbf{x}_i^*, \mathbf{y}_i^*, \mathbf{g}^x, \mathbf{g}^y) = 0 \; \forall i = 1, \ldots, n \tag{7.32}$$

The observed input-output vectors $(\mathbf{x}_i, \mathbf{y}_i), i = 1, ..., n$, are perturbed in direction $(\mathbf{g}^x, \mathbf{g}^y) \in \Re_+^{m+s}$ by random inefficiency $u_i$ and noise $v_i$, which form the composite error term $\varepsilon_i = u_i + v_i$ (note the positive sign of the inefficiency term $u_i$). Specifically, the observed data are perturbed versions of the optimal input-output vectors as follows

$$(\mathbf{x}_i, \mathbf{y}_i) = (\mathbf{x}_i^* + \varepsilon_i \mathbf{g}^x, \mathbf{y}_i^* - \varepsilon_i \mathbf{g}^y) \; \forall i = 1, \ldots, n \tag{7.33}$$

We assume the inefficiency and noise terms satisfy the assumptions discussed in Sect. 7.2. Note that the elements of the direction vector $(\mathbf{g}^x, \mathbf{g}^y)$ represent the impacts of inefficiency and noise on specific input and output variables. If an element of $(\mathbf{g}^x, \mathbf{g}^y)$ is equal to zero, it means that the corresponding input or output variable is immune to both inefficiency and noise in the DGP. The larger the value of an element of $(\mathbf{g}^x, \mathbf{g}^y)$ in the DGP, the larger the impact of inefficiency and noise on the corresponding input or output variable is. Interestingly, Proposition 3 in KJP shows

[18] The wrong curvature violates some of the most elementary properties of production technologies. For example, the Cobb-Douglas or translog specifications of the distance function will violates the basic properties of null jointness and unboundedness (see, e.g., Färe et al. 2005). Another problem concerns the economies of scope (e.g., Panzar and Willig 1981). For example, the Cobb-Douglas distance function cannot capture the economies of scope at any parameter values. Since the economic rationale for joint production is rooted to economies of scope, it is contradictory to apply a technology that exhibits economies of specialization for modeling joint production.

that in the DGP described above the value of the DDF equals the composite error term:

$$\overrightarrow{D}_T(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) = \varepsilon_i \ \forall i.$$

This result provides implicitly a regression equation for estimating the DDF. We can resort to a similar stepwise procedure as described in Sect. 7.5.

The first step is to estimate the conditional mean distance defined as

$$d(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) = \overrightarrow{D}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) - \mu \tag{7.34}$$

Let $\Delta$ denote the set of functions that satisfy the axioms of free disposability, convexity, and the translation property.[19] We can adapt the CNLS estimator to the DDF setting by postulating the following infinite dimensional least squares problem

$$\min_{d} \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y)^2$$

subject to

$$d \in \Delta \tag{7.35}$$

Formulation (7.35) is a complex, infinite dimensional optimization problem that cannot be solved by brute-force numerical methods. The main challenge is to find a way to parameterize the infinitely large set of functions that satisfy the stated regularity conditions. Here again we apply insights from Kuosmanen (2008) and show an equivalent finite dimensional representation in terms of quadratic programming. Consider the following QP problem

$$\min_{\alpha,\boldsymbol{\beta},\boldsymbol{\gamma},\varepsilon} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2$$

subject to

$$\boldsymbol{\gamma}'_i \mathbf{y}_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i - \varepsilon_i^{CNLS} \ \forall i = 1, \ldots, n$$
$$\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i - \boldsymbol{\gamma}'_i \mathbf{y} \leq \alpha_h + \boldsymbol{\beta}'_i \mathbf{x}_i - \boldsymbol{\gamma}'_h \mathbf{y}_i \ \forall h, i = 1, \ldots, n$$
$$\boldsymbol{\gamma}'_{ii} \mathbf{g}^y + \boldsymbol{\beta}'_{ii} \mathbf{g}^x = 1 \ \forall i = 1, \ldots, n$$
$$\boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i = 1, \ldots, n$$
$$\boldsymbol{\gamma}_i \geq \mathbf{0} \ \forall i = 1, \ldots, n \tag{7.36}$$

Note that the residual $\hat{\varepsilon}_i^{CNLS}$ here represents the estimated value of $d_i$(i.e., $\overrightarrow{D}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) + u_i$). We also introduce new firm-specific coefficients $\boldsymbol{\gamma}_i$ that represent marginal effects of outputs to the DDF. The first constraint defines the distance

[19] The translation property, Chambers et al. (1998), states that if we move from the initial point $(\mathbf{x}, \mathbf{y})$ in the direction $(\mathbf{g}^x, \mathbf{g}^y)$ by factor $\alpha$, i.e., to the point $(\mathbf{x} + \alpha\mathbf{g}^x, \mathbf{y} - \alpha\mathbf{g}^y)$, then the distance to the frontier decreases by $\alpha$. This property is crucial for the internal consistency of the DDF and can be seen as an additive analogue of the linear homogeneity property of the input distance function.

to the frontier as a linear function of inputs and outputs. The linear approximation of the frontier is based on the tangent hyperplanes, analogous to the original CNLS formulation. The second set of constraints is the system of Afriat inequalities that impose global concavity. The third constraint is a normalization constraint that ensures the translation property. The last two constraints impose monotonicity in all inputs and outputs. It is straightforward to show that the CNLS estimator of function $d$ satisfies the axioms of free disposability, convexity, and the translation property (see Theorem 3 in KJP).

After solving the CNLS problem, one can proceed to estimate the deterministic frontier by Corrected CNLS as described in Sect. 7.4.2 or the stochastic frontier by StoNED as described in Sect. 7.5.2. Note that the CNLS estimator described above does not estimate the DDF directly, but rather $\overrightarrow{D}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{g}^x, \mathbf{g}^y) + E(u_i)$. If the inefficiency term is homoscedastic, then the techniques described in Sect. 7.5.2 apply for the estimation of $E(u_i) = \mu$. The case of heteroskedastic inefficiency term is discussed in Sects. 7.8.2 and 7.8.3 below. Subsequently, the estimate of the DDF is obtained by shifting the CNLS estimate of function $d$ in direction $(\mathbf{g}^x, \mathbf{g}^y)$ by the estimated expected inefficiency.

To connect the multi-output DDF to the single output case, it is worth noting in the single output case, specifying the direction vector as $g^y = 1$ and $\mathbf{g}^x = \mathbf{0}$, the CNLS problem (7.36) reduces to

$$\min_{\alpha,\boldsymbol{\beta},\varepsilon} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2$$

subject to

$$\begin{aligned} & y_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i - \varepsilon_i^{CNLS} \ \forall i = 1, \ldots, n \\ & \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \ \forall h, i = 1, \ldots, n \\ & \boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i = 1, \ldots, n \end{aligned} \tag{7.37}$$

This formulation is equivalent to the CNLS formulation (7.3) developed in Kuosmanen (2008), except for the sign of the residual $\hat{\varepsilon}_i^{CNLS}$ in the first constraint. Note that the DDF has positive values below the frontier and negative values above the frontier, which explains the negative sign.

### 7.6.4 Convex Nonparametric Quantile Regression and Asymmetric Least Squares

While CNLS estimates the conditional mean $E(y_i \,|\mathbf{x}_i)$, quantile regression aims at estimating the conditional median or other quantiles of the response variable (Koenker and Bassett 1978; Koenker 2005).[20] Denoting the pre-assigned quantile by parameter

[20] In the DEA literature, the quantile frontiers are commonly referred to as robust order-$m$ and order-$\alpha$ frontiers (e.g., Aragon et al. 2005; Daouia and Simar 2007). However, while quantile frontiers are

$q \in (0,1)$, we can modify the CNLS problem (7.3) to estimate convex nonparametric quantile regression (CNQR) (Wang et al. 2014) as follows:[21]

$$\min_{\alpha,\boldsymbol{\beta},\varepsilon^+,\varepsilon^-} q\sum_{i=1}^{n}\varepsilon_i^+ + (1-q)\sum_{i=1}^{n}\varepsilon_i^-$$

subject to

$$y_i = \alpha_i + \boldsymbol{\beta}'_i\mathbf{x}_i + \varepsilon_i^+ - \varepsilon_i^- \;\forall i$$

$$\alpha_i + \boldsymbol{\beta}'_i\mathbf{x}_i \le \alpha_h + \boldsymbol{\beta}'_h\mathbf{x}_i \;\forall h,i$$

$$\boldsymbol{\beta}_i \ge \mathbf{0}\;\forall i$$

$$\varepsilon_i^+ \ge 0\;\forall i$$

$$\varepsilon_i^- \ge 0\;\forall i \tag{7.38}$$

The CNQR problem differs from CNLS in that the composite error term is now broken down to two non-negative components $\varepsilon_i^+, \varepsilon_i^- \ge 0$. The objective function minimizes the asymmetric absolute deviations from the frontier instead of symmetric quadratic deviations. The pre-assigned weight $q$ defines the quantile to be estimated. For example, by setting $q = 0.05$, the piece-wise linear CNQR function will allow at most 5 % of observations to lie above the fitted function and envelope at most 95 % of the observed data points. As the sample size approaches to infinity, the $q$-order frontier will envelop exactly $q$ percent of the observed data points (Wang et al. 2014, Theorem 1). Two important special cases are worth noting. First, if we set $q = 0.5$, then CNQR estimates the conditional median (whereas CNLS estimates the conditional mean). Secondly, as $q$ approaches to zero, the negative deviations $\varepsilon_i^-$ get a larger weight, and the CNQR approaches to the DEA frontier.

An appealing feature of the CNQR formulation is that its objective function and all constraints are linear functions of unknown parameters, and hence the CNQR problem can be solved by standard linear programming (LP) algorithms. However, a major drawback compared to CNLS is that the optimal solution to the CNQR problem is not necessarily unique, not even for the observed data points $(\mathbf{x}_i, y_i), i = 1,\ldots,n$. In econometrics, non-uniqueness of quantile regression is usually assumed away by assuming the regressors $\mathbf{x}$ are randomly drawn from a continuous distribution. In practice, however, input vectors $\mathbf{x}$ are not randomly drawn, and there may be two or more firms use exactly the same amounts of inputs (i.e., $\mathbf{x}_i = \mathbf{x}_j$ for firms $i$ and $j$). In our experience, non-uniqueness of CNQR seems to be particularly a problem in samples where inputs $\mathbf{x}$ are discrete variables. Wang et al. (2014) recognize non-uniqueness of the CNQR estimator, illustrating the problem with a numerical example.

One possible way to resolve the non-uniqueness problem is to apply the asymmetric least squares criterion suggested by Newey and Powell (1987), and reformulate

more robust to outliers than the conventional DEA frontiers, the quantile DEA approaches typically assume away noise.

[21] Similar quantile formulation was first considered by Banker et al. (1991), who refer to it as "stochastic DEA".

the CNQR problem as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} q\sum_{i=1}^{n}(\varepsilon_i^+)^2+(1-q)\sum_{i=1}^{n}(\varepsilon_i^-)^2$$

subject to

$$\begin{aligned}
&y_i=\alpha_i+\boldsymbol{\beta}'_i\mathbf{x}_i+\varepsilon_i^+-\varepsilon_i^-\ \forall i\\
&\alpha_i+\boldsymbol{\beta}'_i\mathbf{x}_i\le\alpha_h+\boldsymbol{\beta}'_h\mathbf{x}_i\ \forall h,i\\
&\boldsymbol{\beta}_i\ge\mathbf{0}\ \forall i\\
&\varepsilon_i^+\ge 0\ \forall i\\
&\varepsilon_i^-\ge 0\forall i
\end{aligned}\tag{7.39}$$

To our knowledge, this asymmetric least squares formulation has not been considered before; we will henceforth refer to it as convex asymmetrically weighted least squares (CAWLS). The CAWLS problem differs from CNQR only in terms of the objective function, which now minimizes the asymmetric squared deviation instead of the absolute deviations. In the case of the linear regression, Newey and Powell (1987) show that the properties of the asymmetric least squares estimator are analogous to those of the quantile regression, but the asymmetric least squares can be more convenient for statistical inferences. In the present context, we hypothesize that the use of the quadratic loss function similar to CNLS ensures that the optimal solution to the CAWLS problem is always unique for the observed data points $(\mathbf{x}_i, y_i), i = 1,\ldots,n$. We leave confirming or rejecting this hypothesis as an open question for future research. Besides the question of uniqueness, the statistical properties of both CNQR and CAWLS would require further research.

CNQR and CAWLS formulations allow one to estimate the $q$-quantile or $q$-expectile frontiers directly, without a need to impose parametric distributional assumptions for the inefficiency and noise terms or resort to stepwise estimation along the lines described in Sect. 7.5. This is one of the attractive properties of CNQR and CAWLS. For the purposes of efficiency analysis, however, the use of quantiles or asymmetric weighted least squares is not a panacea. It is important to stress that the distance from the frontier, measured as $\hat{\varepsilon}_i^{CNQR}=\hat{\varepsilon}_i^+-\hat{\varepsilon}_i^-$ or $\hat{\varepsilon}_i^{CAWLS}=\hat{\varepsilon}_i^+-\hat{\varepsilon}_i^-$ (note: in both cases the residuals satisfy $\hat{\varepsilon}_i^+\hat{\varepsilon}_i^-=0\ \forall i$), should not be interpreted as a measure of inefficiency, as the distance to frontier also includes noise. To estimate conditional expected value of inefficiency along the lines of JLMS, we still need to resort to stepwise estimation. One possibility is to replace CNLS by CNQR or CAWLS as the first step of the StoNED procedure outlined in Sect. 7.5. Of course, residuals $\hat{\varepsilon}_i^{CNQR}$ or $\hat{\varepsilon}_i^{CAWLS}$ can be used as such for relative performance rankings, but such performance rankings obviously depend on the chosen parameter value of $q$. Wang et al. (2014) examine the specification of $q$ for frontier estimation, showing that the optimal value of $q$ is a monotonically decreasing function of the signal to noise ratio $\lambda=\sigma_u/\sigma_v$. One may set the value of $q$ based on subjective judgment, but in real world applications (consider, e.g., regulation of electricity distribution

networks; see Kuosmanen 2012; Kuosmanen et al. 2013), some objective criteria for specifying $q$ would be important.

One appealing feature of the $q$-quantile and $q$-expectile frontiers is that they are robust to heteroscedasticity. Therefore, testing of and dealing with heteroscedasticity provide one promising application area for the CNQR and CAWLS techniques. If the composite error term is homoscedastic, then the quantile and expectile frontiers should have similar shapes at different values of $q$. Newey and Powell (1987) apply this idea for testing heteroscedasticity. We return to this issue in more detail in Sect. 7.8.

## 7.7 Contextual Variables

A firm's ability to operate efficiently often depends on operational conditions and practices, such as the production environment and the firm specific characteristics for example technology selection or managerial practices. Banker and Natarajan (2008) refer to both variables that characterize operational conditions and practices as *contextual variables*. Currently two-stage DEA (2-DEA) is widely applied to investigate the importance of contextual variables as summarized by the citations included in Simar and Wilson (2007). However, its statistical foundation has been subject to sharp debate between Simar and Wilson (2007, 2011) and Banker and Natarajan (2008) (see also Hoff 2007; McDonald 2009). In this section we shed some new light on this debate following Johnson and Kuosmanen (2011, 2012).

It is important to note that Simar and Wilson (2007, 2011) do not consider stochastic noise in their DGP. In contrast, Banker and Natarajan (2008) introduce a noise term that has a doubly-truncated distribution, following the DEA+ approach by Gstach (1998). In this setting, Johnson and Kuosmanen (2012) show that the 2-DEA estimator of contextual variables is consistent under more general assumption that those stated by Banker and Natarajan (2008) and criticized by Simar and Wilson (2011). Further, Johnson and Kuosmanen (2012) employ the least squares formulation of DEA to develop a one-stage DEA method (1-DEA) for estimating the effects of the contextual variables. Relaxing the peculiar assumption of truncated noise,[22] Johnson and Kuosmanen (2011) develop *stochastic (semi-) nonparametric envelopment of z-variables data* (StoNEZD).

[22] We label this assumption as peculiar because it contradicts standard statistical assumptions, namely, the residual term is often model as normally distributed because a mixture of a large number of unknown distributions is approximately normal in finite samples and asymptotically normal. The large number of unknown distributions is a result of measurement errors, modeling simplifications, and other sources of noise. Thus, the motivation for truncated normal distribution used in Gstach (1998) and Banker and Natarajan (2008) is lacking and peculiar as also noted by Simar and Wilson (2011). Johnson and Kuosmanen (2012) argue this truncation may come from an outlier detection procedure that would remove extreme observations from the analysis. However, in this case 1-DEA (introduced below) would still be preferred to 2-DEA because the bias introduced in two-stage estimation.

Taking the multiplicative model described in Sect. 7.6.1 as our starting point, we introduce the contextual variables, represented by $r$-dimensional vectors $\mathbf{z}_i$ that represent the measured values of operational conditions and practices, to obtain the following semi-nonparametric, partial log-linear equation

$$\ln y_i = \ln f(\mathbf{x}_i) + \boldsymbol{\delta}'\mathbf{z}_i + v_i - u_i. \tag{7.40}$$

In this equation, parameter vector $\boldsymbol{\delta} = (\delta_1 \ldots \delta_r)'$ represents the marginal effects of contextual variables on output. All other variables maintain their previous definitions.

In the following sub-sections we will present two-stage DEA (2-DEA), one-stage DEA, and StoNEZD estimators. First, the 2-DEA estimator is described and the statistical properties of it are discussed. Given the assumptions necessary for the consistency of two-stage DEA method we then present the one-stage alternative. The joint estimation avoids the bias in the DEA frontier being transmitted to the parameter estimates of the coefficients on the contextual variables; however, the frontier estimated is still the minimum envelopment of the data and thus does not account for noise in the production model or input/output data. To account for stochastic noise, StoNEZD is introduced in 7.3.

### *7.7.1 Two-Stage DEA*

The literature on 2-DEA includes a number of variants. This sub-section follows the approach by Banker and Natarajan (2008). The two stages of their 2-DEA method are the following. In the first stage, the frontier production function $f$ is estimated using the nonparametric DEA estimator formally stated as (7.5). The DEA output efficiency estimator of firm $i$ is stated as $\hat{\theta}_i^{\text{DEA}} = y_i / \hat{f}^{DEA}(\mathbf{x}_i)$ and computed as

$$(\hat{\theta}_i^{\text{DEA}})^{-1} = \max_{\theta \in \Re, \lambda \in \Re_+^n} \left\{ \theta \,\middle|\, \theta y_i \leq \sum\nolimits_{h=1}^{n} \lambda_h y_h;\ \mathbf{x}_i \geq \sum\nolimits_{h=1}^{n} \lambda_h \mathbf{x}_h; \sum\nolimits_{h=1}^{n} \lambda_h = 1 \right\} \tag{7.41}$$

In the second stage, the following linear equation is estimated using OLS or ML

$$\ln\ \hat{\theta}_i^{\text{DEA}} = \alpha + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{2-DEA},\ i = 1, \ldots, n, \tag{7.42}$$

where the intercept $\alpha$ captures the expected inefficiency and the finite sample bias of the DEA estimator, and the composite disturbance term $\varepsilon_i^{2-DEA}$ captures the noise term $v_i$ and the deviations of $u_i$ from the expected inefficiency $\mu$. Note that the dependent variable has the "hat" because the DEA efficiency estimate is computed beforehand using (7.41), whereas the parameters on the right hand side of (7.42) are estimated using OLS or ML in a second stage.

Johnson and Kuosmanen (2012) state that the 2-DEA estimator is statistically consistent in the case of truncated noise as shown by Banker and Natarajan (2008), however, the assumptions required for consistency in Banker and Natarajan are unnecessarily restrictive.

Let $\mathbf{Z}$ denote a $n \times r$ matrix of contextual variables. Assume the noise terms are truncated as $|v_i| \leq V^M$ and denote $\mathbf{v} = (v_1, ..., v_n)'$. Denote the domains of vectors $\mathbf{x}$ and $\mathbf{z}$ by $D_x$ and $D_z$, respectively. Then the statistical consistency of the 2-DEA estimator can be established under the relaxed set of assumptions as follows.

**Theorem 7** If the following five assumptions are satisfied

(i) sequence $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), \ i = 1, \ldots, n\}$ is a random sample of independent observations,
(ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ is a positive definite matrix,
(iii) noise term $\mathbf{v}$ has a truncated distribution: $|\mathbf{v}| \leq V^M\mathbf{1}$, $f_v(V^M) > 0$,
(iv) elements of domain $D_z$ are bounded from above or below such that $\boldsymbol{\delta}'\mathbf{z}$ has a finite maximum $\zeta = \max_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$ at a point $\mathbf{z}^{\xi} \in \arg\max_{\mathbf{z}\in D_z} \boldsymbol{\delta}'\mathbf{z}$,
(v) the joint density $f$ is continuous and satisfies $f(\mathbf{x}, \mathbf{z}^{\xi}, 0, V^M) > 0$ for all $\mathbf{x} \in D_x$,

then the 2-DEA estimators are statistically consistent in the following sense

$$\operatorname*{plim}_{n\to\infty} \hat{f}^{\text{DEA}}(\mathbf{x}_i) = f(\mathbf{x}_i) \cdot \exp(V^M + \zeta) \ for \ all \ i = 1, \ldots, n,$$

$$\operatorname*{plim}_{n\to\infty} \boldsymbol{\delta}^{\text{2-DEA}} = \boldsymbol{\delta}$$

Proof. See Johnson and Kuosmanen (2012), Theorem 1.

This theorem by Johnson and Kuosmanen (2012) generalizes the consistency result by Banker and Natarajan (2008) result by relaxing the following two assumptions:

1. inputs and contextual variables are statistically independent,
2. the effect of contextual variables is one-sided: $\mathbf{Z} \geq \mathbf{0}, \boldsymbol{\delta} \leq \mathbf{0}$.

Note that the DEA frontier does not converge to the true frontier f, it converges to $f(\mathbf{x}) \cdot \exp(V^M + \zeta)$ (i.e., the frontier augmented by the maximum noise $V^M$ under the ideal conditions represented by $\mathbf{z}^{\zeta}$) thus estimation of the frontier requires observing firms that are operating efficiently and are operating in the best environment and happen to get a noise drawn close to the upper bound $V^M$.

Consistency is a relatively weak property. In practice a data set will be finite in size and probably not as large as we would like. However, Johnson and Kuosmanen (2012) are able to provide the explicit form of the bias in the 2-DEA estimator. Specifically it depends on the bias of the DEA frontier ($\hat{f}^{\text{DEA}}$) as follows:

$$\text{Bias}(\hat{\boldsymbol{\delta}}^{\text{2-DEA}}) = -(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \left[\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X}))\right], \tag{7.43}$$

where

$$\text{Bias}(\hat{f}^{\text{DEA}}(\mathbf{X})) = \begin{pmatrix} E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_1)) - f(\mathbf{x}_1) \cdot \exp(V^M + \zeta) \\ \vdots \\ E(\ln \hat{f}^{\text{DEA}}(\mathbf{x}_n)) - \ln f(\mathbf{x}_n) \cdot \exp(V^M + \zeta) \end{pmatrix}. \tag{7.44}$$

Thus, the bias of the first-stage DEA estimator carries over to the second-stage OLS regression. Importantly, the bias of the second-stage OLS estimator is due to the correlation of $\mathbf{Z}$ and bias of the first-stage DEA estimator.

In summary we would like to emphasize two critical points about 2-DEA.

1. correlation of inputs and contextual variables does not influence the statistical consistency of 2-DEA estimator as long as the columns of $\mathbf{X}$ and $\mathbf{Z}$ matrices are not linearly dependent.
2. the bias of the DEA frontier in the first-stage carries over to the second-stage OLS estimator through the correlation of the DEA frontier with the contextual variables.

We note that statistical independence of inputs and contextual variables does not necessarily guarantee that Bias($\hat{f}^{\text{DEA}}(\mathbf{X})$) is uncorrelated with $\mathbf{Z}$. Thus, 2-DEA does not suffer from some of the problems noted by Simar and Wilson (2011) and in fact requires significantly weaker assumptions than Banker and Natarajan (2008) suggest. However, the DEA frontier is always biased downward in a finite sample and thus this bias may be transferred to the estimation of the effect of the contextual variables. The following two sub-sections propose alternatives building on the regression interpretation of DEA which do not suffer from this bias.

### *7.7.2 One-Stage DEA*

The fundamental problem of the 2-DEA procedure is that the impact of the contextual variables $\mathbf{Z}$ is not taken into account in the first stage DEA. This problem has been recognized in the SFA literature, where the standard approach is to jointly estimate the frontier and the impacts of the contextual variables (e.g., Wang and Schmidt 2002). In the similar vein, the least squares regression interpretation of DEA described in Sect. 7.4.1 allows us to estimate the DEA frontier and the coefficients $\delta$ jointly. Specifically, we can introduce the contextual variables to the least squares formulation of DEA, stated as the QP problem (7.6), to obtain:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\phi,\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{1-DEA})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}'\mathbf{z}_i + \varepsilon_i^{1-DEA} \ \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \ \forall i \\
&\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \ \forall h, i \\
&\boldsymbol{\beta}_i \geq \mathbf{0} \ \forall i \\
&\varepsilon_i^{1-DEA} \leq V^M \ \forall i
\end{aligned} \tag{7.45}$$

Notable differences compared to the problem (7.7) concern the use of the log-transformation to enforce the multiplicative formulation of the inefficiency term

(compare with Sect. 7.6.1) and the truncation of the residual $\varepsilon_i^{1-DEA}$ at point $V^M$. Note that by setting $V^M = 0$ restricts the noise term to zero, and the 1-DEA formulation reduces to the joint estimation of the effect of the contextual variables and the classic deterministic DEA frontier where all input/output data is observed exactly and residuals are non-positive.

Note further that the parameter vector $\delta$ is common to all observations, and hence it can be harmlessly omitted from the Afriat inequalities that impose convexity. In fact, the contextual variables can be interpreted as inputs that have constant marginal products across all firms[23] (i.e., we can think of matrix $\mathbf{Z}$ as a subset of $\mathbf{X}$ for which $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j \; \forall i, j$).

The statistical properties of the 1-DEA estimator generally depend on the specification of the truncation point $V^M$. Performance of the 1-DEA estimator has been investigated via Monte Carlo simulations in Johnson and Kuosmanen (2012) where the authors find that 1-DEA performs well even when the truncation point is misspecified. However, the assumption of truncated noise (i.e., $|v_i| \leq V^M$) is non-standard and debatable (see, e.g., Simar and Wilson 2011). While the consistency of 2-DEA critically depends on this assumption, the CNLS estimator allows us to harmlessly relax it. The next sub-section discusses the StoNED estimator with z-variables that does not rely on the truncated noise assumption.

### 7.7.3 StoNED With z-Variables (StoNEZD)

Relaxing the assumption of truncated noise, we can apply CNLS to jointly estimate the expected output conditional on inputs and the effects of the contextual variables. Johnson and Kuosmanen (2011) were the first to explore this approach, referring to it as StoNED with z-variables (StoNEZD). StoNEZD incorporates the contextual variables to the stepwise procedure sescribed in Sect. 7.5. In the following, we will focus on the CNLS estimator applied in the first step: steps 2–4 follow as described in Sect. 7.5, and are hence omitted here.

To incorporate the contextual variables in step 1 of the StoNED estimation routine, we can refine the multiplicative CNLS problem as follows:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\phi},\boldsymbol{\varepsilon}} \sum_{i=1}^{n} (\varepsilon_i^{CNLS})^2 \\
&\text{subject to} \\
&\ln y_i = \ln(\phi_i + 1) + \boldsymbol{\delta}' \mathbf{z}_i + \varepsilon_i^{CNLS} \; \forall i \\
&\phi_i + 1 = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \; \forall i \\
&\alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \; \forall h, i \\
&\boldsymbol{\beta}_i \geq \mathbf{0} \; \forall i
\end{aligned} \tag{7.46}$$

[23] This interpretation would vary slightly if the $\delta_i$ is negative. Then the contextual variable would be an output which would reduce the firm's ability to produce $y$.

Note that problem (7.46) is identical to (7.45), except that the truncation constraint $\varepsilon_i \leq V^M \ \forall i$ has been removed. Therefore, the least squares residuals are unrestricted, and hence problem (7.46) is a genuine conditional mean regression estimator.

Denote by $\hat{\boldsymbol{\delta}}^{StoNEZD}$ the coefficients of the contextual variables obtained as the optimal solution to (7.46). Johnson and Kuosmanen (2011) examine the statistical properties of this estimator in detail, showing its unbiasedness, consistency, and asymptotic efficiency.[24] Most importantly, the authors show that the conventional methods of statistical inference from linear regression analysis (e.g., t-tests, confidence intervals) can be applied for asymptotic inferences regarding coefficients $\boldsymbol{\delta}$. Their main result can be summarized as follows:

**Theorem 8** If the following conditions are satisfied

i) sequence $\{(y_i, \mathbf{x}_i, \mathbf{z}_i), i = 1,\ldots,n\}, i = 1,\ldots,n\}$ is a random sample of independent observations,
ii) $\lim_{n\to\infty} \mathbf{Z}'\mathbf{Z}/n$ is a positive definite matrix,
iii) the inefficiency terms $\mathbf{u}$ and the noise terms $\mathbf{v}$ are identically and independently distributed (i.i.d.) random variables with $Var(\mathbf{u}) = \sigma_u^2\mathbf{I}$ and $Var(\mathbf{v}) = \sigma_v^2\mathbf{I}$,

then the StoNEZD estimator for the coefficients of the contextual variables ($\hat{\boldsymbol{\delta}}^{StoNEZD}$) is statistically consistent and asymptotically normally distributed according to:

$$\hat{\boldsymbol{\delta}}^{StoNEZD} \sim_a N\left(\boldsymbol{\delta}, (\sigma_v^2 + \sigma_u^2)(\mathbf{Z}'\mathbf{Z})^{-1}\right).$$

Proof. See Johnson and Kuosmanen (2011), Theorem 2.

This theorem extends the standard result of asymptotic normality of the OLS coefficients to the StoNEZD estimator of the contextual variables. In other words, even though model (7.40) includes a nonparametric function in addition to a linear regression function, the presence of the nonparametric function does not affect the limiting distribution of the parameter estimator in the linear part. In addition, Johnson and Kuosmanen (2011) show that the estimator $\hat{\boldsymbol{\delta}}^{StoNEZD}$ converges at the standard parametric rate, despite the presence of the nonparametric part in the regression equation. Therefore, we can apply the standard techniques from regression analysis such as *t*-tests and confidence intervals for asymptotic inferences.

A simple trick to compute standard errors for $\hat{\boldsymbol{\delta}}^{StoNEZD}$ is to run OLS regression where the contextual variables $\mathbf{Z}$ are regressors and the dependent variable is the difference between the natural log of observed output subtracting the natual log of the input aggregation plus 1, specifically $\ln y_i - \ln(\hat{\phi}_i + 1) = \hat{\boldsymbol{\delta}}'\mathbf{z}_i + \hat{\varepsilon}_i^{CNLS}$. This OLS regression will yield the same coefficients $\hat{\boldsymbol{\delta}}'^{StoNEZD}$ that were obtained as the optimal solution to problem (7.46),[25] but also return the standard errors and other standard diagnostic statistics such as t-ratios, p-values, and confidence intervals.

[24] Johnson and Kuosmanen (2012) report some Monte Carlo evidence of the finite sample performance of the StoNEZD estimator.

[25] Note that this two-stage regression procedure is not subject to the problems of the 2-DEA procedure because we do control for the effects of the contextual variables in the first stage CNLS

## 7.8 Heteroscedasticity

Up to this point we have assumed that the composite error term is homoscedastic, implying the variance parameters $\sigma_u^2$ and $\sigma_v^2$ are constant across all firms. This is a standard assumption both in regression analysis and in the parametric literature of frontier estimation (e.g., Aigner et al. 1977). However, this assumption is not always realistic in applications.

We can relax the assumption of constant $\sigma_u^2$ and $\sigma_v^2$, and allow these parameters to be firm specific (i.e., $\sigma_{u,i}^2$ and $\sigma_{v,i}^2$), and potentially dependent on inputs $\mathbf{x}$ and contextual variables $\mathbf{z}$. We stress that the least squares approach considered in this paper enables us to apply standard econometric techniques of testing and modeling heteroscedasticity considered in the SFA literature (see, e.g., Kumbhakar et al. 1991; Caudill and Ford 1993; Caudill et al. 1995; Battese and Coelli 1995; Hadri 1999; and Kumbhakar and Lovell 2000). The purpose of this section is to provide a brief review of how some of those techniques could be adapted for the purposes of CNLS and StoNED.

The first question to consider is how would heteroscedasticity affect the CNLS and StoNED estimators if we simply ignore it? Like standard OLS, the CNLS estimator remains unbiased and consistent despite heteroscedasticity. A weighted CNLS estimator (to be considered below) might be more efficient, provided that the heteroscedastic variance parameters can be estimated with a sufficient precision. However, heteroscedasticity is not a major problem for CNLS, and trying to improve its performance through explicit modeling and estimation of heteroscedasticity may not be worth the effort. Further research would be needed to investigate this issue.

The stepwise StoNED procedure is more sensitive to heteroscedasticity, as discussed by Kuosmanen and Kortelainen (2012). At this point, we need to distinguish between (i) heteroscedastic inefficiency term and ii) heteroscedasticity noise term. Ignoring type (ii) heteroscedasticity is less harmful in the StoNED estimation because the skewness of the CNLS residuals is still driven by the homoscedastic inefficiency term, the expected value of inefficiency is constant, and hence the shape of the regression function (i.e., the conditional mean $E(y_i|\mathbf{x}_i)$) is identical to that of the frontier production function $f$. Type (i) heteroscedasticity will cause bigger problems, as Kuosmanen and Kortelainen (2012) recognize. If the inefficiency term is heteroscedastic, then the expected value of inefficiency is no longer constant, and the shapes of the regression function and the frontier production function will diverge. To take both types of heteroscedasticity explicitly into account, in Sect. 7.8.2 we will consider a doubly-heteroscedastic model where both inefficiency and noise terms are heteroscedastic. But before proceeding to the explicit modeling of heteroscedasticity, we describe a diagnostic test of the homoscedasticity assumption.

---

regression. It is just a computational trick to calculate the standard errors, but it can also serve as a simple diagnostic check that the solution to problem (7.32) is indeed optimal with respect to the contextual variables.

### 7.8.1 White Test of Heteroscedasticity Applied to CNLS

Although the heteroscedastic inefficiency term would bias the StoNED estimator, it is important to emphasize that we do not need to take the homoscedasticity assumption by faith. Standard econometric tests of heteroscedasticity such as the White or the Breusch-Pagan tests are directly applicable to CNLS residuals. In this sub-section we briefly describe how the White (1980) test can be applied following Kuosmanen (2012).

The null hypothesis of the White test is that composite error term is homoscedastic, that is, $H_0$: $\sigma_{\varepsilon,i} = \sigma_{\varepsilon,j}\ \forall i, j$. The alternative hypothesis states there is heteroscedasticity, that is, $H_1$: $\sigma_{\varepsilon,i} \neq \sigma_{\varepsilon,j}$ for some *i, j*. Note that the alternative hypothesis does not assume any particular model of heteroscedasticity, which makes the White test compatible with the nonparametric approach. Postulating a more specific alternative hypothesis can increase the power of the test. However, the White test provides a useful starting point for more explicit modeling of heteroscedasticity.

The White test can be built upon the OLS regression of the following equation:[26]

$$(\hat{\varepsilon}_i^{CNLS})^2 = \alpha + \sum_{j=1}^{m} \beta_j x_{ij} + \frac{1}{2} \sum_{j=1}^{m} \sum_{h=1}^{j} \gamma_j x_{ij} x_{ih} + \varepsilon_i. \tag{7.47}$$

In words, we explain the squared CNLS residual by a constant, all *m* input variables, and their squared values and cross-products using a flexible quadratic functional form as an approximation of the true but unknown heteroscedasticity effects. The test statistic is

$$W = nR^2,$$

where $R^2$ is the coefficient of determination of the OLS regression of Eq. (7.47). Under the null hypothesis of homoscedasticity, the test statistic $W$ follows the $\chi^2(J)$ distribution with $J$ degrees of freedom, where $J = 1+m+m(m+1)/2$ is the number of $\alpha, \beta, \gamma$ parameters on the right hand side of Eq. (7.47). If the value of test statistic $W$ falls below the critical value of $\chi^2(J)$ at the given level of significance (note: the usual significance levels considered are 5 and 1 %), then the null hypothesis of homoscedasticity is maintained. In that case, the test result provides some additional reassurance that the original model is well specified. On the other hand, if the value of test statistic $W$ exceeds the critical value of $\chi^2(J)$ at the given level of significance, then the null hypothesis is rejected, and hence explicit modeling of heteroscedasticity is needed.

[26] In econometrics, heteroscedasticity is usually modeled as a function of explanatory variables (i.e., inputs **x**). In contrast, the SFA literature usually models heteroscedasticity as a function of **z**-variables that may contain some (or all) of the inputs **x**. For clarity, in this section we follow the econometric convention and focus on heteroscedasticity with respect to inputs **x** and discuss the additional **z**-variables below.

The White test is usually presented in terms of the regressors of the original regression model (i.e., in terms of inputs **x** in the present context). Note that we are mainly concerned about possible heteroscedasticity with respect to inputs, which would cause bias in StoNED estimation. If we are interested in heteroscedasticity with respect to contextual variables **z**, we can also introduce the z-variables to the regression Eq. (7.47). We only need to adjust the degrees of freedom $J$ to include the number of additional parameters for the z-variables, otherwise the test procedure is conducted as described above.

If significant heteroscedasticity is found, the White test does not indicate whether heteroscedasticity is in the inefficiency term or the noise term, or possibly both. To our knowledge, general diagnostic testing of whether heteroscedasticity is in the inefficiency or noise term has attracted little attention in the SFA literature. The doubly-heteroscedastic model (following Hadri 1999; and Wang 2002), to be examined in detail in the next sub-section, does allow us model heteroscedasticity in both inefficiency and noise terms, and also test for significance of the parameter estimates. However, such specification tests are conditional on the assumed model of heteroscedasticity, including the parametric distributional assumptions regarding inefficiency and noise. An appealing feature of the White test is it does not assume any specific model of heteroscedasticity and it does not depend on the distributional assumptions. Further, the parameter estimates of the auxiliary regression (7.47) and the associated diagnostic tools can provide some insights on which specific inputs (or contextual variables) are most likely causes of heteroscedasticity, and whether heteroscedasticity effect appears to be linear or non-linear, and whether the interaction terms (cross-products) are significant. These insights can be useful for specifying parametric models of heteroscedasticity, to be considered in the next sub-section.

Before proceeding, note that quantile estimation (see Sect. 7.6.4) could provide a promising nonparametric route for testing heteroscedasticity. If the composite error term is homoscedastic, then the $q$-quantiles should have approximately same shape for different values of parameter $q$. Provided that the number of input (and output) variables is sufficiently small, plotting the estimated $q$-quantiles at different values of $q$ allow one to visually inspect whether homoscedasticity holds by a reasonable approximation. If homoscedasticity is violated, the $q$-quantile plots can help one to identify in which part of the frontier heteroscedasticity occurs, and which inputs are likely sources of heteroscedasticity. In the context of linear quantile regression, Koenker and Bassett (1982) propose formal tests of heteroscedasticity based on the comparison of the estimated $q$-quantiles at different values of $q$. Newey and Powell (1987) apply a similar idea for the $q$-expectiles, noting that the $q$-expectiles could also be used for testing symmetry of the composite error term (i.e., whether the asymmetric inefficiency term $u$ is significant; compare with Sect. 7.5.5). Adapting these tests to the nonparametric CNQR method for estimating $q$-quantiles and the CAWLS method for estimating $q$-expectiles introduced in Sect. 7.6.4 provides an interesting challenge for future research further discussed in Sect. 7.9.

### 7.8.2 Doubly-Heteroscedastic Model

If the White test indicates significant heteroscedasticity, it is difficult to tell *a priori* whether heteroscedasticity is due to the inefficiency term, the noise term, or possibly both. Therefore, we will consider the general doubly-heteroscedastic model where both the inefficiency and noise term can be heteroscedastic. The doubly-heteroscedastic model was first considered by Hadri (1999). Our formulation below is mainly based on Wang (2002) and Kumbhakar and Sun (2013).

Consider the unified model described in Sect. 7.2. In this section we assume the inefficiency term has a truncated normal distribution and the noise term is normally distributed according to

$$u_i \sim N^+(\mu_i, \sigma^2_{u,i})$$
$$v_i \sim N(0, \sigma^2_{v,i})$$

The pre-truncation mean of the inefficiency term is assumed to be a linear function of inputs:

$$\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i.$$

The pre-truncation standard deviation of the inefficiency term and the standard deviation of the noise term are specified as

$$\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$$
$$\sigma_{v,i} = \exp(\alpha_2 + \rho'\mathbf{x}_i)$$

Note that the exponent functions are commonly used in this context to guarantee that the standard deviations are positive at all input levels. While the specific parametric assumption may appear arbitrary, this model is one of the most flexible and general parametric specifications of heteroscedasticity. Note that the truncated normal distribution where both the pre-truncation mean and variance depend on the input level allows that the location (mean) and the shape (variance)of the inefficiency distribution can change as a function of inputs.

This formulation of heteroscedastic inefficiency term implies that the expected value of inefficiency can be stated as (see Wang 2002; Kumbhakar and Sun 2013)

$$E(u_i \,|u_i > 0) = \sigma_{u,i}\left[\Lambda_i + \frac{\phi(\Lambda_i)}{\Phi(\Lambda_i)}\right], \tag{7.48}$$

where

$$\Lambda_i = \frac{\mu_i}{\sigma_{u,i}}$$

and $\phi$ and $\Phi$ are the density function and the cumulative distribution function of the standard normal $N(0,1)$ distribution, respectively. The expected inefficiency is no longer a constant, but its dependence on inputs $\mathbf{x}$ has a well-defined functional form conditional on the parametric assumptions stated above. This allows us to both estimate the heteroscedasticity effects empirically, and take heteroscedasticity explicitly into account in the StoNED procedure.

### 7.8.3 Stepwise StoNED Estimation Under Heteroscedasticity

To estimate the doubly-heteroskedastic model, we can adjust the stepwise StoNED routine presented in Sect. 7.5 as follows (a more detailed elaboration of each step follows below):

**Step 1** Apply the CNLS estimator (7.3) to estimate the conditional mean output $\hat{g}^{CNLS}(\mathbf{x}_i) = E(y_i \,|\mathbf{x}_i)$.

**Step 2** Apply quasi-likelihood estimation to the CNLS residuals $\varepsilon_i^{CNLS}$ to estimate the parameters of $\mu_i$, $\sigma_{u,i}$, and $\sigma_{v,i}$.

**Step 3** Adjust the conditional mean function by adding the expected inefficiency $E(u_i \,|\mathbf{x}_i, \hat{\mu}_i\ , \hat{\sigma}_{u,i})$ to estimate the frontier for the observed data points using

$$\hat{f}^{StoNED}(\mathbf{x}_i) = \hat{g}^{CNLS}(\mathbf{x}_i) + E(u_i \,|\mathbf{x}_i, \hat{\mu}_i\ , \hat{\sigma}_{u,i}).$$

Then apply Eq. (7.5) to estimate the frontier $\hat{f}_{\min}^{StoNED}(\mathbf{x})$ for unobserved points.

**Step 4** Apply JLMS method to estimate firm-specific inefficiency using the conditional mean $E(u_i \,|\hat{\varepsilon}_i^{CNLS})$.

In step 1, we estimate the conditional mean function $g(\mathbf{x})$. The CNLS estimator remains unbiased and consistent estimator of the conditional mean $g$, despite heteroscedastic composite error term (similar to OLS). However, note that in the case of the doubly-heteroscedastic model

$$g(\mathbf{x}_i) = E(y_i \,|\mathbf{x}_i) = f(\mathbf{x}_i) - E(u_i \,|\mathbf{x}_i).$$

Note that the shape of function $g$ can differ from that of frontier $f$ because $E(u_i \,|\mathbf{x}_i)$ is a function of inputs $\mathbf{x}$. We will take this into account in step 3 where we shift function $g$ upward, not by a constant $\mu$, but rather, by the estimated $E(u_i \,|\mathbf{x}_i)$.[27] It is also worth noting that function $g$ is not necessarily monotonic increasing and concave even if the production function $f$ satisfies these axioms because $-E(u_i \,|\mathbf{x}_i)$ can be a non-monotonic and non-concave function of inputs (note: there does exist parameter values for which $-E(u_i \,|\mathbf{x}_i)$ is indeed monotonic and concave in the domain of non-negative $\mathbf{x}$). To apply CNLS in step 1, we need to assume that the curvature of the production function $f$ dominates and that function $g$ is monotonic increasing and concave (at least by approximation). Even if one assumes that $f$ exhibits CRS, it is recommended to apply the VRS specification in step 1 to allow for the nonlinear effects of $E(u_i \,|\mathbf{x}_i)$, and impose CRS later in step 3.

[27] In the context of SFA, Kumbhakar and Lovell (2000) state strongly that the stepwise MOLS procedure cannot be used in the case of heteroscedastic inefficiency. They correctly note that the OLS estimator used in the first step yields biased estimates of not only the intercept but also the slope coefficients of the frontier. However, Kumbhakar and Lovell seem to overlook the possibility of eliminating the bias by shifting function $g$ upward by a conditional expectation of inefficiency that depends on inputs $\mathbf{x}$.

Having estimated the parameters of the inefficiency and noise terms, it is possible to test if monotonicity and concavity assumptions of $g$ hold. If $g$ does not satisfy monotonicity and concavity, we can substitute CNLS by techniques depending on which axiom does not hold. Specifically, if the concavity assumption is violated, it is possible to apply isotonic nonparametric least squares (INLS) suggested by Keshvari and Kuosmanen (2013). Another possibility is to estimate order-$q$ quantile frontier using either CNQR or CAWLS techniques introduced in Sect. 7.6.4. Specifying the correct value for $q$ will ensure that the quantile frontier inherits the monotonicity and concavity properties of frontier $f$ even if the heteroscedastic inefficiency term is a non-monotonic or non-convex function of inputs. Indeed, we do not insist on estimating the conditional mean in step 1, the conditional quantile is equally suitable.

In step 2 it is natural to resort to the pseudolikelihood method since we utilize a rather heavily parametrized model of heteroscedasticity. As already noted in Sect. 7.5, a simple practical trick to conduct quasi-likelihood estimation is to use the standard ML algorithms available for SFA in standard software packages (e.g., Stata, Limdep, or R). In this case we specify the CNLS residuals $\hat{\varepsilon}_i^{CNLS}$ as the dependent variable (i.e., the output) and a constant term as an explanatory variable (input), and the ML algorithm performs the quasilikelihood estimation. For example, the frontier modeling tools of Stata allows one to include "explanatory variables for technical inefficiency variance function (uhet)" and "explanatory variables for idiosyncratic error variance function (vhet)" if the distribution of inefficiency term is specified as half-normal or exponential. It is also possible to include covariates to the truncated normal specification of the inefficiency term, but in this specification the noise term is assumed to be homoscedastic. Hung-Jen Wang has developed a Stata package for the model described in Wang (2002), which can be used for estimating the model estimating the heteroscedasticity model described above.[28]

Having estimated the underlying parameters of $\mu_i, \sigma_{u,i}, \sigma_{v,i}$, it is recommended to apply standard specification tests available for ML (i.e., likelihood-ratio, Lagrange multiplier, or Wald test) to test restrictions $\boldsymbol{\beta} = \mathbf{0}, \boldsymbol{\gamma} = \mathbf{0}$, and $\boldsymbol{\rho} = \mathbf{0}$. For example, if the null hypothesis of $\boldsymbol{\rho} = \mathbf{0}$ is not rejected, then the assumption of homoscedastic noise term can be maintained. Similarly, if $\alpha_0 = 0, \boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\gamma} = \mathbf{0}$, then the model of heteroscedastic truncated normal inefficiency term reduces to a homoscedastic half-normal inefficiency term. If the specification tests provide evidence that some of the heteroscedasticity effects are not significant, we would recommend excluding those effects from the heteroscedasticity model and estimating step 2 again.

One additional issue is in the context of linear regressionthat efficiency of the least squares estimator can be improved by applying weighted least squares or generalized least squares. Having estimated the firm specific $\sigma_{u,i}, \sigma_{v,i}$, it is possible to return back to step 1 and apply a weighted version of the CNLS estimator. Defining $\hat{\sigma}^2_{\varepsilon,i}$

[28] The Stata package is available from Wang's homepage: http://homepage.ntu.edu.tw/~wangh/.

$= \hat{\sigma}_{u,i}^2 + \hat{\sigma}_{v,i}^2$, we can modify the objective function of the CNLS problem as

$$\min \sum_{i=1}^{n} \frac{(\varepsilon_i^{CNLS})^2}{\hat{\sigma}_{\varepsilon,i}^2}$$

maintaining the original constraints of (7.3). Interpreting the given $1/\hat{\sigma}_{\varepsilon,i}^2$ as firm-specific weights, this weighted least squares formulation of CNLS is directly analogous to the generalized least squares (GLS) estimator of the linear regression model.[29] However, as yet there is no evidence that the use of weighted least squares can improve efficiency of the CNLS estimator. Intuitively, the direct analogue with GLS would suggest that weighted least squares can be more efficient than the unweighted CNLS under heteroscedasticity. On the other hand, recall that CNLS approximates the underlying function $g$ by a piece-wise linear curve. Since the hyperplane segments of the unweighted CNLS formulation provide local approximation, assigning larger or smaller weights to certain regions of the frontier may not have much effect on the piece-wise linear approximation. In our limited experience, introducing the weights $1/\hat{\sigma}_{\varepsilon,i}^2$ does not necessarily have any notable impact on the results. Further, we need to be able to estimate $\sigma_{\varepsilon,i}^2$ with a sufficient precision. Overall, we are somewhat skeptical whether the possible benefit in terms of improved efficiency of the CNLS estimator can outweigh the cost of additional effort of conducting the weighted least squares estimation. This forms an interesting open question for future research.

In step 3 we adjust the conditional mean function $g$ estimated in step 1 (or alternatively, the conditional $q$-quantile) for the estimated expected inefficiency to estimate the frontier $f$. Note that the conditional mean $E(u_i \,|\mathbf{x}_i)$ is no longer a constant, but a function that depends on inputs $\mathbf{x}$. Using Eq. (7.48), we can write the estimated expected inefficiency as the function of inputs and parameter estimates as

$$E(u_i \,|\mathbf{x}_i, \hat{\mu}_i\,, \hat{\sigma}_{u,i}) = \hat{\mu}_i + \hat{\sigma}_{u,i} \frac{\phi(\hat{\Lambda}_i)}{\Phi(\hat{\Lambda}_i)}$$

$$= (\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i) + \exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i) \left[ \phi\left( \frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)} \right) / \Phi\left( \frac{\hat{\alpha}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}_i}{\exp(\hat{\alpha}_1 + \hat{\boldsymbol{\gamma}}'\mathbf{x}_i)} \right) \right]$$

This expression reveals that in the doubly-heteroscedastic model the expected value of inefficiency has a linear part originating from the mean $\mu_i = \alpha_0 + \boldsymbol{\beta}'\mathbf{x}_i$, and a nonlinear part driven by $\sigma_{u,i} = \exp(\alpha_1 + \boldsymbol{\gamma}'\mathbf{x}_i)$. Having estimated the parameters of the inefficiency term, it is useful to evaluate whether $-E(\hat{u}_i \,|\mathbf{x}_i)$ is monotonically increasing and concave within the observed range of inputs (e.g., plot the values of $-E(\hat{u} \,|\mathbf{x})$ at different levels of $\mathbf{x}$ to visually inspect possible violations of monotonicity and concavity). To ensure that the estimated frontier function satisfies the

[29] Note that in the CNLS context we prefer to introduce weights to the objective function instead of applying variable transformations (as in GLS) because the monotonicity and concavity constraints must hold for the original input variables $\mathbf{x}$.

postulated axioms despite minor violations of monotonicity and concavity (which may be just artifacts of the arbitrary parametric specification of the heteroscedasticity model), we apply the minimum extrapolation principle and utilize the DEA method stated in Eq. (7.5) to obtain the convex monotonic hull of the fitted values $\hat{f}^{StoNED}(\mathbf{x}_i)$ of observations $i = 1,\ldots, n$, which yields the frontier estimator $\hat{f}^{StoNED}_{\min}(\mathbf{x})$.

In step 4, we can compute firm specific inefficiency estimates using the JLMS conditional mean $E(u_i \,|\hat{\varepsilon}_i^{CNLS})$ using the firm specific parameter estimates $\hat{\mu}_i, \hat{\sigma}_{u,i}, \hat{\sigma}_{v,i}$. Note that the expected inefficiency $E(u_i \,|\mathbf{x}_i, \hat{\mu}_i, \hat{\sigma}_{u,i})$ applied for shifting the conditional mean function $g$ to estimate frontier $f$ does not depend on the heteroscedasticity of the noise term. However, the JLMS efficiency does also depend on the heteroscedasticity of the noise term $\hat{\sigma}_{v,i}$. Kumbhakar and Sun (2013) discuss this issue in more detail, showing that the marginal effect of inputs on the conditional JLMS efficiency also depend on the heteroscedasticity of the noise term.

## 7.9 Directions for Future Research

This chapter has provided an updated and elaborated presentation of the CNLS and StoNED methods. Bridging the gap between the established DEA and SFA paradigms, these methods represent a major paradigm shift towards a unified and integrated methodology of frontier estimation and efficiency analysis that has a considerably broader scope than the conventional DEA and SFA tools. This chapter did not only review previously published method developments and their extensions, but also presented some new innovations, including the first extension of the StoNED method to the general case of multiple inputs and multiple outputs, and the first detailed examination of how heteroscedastic inefficiency and noise terms can be modeled within the CNLS and StoNED estimation frameworks.

We see CNLS and StoNED not only as the state of the art in axiomatic nonparametric frontier estimation and efficiency analysis under stochastic noise, but also a promising way forward. Kuosmanen and Kortelainen (2012) stated explicitly 12 promising avenues of future research on the StoNED methodology. In the following we will provide an updated version of a 12 point research program, indicating the work that has already been done as well as work that remains to be done.

1. Adapting the known econometric and statistical methods for dealing with heteroskedasticity, endogeneity, sample selection, and other potential sources of bias, to the context of CNLS and StoNED estimators.

In this chapter we presented the first detailed examination about the modeling of heteroscedasticity in the inefficiency and noise terms. Kuosmanen et al. (2013) examine the endogeneity problem from a novel perspective employing directional distance functions. Obviously, a lot of further work is needed in this area. Alternative models of heteroscedasticity as well as estimation techniques deserve careful attention. The convex nonparametric quantile regression and the convex asymmetrically weighted

least squares methods discussed in Sect. 7.6.4 and the generalized least squares estimator discussed in Sect. 7.8.3 provide potential methods for modeling and testing heteroskedasticity. The use of instrumental variables in CNLS for modeling measurement errors, sample selection, and other types of endogeneity bias should be investigated.

2. Extending the proposed approach to a multiple output setting.

In this chapter we also presented the first extension of the StoNED method to the general case of multiple inputs and multiple outputs using the directional distance function (see also Kuosmanen et al. 2013). Further work is also needed in this area. Alternative representations of the joint production technology, including the radial input and output distance functions, should be investigated. The main challenge in modeling joint production is not the formulation of the mathematical programming problem for the CNLS estimator (the usual DEA problem) or deconvoluting the composite error term (the usual SFA problem). The main challenge is the probabilistic modeling of the data generating process in the case of joint production, involving multiple endogenous inputs and outputs. Kuosmanen et al. (2013) provides a useful starting point in this respect.

3. Extending the proposed approach to account for relaxed concavity assumptions (e.g., quasiconcavity).

Keshvari and Kuosmanen (2013) presented the first extension in this direction, applying isotonic regression that relaxes the concavity assumption of CNLS. This approach estimates a step function analogous to free disposable hull (FDH) in the middle of the data cloud. The insights of Keshvari and Kuosmanen could be useful for examining the intermediate cases between the non-convex step function and the fully convex CNLS, allowing one to postulate quasiconcavity or quasiconvexity in terms of some variables (e.g., inputs, or input prices in the estimation of the cost function). Many opportunities for future research exist in this direction.

4. Developing more efficient computational algorithms or heuristics for solving the CNLS problem.

Lee et al. (2013) is the first contribution in this direction. The algorithm developed in that paper first solves a relaxed CNLS problem containing an initial set of constraints, those that are likely to be binding, and then iteratively adds a subset of the violated concavity constraints until a solution that does not violate any constraint is found. We believe the computational efficiency can be improved considerably by clever algorithms and heuristics (see, e.g., Hannah and Dunson 2013). This is an important avenue for future research in the era of "big data".

5. Examining the statistical properties of the CNLS estimator, especially in the multivariate case.

Seijo and Sen (2011) and Lim and Glynn (2012) were the first to address this challenge, proving statistical consistency of the CNLS estimator in the general multivariate case under slightly different assumptions about the data generating process. Further research on both the finite sample properties (e.g., unbiasedness or bias,

efficiency, mean squared error) and the asymptotic properties (e.g., rates of convergence, limiting distributions) under different assumption of the data generating process would be needed. In this respect, Groeneboom et al. (2001a, 2001b) provide an excellent starting point. The statistical properties of the convex nonparametric quantile regression (CNQR) and the convex asymmetrically weighted least squares (CAWLS) methods introduced in Sect. 7.6.4 also deserve further research.

6. Investigating the axiomatic foundation of the CNLS and StoNED estimators.

CNLS regression builds upon the same axioms as DEA, and StoNED estimation applies the minimum extrapolation principle to obtain a unique frontier function that satisfies the postulated axioms. However, it would be compelling if the technology characterized by CNLS and/or StoNED could be stated rigorously from the axiomatic point of view as the intersection of all sets that satisfy the stated axioms and satisfy axiom X. It remains unknown whether axiom X exists, and how it could be formulated explicitly.

7. Implementing alternative distributional assumptions and estimating the distribution of the inefficiency term by semi- or nonparametric methods in the cross-sectional setting.

In this chapter (Sect. 7.5.2) we have provided an extensive review of possibilities, including parametric and semi-parametric alternatives. In principle, the quasilikelihood method is applicable to any parametric specification of inefficiency distribution. The most promising way forward seems to be the nonparametric kernel deconvolution of the CNLS residuals, following the works by Hall and Simar (2002) and Horrace and Parmeter (2011). One challenge that remains is to adapt the JLMS conditional mean inefficiency to the semi-parametric setting where no parametric distribution is specified for the inefficiency term.

8. Distinguishing time-invariant inefficiency from heterogeneity across firms, and identifying inter-temporal frontier shifts and catching up in panel data models.

Kuosmanen and Kortelainen (2012) present a simple fixed effects approach to modeling panel data, assuming time-invariant inefficiency. In this chapter we considered the parallel random effects approach, following Eskelinen and Kuosmanen (2013). Ample opportunities for extending these basic techniques to more sophisticated semi-parametric models allowing for technical progress and time-varying inefficiency are available. Indeed, panel data models have been extensively studied both in general econometrics and in the SFA literature. Both the insights and practical solutions from panel data econometrics can be imported to the CNLS and StoNED framework.

9. Extending the proposed approach to the estimation of cost, revenue, and profit functions as well as to distance functions.

Kuosmanen and Kortelainen (2012) consider the estimation of cost function in the single output case under CRS. They made these restrictive assumptions because the cost function must be a concave function of input prices. However, if the standard convexity axiom of the production possibility set holds, then the cost function is a convex function of outputs. A challenge that remains is to formulate the CNLS problem such that we can estimate a function that is convex in one subset of variables (i.e.,

outputs), but concave in another subset of variables (i.e., input prices). Kuosmanen (2012) estimates a multi-output cost function using StoNED, but the input prices were excluded by assuming that all firms take the same input prices as given.

10. Developing a consistent bootstrap algorithm and/or other statistical inference methods.

An earlier version of Kuosmanen and Kortelainen (2012) proposed to adapt the parametric bootstrap method proposed by Simar and Wilson (2010) for drawing statistical inferences in the StoNED setting. However, the anonymous reviewers were not convinced that the proposed boostrap method is necessarily consistent when applied to the CNLS residuals. Indeed, one should be wary of naïve bootstrap and resampling approaches that produce invalid and misleading results. Since Kuosmanen and Kortelainen were not able to prove consistency of Simar and Wilson's bootstrap procedure in the CNLS case, the suggestion was excluded from the published version. We stress that adapting one of the known variants of the bootstrap method to the context of CNLS and StoNED would be straightforward. The challenge is to prove that the chosen version of bootstrap method is consistent under the stated assumptions about the data generating process. Another promising approach is to test if CNLS estimates differ significantly from the corresponding estimates obtained using parametric methods (see Sen and Meyer 2013). As for the contextual variables, Johnson and Kuosmanen (2012) prove that conventional inference techniques from linear regression analysis (e.g., t-tests, p-values, confidence intervals) can be applied for the parametric part (i.e., the coefficients of the contextual variables).

11. Conducting further Monte Carlo simulations to examine the performance of the proposed estimators under a wider range of conditions, and comparing the performance with other semi- and nonparametric frontier estimators.

Several published studies provide Monte Carlo evidence on the finite sample performance of CNLS and StoNED estimators. Kuosmanen (2008) and Kuosmanen and Kortelainen (2012) provide the first simulation results for CNLS and StoNED, respectively, focusing on the precision in estimating the frontier production function $f$. Johnson and Kuosmanen (2011) present MC simulations regarding the estimation of the parametric $\delta$ representing the effect of a single contextual variable $z$ that may be correlated with input $x$. Andor and Hesse (2014) provide an extensive comparison of the performances of DEA, SFA, and StoNED, mainly focusing on the estimation of the firm specific inefficiency $u_i$. However, note that all estimators considered are inconsistent in the noisy setting considered because $u_i$ is just a single realization of a random variable. Kuosmanen et al. (2013) compare performances of DEA, SFA and StoNED in terms of estimating a frontier cost function. They calibrate their simulations to match the empirical characteristics of the Finnish electricity distribution firms. Their simulations demonstrate that if the premises stated by the Finnish energy regulator hold, then the StoNED estimator has superior performance compared to its restricted special cases, DEA and SFA. As for further research, it would be interesting to compare performance of CNLS and StoNED with those of other semi- and nonparametric frontier estimation techniques such as kernel regression and local maximum likelihood.

12. Applying the proposed method to empirical data, and adapting the method to better serve the needs of specific empirical applications.

The first published application of the StoNED method was Kuosmanen and Kuosmanen (2009), who estimated the production function from the data of 332 Finnish dairy farms in order to assess sustainability performance of farms. Subsequently, there have been several applications in the energy sector, both in production and distribution of electricity. Mekaroonreung and Johnson (2012) applied StoNED to estimate the shadow prices of SO2 and NOx from the data of U.S. coal-fired power plants. Thus far, the most significant real-world application of StoNED has been the study by Kuosmanen (2012) [see also Kuosmanen et al. (2013), Dai and Kuosmanen (2014), and Saastamoinen and Kuosmanen (2014)]. Based on the results of this study, the Finnish energy market regulator adopted the StoNED method in systematic use in the regulation of the Finnish electricity distribution industry, with the total annual turnover of more than € 2 Billion. Another real-world application of StoNED is Eskelinen and Kuosmanen (2013), who assessed inter-temporal performance of sales teams using monthly data of Helsinki OP-Pohjola Bank, in close collaboration with the central management of the bank. The results and insights gained in this study were communicated to the team managers and were utilized for setting performance targets for sales teams. These empirical applications illustrate the flexibility and adaptability of the StoNED methodology to suit the specific needs of the application. The applications also provide motivation for developing further methodological extensions to meet the requirements of future applications.

In conclusion, we hope the 12-point program discussed above might inspire future methodological research along the lines described or along new avenues that have escaped our attention. We also hope that the methodological tools currently available would find inroads to empirical applications. In our experience from both Monte Carlo simulations and real empirical applications, CNLS and StoNED has proved dependable, reliable and robust, with an ability to produce results and insights that could not be found using the conventional methods.

## References

Afriat SN (1967) The construction of a utility function from expenditure data. Int Econ Rev 8:67–77

Afriat SN (1972) Efficiency estimation of production functions. Int Econ Rev 13(3):568–598

Aigner D, Chu S (1968) On estimating the industry production function. Amer Econ Rev 58:826–839

Aigner D, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econom 6:21–37

Almanidis P, Sickles RC (2012) The skewness issue in stochastic frontier models: fact of fiction? In: Keilegom I van, Wilson PW (eds) Exploring research frontiers in contemporary statistics and econometrics. Springer Verlag, Berlin

Alminidis P, Qian J, Sickles R (2009). Stochastic frontiers with bounded inefficiency, mimeo, Rice University

Andor M, Hesse F (2014) The StoNED age: the departure into a new era of efficiency analysis? –a Monte Carlo comparison of StoNED and the "Oldies" (SFA and DEA). J Productiv Anal 41(1):85–109

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

Banker RD (1993) Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. Manag Sci 39:1265–1273

Banker RD, Datar S, Kemerer C (1991) A Model to Evaluate Variables Impacting the Productivity of Software Maintenance Projects. Management Science 37(1):1–18

Banker RD, Maindiratta A (1986) Piece-wise loglinear estimation of efficient production surfaces. Manag Sci 32(1):126–135

Banker RD, Maindiratta A (1992) Maximum likelihood estimation of monotone and concave production frontiers. J Product Anal 3:401–415

Banker RD, Natarajan R (2008) Evaluating contextual variables affecting productivity using data envelopment analysis. Oper Res 56(1):48–58

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 30(9):1078–1092

Battese GE, Coelli TJ (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. Empir Econ 20(2):325–332

Carree MA (2002) Technological inefficiency and the skewness of the error component in stochastic frontier analysis. Econ Lett 77:101–107

Caudill S, Ford J (1993) Biases in frontier estimation due to heteroscedasticity. Econ Lett 41(1): 17–20

Caudill S, Ford J, Gropper D (1995) Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity. J Bus Econ Statist 13(1):105–111

Chambers RG, Chung YH, Färe R (1996) Benefit and distance functions. J Econ Theory 70(2): 407–419

Chambers RG, Chung and Y, Färe R (1998) Profit, distance functions and Nerlovian efficiency. J Optim Theory Appl 98:351–364

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Seiford L, Stutz J (1982) A multiplicative model for efficiency analysis. Socio-Econ Plan Sci 16:223–224

Chen X (2007). Large sample sieve estimation of semi-nonparametric models. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6. Elsevier, North Holland

Cobb CW, Douglas PH (1928) A theory of production. Amer Econ Rev 18:139–165

Coelli T (1995) Estimators and hypothesis tests for a stochastic frontier function: a Monte Carlo analysis. J Product Anal 6:247–268

Coelli T, Perelman S (1999) A comparison of parametric and non-parametric distance functions: With application to European railways. Eur J Oper Res 117(2):326–339

Coelli T, Perelman S (2000) Technical efficiency of European railways: a distance function approach. Appl Econ 32(15):1967–1976

D'Agostino R, Pearson ES (1973). Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{b_1}$. Biometrika 60(3):613–622

Dai X, Kuosmanen T (2014) Best-practice benchmarking using clustering methods: application to energy regulation. Omega 42(1):179–188

Dantzig GB, Fulkerson DR, Johnson SM (1954) Solution of a large-scale traveling salesman problem. Oper Res 2:393–410

Dantzig GB, Fulkerson DR, Johnson SM (1959) On a linear-programming combinatorial approach to the traveling-salesman problem. Oper Res 7:58–66

Daouia A, Simar L (2007) Nonparametric efficiency analysis: a multivariate conditional quantile approach. J Econom 140:375–400

Eskelinen J, Kuosmanen T (2013) Intertemporal efficiency analysis of sales teams of a bank: stochastic semi-nonparametric approach. J Bank Financ 37(12):5163–5175
Fan J (1991) On the optimal rates of convergence for nonparametric deconvolution problems. Ann Stat 19:1257–1272
Fan Y, Li Q, Weersink A (1996) Semiparametric estimation of stochastic production frontier models. J Bus Econ Stat 14:460–468
Farrell MJ (1957) The measurement of productive efficiency. J Royal Stat Soc Ser A 120:253–281
Färe R, Grosskopf S, Norris M, Zhang Z (1994) Productivity growth, technical progress, and efficiency change in industrialized countries. Amer Econ Rev 84(1):66–83
Färe R, Grosskopf S, Noh D-W, Weber W (2005) Characteristics of a polluting technology: theory and practice. J Econom 126:469–492
Gabrielsen A (1975). On estimating efficient production functions. Working Paper No. A-85, Chr. Michelsen Institute, Department of Humanities and Social Sciences, Bergen, Norway
Greene WH (1980) Maximum likelihood estimation of econometric frontier functions. J Econom 13:26–57
Greene WH (2008) The econometric approach to efficiency analysis. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity growth. Oxford University Press Inc., New York, pp 92–250
Groeneboom P, Jongbloed G, Wellner JA (2001a) A canonical process for estimation of convex functions: the "Invelope" of integrated brownian motion +t4. Ann Stat 29:1620–1652
Groeneboom P, Jongbloed G, Wellner JA (2001b) Estimation of a convex function: characterizations and asymptotic theory. Ann Stat 29:1653–1698
Gstach D (1998) Another approach to data envelopment analysis in noisy environments: DEA+. J Product Anal 9(2):161–176
Hadri K (1999) Estimation of a doubly heteroscedastic stochastic frontier cost function. J Bus Econ Statist 17(3):359–363
Hall P, Simar L (2002) Estimating a changepoint, boundary, or frontier in the presence of observation error. J Amer Stat Assoc 97:523–534
Hannah LA, Dunson DB (2013) Multivariate convex regression with adaptive partitioning. J Mach Learn Res 14:3207–3240
Hanson DL, Pledger G (1976) Consistency in concave regression. Ann Stat 4(6):1038–1050
Hildreth C (1954) Point estimates of ordinates of concave functions. J Am Stat Assoc 49:598–619
Hoff A (2007) Second stage DEA: comparison of approaches for modeling the DEA score. Eur J Oper Res 181:425–435
Horrace W, Parmeter C (2011) Semiparametric deconvolution with unknown error variance. J Product Anal 35(2):129–141
Johnson AL, Kuosmanen T (2011) One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. J Product Anal 36(2):219–230
Johnson AL, Kuosmanen T (2012) One-stage and two-stage DEA estimation of the effects of contextual variables. Eur J Oper Res 220:559–570
Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On estimation of technical inefficiency in the stochastic frontier production function model. J Econom 19:233–238
Keshvari A, Kuosmanen T (2013) Stochastic non-convex envelopment of data: applying isotonic regression to frontier estimation. Eur J Oper Res 231:481–491
Koenker R (2005). Quantile regression. Cambridge University Press. Cambridge, UK
Koenker R, Bassett GW (1978) Regression quantiles. Econometrica 46(1):33–50
Koenker R, Bassett GW (1982) Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50:43–61
Krugman P (1992) The age of diminished expectations: US economic policy in the 1980s. MIT Press, Cambridge
Kumbhakar SC, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, New York

Kumbhakar SC, Ghosh S, McGuckin JT (1991) A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. J Bus Econ Statist 9(3):279–286
Kumbhakar SC, Sun K (2013) Derivation of marginal effects of determinants of technical inefficiency. Economics Letters 120(2):249–253
Kuosmanen T (2006): Stochastic nonparametric envelopment of data: combining virtues of SFA and DEA in a unified framework, MTT Discussion Paper No. 3/2006.
Kuosmanen T (2008) Representation theorem for convex nonparametric least squares. Econom J 11:308–325
Kuosmanen T (2012) Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. Energy Econ 34:2189–2199
Kuosmanen T, Fosgerau M (2009) Neoclassical versus frontier production models? Testing for the skewness of regression residuals. Scand J Econ 111(2):351–367
Kuosmanen T, Johnson AL (2010) Data envelopment analysis as nonparametric least-squares regression. Oper Res 58:149–160
Kuosmanen T, Kortelainen M (2012) Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints. J Product Anal 38(1):11–28
Kuosmanen T, Kuosmanen N (2009) Role of benchmark technology in sustainable value analysis: an application to Finnish dairy farms. Agric Food Sci 18(3–4):302–316
Kuosmanen T, Johnson AL, Parmeter C (2013) Orthogonality conditions for identification of joint production technologies: axiomatic nonparametric approach to the estimation of stochastic distance functions, unpublished working paper (available from the authors by request).
Kuosmanen T, Saastamoinen A, Sipiläinen T (2013) What is the best practice for benchmark regulation of electricity distribution? Comparison of DEA, SFA and StoNED methods. Energy Policy 61:740–750
Lee C-Y, Johnson AL, Moreno-Centeno E, Kuosmanen T (2013) A more efficient algorithm for convex nonparametric least squares. Eur J Oper Res 227(2):391–400
Lim E, Glynn PW (2012) Consistency of multidimensional convex regression. Oper Res 60(1):196–208
Lovell CAK, Richardson S, Travers P, Wood LL (1994) Resources and functionings: a new view of inequality in Australia. In: Eichhorn W (ed) Models and measurement of welfare and inequality. Springer, Berlin, pp 787–807
Marschak J, Andrews W (1944) Random simultaneous equations and the theory of production. Econometrica 12:143–205
McDonald J (2009) Using least squares and tobit in second stage DEA efficiency analyses. Eur J Oper Res 197:792–798
Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. Int Econ Rev 18(2):435–445
Mekaroonreung M, Johnson AL (2012) Estimating the shadow prices of SO2 and NOx for U.S. coal power plants: a convex nonparametric least squares approach. Energy Econ 34(3):723–732
Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica 55(4):819–847
Ondrich J, Ruggiero J (2001) Efficiency measurement in the stochastic frontier model. Eur J Oper Res 129:434–442
Panzar JC, Willig RD (1981) Economies of scope. American Economic Review 71(2):268–272
Ruggiero J (2004) Data envelopment analysis with stochastic data. J Oper Res Soc 55(9):1008–1012
Saastamoinen A, Kuosmanen T (2014) Quality frontier of electricity distribution: supply security, best practices, and underground cabling in Finland. Energy Econ. (in press). doi:10.1016/j.eneco.2014.04.016
Schmidt P, Lin T (1984) Simple tests of alternative specifications in stochastic frontier models. J Econometric 24:349–361
Schmidt P, Sickles RC (1984) Production frontiers and panel data. J Bus Econ Stat 2(4):367–374

Seijo E, Sen B (2011) Nonparametric least squares estimation of a multivariate convex regression function. Ann Stat 39(3):1633–1657

Sen B, Meyer M (2013). Testing against a parametric regression function using ideas from shape restricted estimation, arXiv preprint arXiv:1311.6849. http://arxiv.org/pdf/1311.6849.pdf. Accessed 16 June 2014

Simar L, Wilson PW (1998) Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Manag Sci 44(1):49–61

Simar L, Wilson PW (2000) A general methodology for bootstrapping in non-parametric frontier models. J Appl Stat 27(6):779–802

Simar L, Wilson PW (2007) Estimation and inference in two-stage, semi-parametric models of production processes. J Econom 136(1):31–64

Simar L, Wilson PW (2010) Inferences from cross-sectional, stochastic frontier models. Econom Rev 29(1):62–98

Simar L, Wilson PW (2011) Two-stage DEA: Caveat emptor. J Product Anal 36(2):205–218

Timmer CP (1971) Using a probabilistic frontier production function to measure technical efficiency. J Polit Econ 79:767–794

Varian HR (1984) The nonparametric approach to production analysis. Econometrica 52:579–598

Verbeek M (2008) A guide to modern econometrics. Wiley, England

Wang H, Schmidt P (2002) One step and two step estimation of the effects of exogenous variables on technical efficiency levels. J Product Anal 18:129–144

Wang Y, Wang S, Dang C, Ge W (2014) Nonparametric quantile frontier estimation under shape restriction. Eur J Oper Res 232:671–678

White H (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4):817–838

Winsten CB (1957) Discussion on Mr. Farrell's Paper. J Royal Stat Soc Ser A 120(3):282–284

# Chapter 8
# Translation Invariance in Data Envelopment Analysis

**Jesus T. Pastor and Juan Aparicio**

**Abstract** In this chapter we present an overview of the different approaches that have considered translation invariant Data Envelopment Analysis (DEA) models. Translation invariance is a relevant property for dealing with non-positive input and/or non-positive output values. We start by considering the classical approach and continue revising recent contributions. We also consider non-translation invariant DEA models that are able to deal with negative data at the expense of modifying the model itself. Finally, we propose to study translation invariance in a general framework through a recently introduced distance function: the linear loss distance function.

**Keywords** Data envelopment analysis · Translation invariance · Negative data · Linear loss distance function

## 8.1 Introduction

Charnes et al. (1978, 1979) defined the first DEA model, the so called CCR model—or constant returns to scale (CRS) radial model-, requiring strict positivity of all the input and output values. Later on, Charnes et al. (1986) relaxed this strong requirement and, based on the ratio-form of the CCR model, showed that for each unit under scrutiny it is enough to have at least one positive input and at least one positive output. Ali and Seiford (1990) showed that resorting to the additive model (Charnes et al. 1985) is a way to deal with units with all its input and output values at level 0. In 1994 Pastor was the first to extend the latter results for dealing with negative data. His findings were published in Lovell and Pastor (1995) and in Pastor (1996). In the above-mentioned papers the three basic DEA models –the CCR, the BCC (Banker et al. 1984) and the additive- appear classified according to their translation invariant characteristics. Moreover, the result of the additive model was extended to the family of weighted additive models. Historically, prior to 1995, the

J. T. Pastor (✉) · J. Aparicio
Center of Operations Research (CIO), University Miguel Hernandez of Elche, Elche, Spain
e-mail: jtpastor@umh.es

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_8

way of dealing with negative data in a DEA framework was quite arbitrary. The easiest way was simply to delete the units with negative data from the sample. A more sophisticated way, although equally unjustified, was to perform an appropriate change of variables. Nonetheless, there are only a few manuscripts dealing with negative data prior to the paper by Lovell and Pastor (1995). References to four of them as well as their associated applications can be found in Pastor and Ruiz (2007). See also Thanassoulis et al. (2008).

During the decade starting in 1995, only four new contributions to the treatment of negative data by means of DEA models were proposed. The first one is by Cooper et al. (1999) and is known as the RAM (range adjusted measure), which is nothing other than a specific weighted additive model satisfying the translation invariant property. The second one assumes that each unrestricted in sign variable is an interval-scale variable that has a known decomposition as a sum of two ratio-scale variables; the first one non-negative and the second one non-positive (see Halme et al. 2002). The third one modifies the facets of the DEA frontier associated to a DEA radial output-oriented model if a unit with negative outputs is rated as efficient, which is judged as unacceptable (see Seiford and Zhu 2002). The last one introduces a new variable returns to scale DEA model, called range directional model (RDM), inspired by a directional distance function model that is fully translation invariant (see Silva Portela et al. 2004). All of them are described in detail and discussed in Pastor and Ruiz (2007). Nevertheless we will go back to the second and the fourth model later on.

During the period starting in 2005, at least ten more papers have tackled the problem of how to deal with negative data by means of DEA models. Let us start with the only three proposed models that are translation invariant. In 2007, Sharp et al. reformulated the objective function of the DEA model associated to the slacks based measure by Tone (2001), also known as Enhanced Russell Graph Measure (Pastor et al. 1999), obtaining a translation invariant model. In 2011, Cooper et al. published a new DEA efficiency model, known as BAM (bounded adjusted measure), which constitutes the first example of a weighted additive model with variable weights that happens to be translation invariant. Recently, Hadi-Vencheh and Esmaeilzadeh (2013) introduced a new super-efficiency model with negative data based on RDM. The rest of the papers that follow are not translation invariant.

In 2011, Kerstens and Van de Woestyne proposed considering a modified version of the proportional directional distance function instead of the RDM model, justifying this new approach through its economic appeal. In the same year, Kazemi Matin and Azizi designed a two phase approach in the same spirit as the paper by Seiford and Zhu (2002), i.e., to avoid efficient projections with negative outputs, but resorting to an additive model and not to a radial model.

The last five papers that follow are based on traditional radial measures. The most influential one, by Emrouznejad et al. (2010), defines several DEA models based on a semi-oriented radial measure (SORM) that allows the presence of unrestricted in sign variables. SORM constitutes a new less restrictive version of the model proposed by Halme et al. (2002). Besides being less restrictive it is also less reliable, as already pointed out by Jahanshahloo and Piri (2013), who proposed an extension of

SORM for accommodating not only negative values but also integer variables. A year before, Hadad et al. (2012) published an application of SORM for the Indonesian banking sector including a robustness analysis based on RDM. Kordrostami and Noveiri (2012) extended the SORM model so as to accommodate what they call flexible variables, i.e., variables that can be used at the same time as an input and as an output. Finally, Cheng et al. (2013) propose a variant of the traditional radial models, called VRM, which are nothing other than particular oriented cases of the model published two years earlier by Kerstens and Van de Woestyne (2011).

## 8.2 Translation Invariance for Dealing with Data Which Have Value Zero

As mentioned in the Introduction, Ali and Seaford (1990) were able to relax the weak requirement about non-negative data in connection with the CCR model introduced by Charnes et al. (1986) at the expense of forgetting the CCR model. They proposed considering the additive model instead of a DEA radial model, showing that it is possible to deal with units with all their inputs at level zero and/or all their outputs at level zero. The relevant feature of the additive model is that it works under a variable returns to scale (VRS) technology. They further considered an initial dataset of $n$ units to be rated. For the specific subset of units with all its inputs at level zero and/or all its outputs at level zero the corresponding additive problem could not be solved. Consequently, these authors proposed to consider a derived specific subset of units to be rated, which gave rise to a subset of specific derived additive problems. Each derived unit is obtained by translating the data of the original one by means of a fixed translation vector. For a problem with $m$ inputs and $s$ outputs they considered an $(m + s)$ translation vector with nonnegative components. It may happen that the original DEA problem and the corresponding translated problem are equivalent, i.e., both problems rate each original unit and its corresponding translated unit in exactly the same way. In this case, the model satisfies the translation invariance property or is a translation invariant model, and through translation we can transform the original dataset into a new derived one with any data greater than zero. This was exactly what happened with the envelopment form[1] of the additive model, as proved by Ali and Seiford (1990). They also considered a VRS radial model—the BCC input-oriented model (Banker et al. 1984)—obtaining less interesting results. Here is a summary of their findings.

[1] Each DEA linear program has a linear dual. Usually the primal or envelopment form evaluates each unit by measuring its "distance" with respect to the frontier, while the dual or multiplier form determines the supporting hyperplane where the unit under evaluation is projected (see Ali and Seiford 1993). Moreover, linear programming theory shows us that the objective function values of both dual programs are the same if they are finite.

**Proposition 1** The envelopment form of the additive model by Charnes et al. (1985) is translation invariant.

**Proposition 2** The BCC input-oriented model is not translation invariant. Nevertheless it is translation invariant for the subset of efficient points, i.e., an efficient unit of the original problem is also efficient for the translated problem, and conversely.

The first result establishes that the VRS additive model by Charnes et al. (1985) meets the property of translation invariance, while the second result states that although the BCC is also a VRS DEA model, it is not a translation invariant model for inputs and outputs. Nonetheless, the subset of technically efficient units of the translated model is exactly the translated subset of technically efficient units of the initial model.

## 8.3 Translation Invariant Models for Dealing with Non-Positive Data

The analysis carried out by Ali and Seiford (1990) for dealing with the presence of multiple zero values is perfectly valid for dealing with non-positive data since in both situations the clue is data displacement. Nevertheless, the most relevant contribution on translation invariance, and also on units invariance, is the paper by Lovell and Pastor (1995) (see also Pastor 1996), where the basis for achieving translation invariant models in DEA were stated. In addition, the paper shows that the unique family of models that are fully translation invariant is the so called weighted additive model. On the other hand, any CRS DEA model can never be translation invariant because its efficient frontier depends on the location of the origin of coordinates.

Overall, the property of translation invariance allows a model using the original unrestricted in sign data and the same model using the translated positive data to be equivalent, which means that both have the same optimal value and that the role-efficient unit(s) of any unit with translated data are exactly the translated role-unit(s) of the same unit based on the original data. In particular, the subset of translated efficient unit(s) corresponds to the subset of efficient unit(s) of the model with translated data.

Both papers, Ali and Seiford (1990) and Lovell and Pastor (1995), reached the same conclusion: the basic key to satisfying translation invariance in inputs and outputs is the independence of the VRS frontier from the origin of coordinates. The special configuration of the DEA technology under VRS, in contrast to CRS technology, permits that data translation does not affect the structure of the technical efficient frontier. This point can be illustrated graphically.

Suppose we have observed five units, A = (1,2), B = (2,3.5), C = (3,4), D = (2.5,1), and E = (4,2.5), each of which consumes one input to produce one output. Figure 8.1 shows the representation of the DEA VRS technology estimated through the sample of five units. Additionally, Fig. 8.1 shows a displacement of the VRS frontier obtained by substracting 6 units from each input and 5 units from each output. This translation moves the sample of units from the first to the third quadrant.

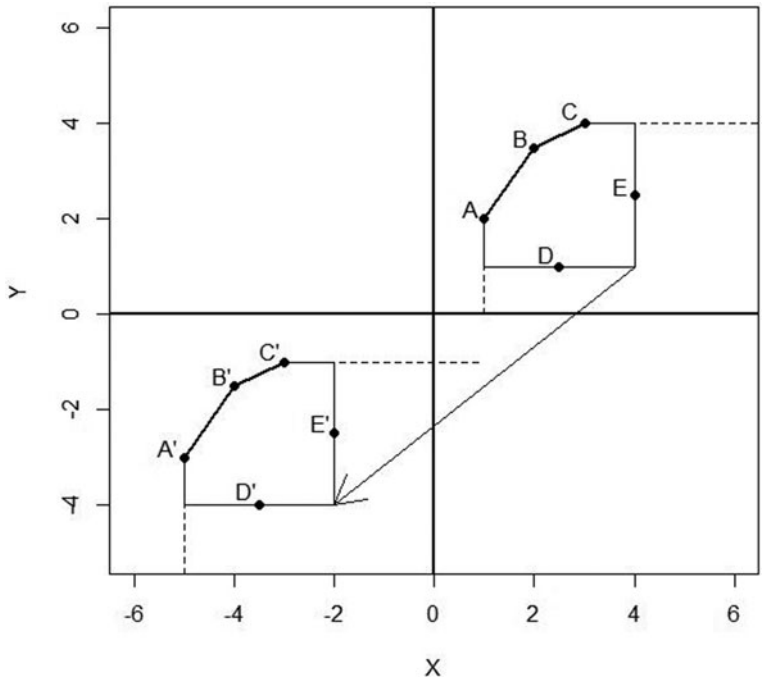



**Fig. 8.1** Graphical example for the VRS case

The translated points are identified as A', B', C', D', and E'. Similarly, we show the estimated DEA CRS frontier using the same five units in Fig. 8.2. Here we prefer to show an easier case, where the translated points, obtained by adding 5 units to the input of each observation while maintaining the output values, belong to the first quadrant.

The graphical intuitive idea behind Fig. 8.1 is that under VRS, the supporting hyperplanes containing facets of the translated efficient frontier are parallel to the original ones. As a consequence, the translated original efficient points are the efficient points of the translated model. On the contrary, in Fig. 8.2 under CRS, all the supporting hyperplanes of the technology must pass through the origin of coordinates. Therefore, any translation changes the slope of the supporting hyperplanes and, as a consequence, even the points on the efficient frontier may change. Consequently, the targets and the corresponding optimal value of each DEA problem will change accordingly which means that translation invariance is not fulfilled. In this particular case, we note that, in contrast to the VRS case, under CRS the subset of technically efficient units of the translated model, {B',C'}, does not match with the translated subset of technically efficient units of the initial model, {A'}. Consequently, the role unit(s) for each inefficient unit and its corresponding translated unit are not related.

Mathematically, the single added constraint that transforms the envelopment form of a CRS DEA model into a VRS model is $\Sigma_{j=1}^{n}\lambda_{j0} = 1$ (the sum of intensity variables equals one). For example, the well-known additive model by Charnes et al. (1985)

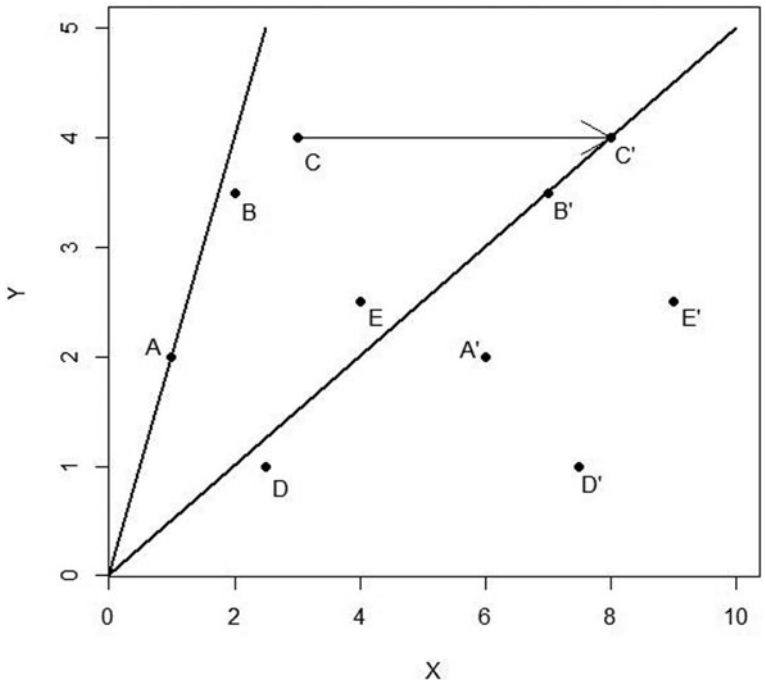



**Fig. 8.2** Graphical example for the CRS case

includes this restriction:

$$
\begin{aligned}
Max \quad & \sum_{i=1}^{m} s_{i0}^{-} + \sum_{r=1}^{s} s_{r0}^{+} \\
s.t. \quad & \\
& \sum_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - s_{i0}^{-}, \quad i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_{j0} y_{rj} = y_{r0} + s_{r0}^{+}, \quad r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& s_{i0}^{-}, s_{r0}^{+}, \lambda_{j0} \geq 0, \qquad \forall i, r, j
\end{aligned}
\tag{8.1}
$$

In this way, if we translate the original data by adding to the inputs of each unit the vector $h \in R^m$, the first subset of constraints in (8.1) would be transformed into the next one: $\Sigma_{j=1}^{n} \lambda_{j0}(x_{ij} + h_i) = (x_{i0} + h_i) - s_{i0}^{-}, i = 1, \ldots, m$, which is equivalent to $\Sigma_{j=1}^{n} \lambda_{j0} x_{ij} + h_i \Sigma_{j=1}^{n} \lambda_{j0} = x_{i0} + h_i - s_{i0}^{-}, i = 1, \ldots, m$. Since $\Sigma_{j=1}^{n} \lambda_{j0} = 1$, each of these constraints can be finally rewritten as $\Sigma_{j=1}^{n} \lambda_{j0} x_{ij} + h_i = x_{i0} + h_i - s_{i0}^{-} \Leftrightarrow \Sigma_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - s_{i0}^{-}$, which coincides with the first constraint of the model associated with the original data. The same steps may be undertaken with the second subset of constraints in (8.1) in case we translate the outputs of each unit by adding

the vector $k \in R^s$, arriving at the same conclusion. Finally, another relevant feature of (8.1) that allows the satisfaction of the translation invariance property is the fact that the objective function of the additive model remains unchanged after performing any translation. Next, we analyze the main results found by Lovell and Pastor (1995), which do not focus on the additive model, but on the weighted additive model. Each weighted additive model is defined by multiplying each slack of the objective function by a real number, which constitutes the attached weight. They constitute the family of weighted additive models with constant weights. Mathematically, the weighted additive model maximizes a weighted $\ell_1$-distance to the efficient frontier from the assessed unit rather than maximizing the $\ell_1$-distance as the additive model does. The envelopment form of a weighted additive model with weights $w_0^- \in R_+^m$ for inputs and $w_0^+ \in R_+^s$ for outputs is as follows.

$$
\begin{aligned}
Max \quad & \sum_{i=1}^{m} w_{i0}^- s_{i0}^- + \sum_{r=1}^{s} w_{r0}^+ s_{r0}^+ \\
s.t. \quad & \\
& \sum_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - s_{i0}^-, \quad i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_{j0} y_{rj} = y_{r0} + s_{r0}^+, \quad r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& s_{i0}^-, s_{r0}^+, \lambda_{j0} \geq 0, \qquad \forall i, r, j
\end{aligned}
\tag{8.2}
$$

The next proposition establishes that the family of weighted additive models with constant weights is translation invariant.

**Proposition 3** Any weighted additive model with constant weights is translation invariant.

As said before, Cooper et al. (1999) introduced another well-known weighted additive model known as RAM. It is a weighted additive model with non-constant weights. In fact, each of its weights is inversely proportional to the range of the corresponding variable. In particular, $w_{i0}^- = 1/((m+s)R_i^-)$, $i = 1, \ldots, m$, and $w_{r0}^+ = 1/((m+s)R_r^+)$, $r = 1, \ldots, s$, where $R_i^- = \max\limits_{j=1,\ldots,n} \{x_{ij}\} - \min\limits_{j=1,\ldots,n} \{x_{ij}\}$ and $R_r^+ = \max\limits_{j=1,\ldots,n} \{y_{rj}\} - \min\limits_{j=1,\ldots,n} \{y_{rj}\}$. In this case it is easy to prove that the defined weights for inputs and outputs are translation invariant and, consequently, we are able to state the following.

**Proposition 4** Any weighted additive model with non-constant weights is translation invariant if, and only if, the weights are translation invariant.

**Corollary 4.1** The RAM model is translation invariant.

We can find examples in literature of weighted additive models that are not translation invariant. This is the case of the Measure of Inefficiency Proportions (MIP) by Charnes et al. (1987). The corresponding weights are defined as $w_{i0}^- = 1/x_{i0}$, $i = 1, \ldots, m$, and $w_{r0}^+ = 1/y_{r0}$, $r = 1, \ldots, s$, which are clearly modified by any translation.

Lovell and Pastor (1995) introduced the so-called normalized weighted additive DEA model that is, by definition, a weighted additive model with the following set of weights: $w_{i0}^{-} = 1/\sigma_i^{-}$, $i = 1, \ldots, m$, and $w_{r0}^{+} = 1/\sigma_r^{+}$ , $r = 1, \ldots, s$, where $\sigma_i^{-}$ and $\sigma_r^{+}$ denote the sample standard deviation of input i-th and output r-th, respectively. The next result is a direct consequence of the last proposition.

**Corollary 4.2** The normalized weighted additive DEA model is translation invariant.

As mentioned in the introduction, the most recently defined weighted additive model with non-constant weights introduced by Cooper et al. (2011) is known as BAM. In this case, $w_{i0}^{-} = 1/((m+s)L_{i0})$, $i = 1, \ldots, m$, and $w_{r0}^{+} = 1/((m+s)U_{r0})$, $r = 1, \ldots, s$, where $L_{i0} = x_{i0} - \min\limits_{j=1,\ldots,n}\{x_{ij}\}$ and $U_{r0} = \max\limits_{j=1,\ldots,n}\{y_{rj}\} - y_{r0}$. It is easy to verify that these weights are also translation invariant. Hence, we can formulate the next corollary.

**Corollary 4.3** The BAM model is translation invariant.

On the other hand, Ali and Seiford (1990) and Lovell and Pastor (1995) also studied the traditional radial models, trying to determine relationships between the translation invariance property and the optimal value of these models. As already mentioned, Ali and Seiford (1990) established a partial result as shown above in Proposition 2. Lovell and Pastor (1995) went a step further and realized that in an oriented BCC model the corresponding efficiency score does not remain invariant when the variables—inputs or outputs—associated to the orientation are translated simply because the score is related to the position of the origin of coordinates. Nonetheless, they were able to establish that the BCC model is partially translation invariant, as the next proposition shows.

**Proposition 5** The input (output)-oriented BCC model is translation invariant with respect to output (input) changes only.

Some extensions of the last proposition considering non-discretionary variables as well as non-increasing or non-decreasing DEA models can be found in Pastor and Ruiz (2007).

Regarding other approaches that have tried to define translation invariant models in DEA, a distance function that has captured the interest of researchers is the directional distance function (DDF) (see Chambers et al. 1998, for properties of the DDF). As pointed out before, the first remarkable example is not a pure DDF model but a closely related one, known as the Range Directional Model (RDM) by Silva Portela et al. (2004). These authors proposed a specific model based on the DDF, which is able to handle datasets with negative data and to generate efficiency scores that may be readily utilized without the need to transform the data. Indeed, the generated efficiency scores have a similar interpretation as the traditional input-oriented radial measures since in the case of the RDM the origin of coordinates is substituted by the zenith point, an ideal point defined as $\left(\min\limits_{j=1,\ldots,n}\{x_{1j}\}, \ldots, \min\limits_{j=1,\ldots,n}\{x_{mj}\}, \max\limits_{j=1,\ldots,n}\{y_{1j}\}, \ldots, \max\limits_{j=1,\ldots,n}\{y_{sj}\}\right)$, and the evaluated $\text{DMU}_0$ is projected towards the efficient frontier along the direction that connects

$DMU_0$ with the zenith point. The RDM model has the following formulation.

$$
\begin{aligned}
&Max \quad \beta \\
&s.t. \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} x_{ij} \leq x_{i0} - \beta L_{i0}, \quad i = 1, \ldots, m \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq y_{r0} + \beta U_{r0}, \quad r = 1, \ldots, s, \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} = 1, \\
&\qquad \lambda_{j0} \geq 0, \qquad\qquad\qquad\qquad j = 1, \ldots, n
\end{aligned}
\tag{8.3}
$$

where $L_{i0}$, $i = 1, \ldots, m$, and $U_{r0}$, $r = 1, \ldots, s$ are defined as above. The specific definition of the directional vector guarantees that RDM is translation invariant. The proof follows the same lines as the proof of Corollary 4.1. On the other hand, an implementation of the RDM to estimate productivity change over time can be found in Silva Portela and Thanassoulis (2010), where a unique global VRS frontier is required to define both a Malmquist-type index based on the RDM efficiency measure and a Luenberger productivity indicator based on the RDM inefficiency measure. The proposed approach is applied to a sample of bank branches with unrestricted in sign data.

As with the RDM, other existing measures also have a wide set of interesting properties. For instance, the ERG (Enhanced Russell Graph) by Pastor et al. (1999), also known as the SBM (Slacks-Based Measure) by Tone (2001), is a measure that satisfies the following properties: (1) it is always between zero and one; (2) the measure is equal to one if and only if the rated DMU is Pareto-Koopmans efficient; (3) it is units invariant; and (4) it is strongly monotonic in inputs and outputs. The original formulation of the ERG is as follows:

$$
\begin{aligned}
&Min \quad \frac{\frac{1}{m} \sum_{i=1}^{m} \theta_{i0}}{\frac{1}{s} \sum_{r=1}^{s} \phi_{r0}} \\
&s.t. \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} x_{ij} = \theta_{i0} x_{i0}, \quad i = 1, \ldots, m \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} y_{rj} = \phi_{r0} y_{r0}, \quad r = 1, \ldots, s, \\
&\qquad \sum_{j=1}^{n} \lambda_{j0} = 1,
\end{aligned}
\tag{8.4}
$$

$$\lambda_{j0} \geq 0, \quad j = 1, \ldots, n$$

$$\theta_{i0} \leq 1, \quad i = 1, \ldots, m$$

$$\phi_{r0} \geq 1, \quad r = 1, \ldots, s$$

If we consider the change of variables $s_{i0}^- = x_{i0}(1 - \theta_{i0})$, $i = 1, \ldots, m$, and $s_{r0}^+ = y_{r0}(\phi_{r0} - 1)$, $r = 1, \ldots, s$, then the objective function of (4) can be rewritten as $\frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_{i0}^-}{x_{i0}}}{1 + \frac{1}{s}\sum_{i=1}^{m} \frac{s_{r0}^+}{y_{r0}}}$ and the constraints as $\Sigma_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - s_{i0}^-$, $i = 1, \ldots, m$, and $\Sigma_{j=1}^{n} \lambda_{j0} y_{rj} = y_{r0} + s_{r0}^+$, $r = 1, \ldots, s$, with $s_{i0}^- \geq 0$, $i = 1, \ldots, m$, and $s_{r0}^+ \geq 0$, $r = 1, \ldots, s$. This alternative formulation is exactly the SBM (see Tone 2001). Note also that the value of the ERG may be interpreted as the ratio between the average efficiency of inputs and the average efficiency of outputs.

On the other hand, in the case of strictly positive inputs, we have that $s_{i0}^- = x_{i0} - \Sigma_{j=1}^{n} \lambda_{j0} x_{ij} \leq x_{i0}$, since $x_{ij} > 0$ and $\lambda_{j0} \geq 0$. However, this is not necessarily the case for negative inputs, as pointed out by Sharp et al. (2007). Consequently, there is the possibility that the optimal value in model (4) is negative and, therefore, meaningless. Taking this fact into account, Sharp et al. (2007) introduced a modification of the SBM in order to define an efficiency measure capable of handling negative data. The model proposed by Sharp et al. (2007) follows.

$$Min \quad \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{w_i^- s_{i0}^-}{L_{i0}}}{1 + \frac{1}{s}\sum_{i=1}^{m} \frac{w_r^+ s_{r0}^+}{U_{r0}}}$$

$$s.t.$$

$$\begin{aligned}
&\sum_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - s_{i0}^-, && i = 1, \ldots, m \\
&\sum_{j=1}^{n} \lambda_{j0} y_{rj} = y_{r0} + s_{r0}^+, && r = 1, \ldots, s, \\
&\sum_{j=1}^{n} \lambda_{j0} = 1, && \\
&\lambda_{j0} \geq 0, && j = 1, \ldots, n \\
&s_{i0}^- \geq 0, && i = 1, \ldots, m \\
&s_{r0}^+ \geq 0, && r = 1, \ldots, s
\end{aligned} \tag{8.5}$$

where $L_{i0}$, $i = 1, \ldots, m$, and $U_{r0}$, $r = 1, \ldots, s$, are defined as before, and $w_i^-, i = 1, \ldots, m$, and $w_r^+, r = 1, \ldots, s$, are user pre-specified constant weights.

Although model (8.5) is not linear, it can be easily transformed into a linear program through a specific change of variables (see Sharp et al. 2007). Despite the

fact that the modified-SBM has good properties, among them translation invariance, one drawback is that the previously shown interpretation for SBM is no longer valid for the modified model.

## 8.4 Non-Translation Invariant Models for Dealing with Negative Data

One of main implications of translation invariance is that models satisfying this property are valid for dealing with several inputs and/or outputs taking negative values. However, there are other alternative approaches defined in literature, which aim to work well when data present some negative values. In this section, we review these approaches.

We start our revision with the work by Halme et al. (2002), which has gone largely overlooked in literature. They considered the problem of working with interval scale data in the CCR and BCC models. In most cases, the negative observations in the data results from the fact that variables are measured on an interval scale and these types of inputs and outputs are usually derived from the difference of two ratio scale variables. For example, profit, which could be negative, is the difference between two non-negative magnitudes: income and cost. Note also that both income and cost are ratio scale variables. In this way, the mentioned authors suggest that the interval scale variable, like profit, should be replaced by the two corresponding ratio scale variables. The two new variables could be interpreted as one output (income) and one input (cost). If we assume that $t \leq m$ inputs and $p \leq s$ outputs are interval scale variables, then we should replace each by two ratio scale variables. In this particular case, Halme et al. (2002) suggest solving the following model:

$$
\begin{aligned}
Max \quad & \beta_0 + \varepsilon \left( \sum_{i=1}^{m+p} s_{i0}^- + \sum_{r=1}^{s+t} s_{r0}^+ \right) \\
s.t. \quad & \\
& \sum_{j=1}^{n} \lambda_{j0} x_{ij} = x_{i0} - \beta_0 x_{i0} - s_{i0}^-, \quad i = 1, \ldots, m+p \\
& \sum_{j=1}^{n} \lambda_{j0} y_{rj} = y_{r0} + \beta_0 y_{r0} + s_{r0}^+, \quad r = 1, \ldots, s+t \\
& \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& \lambda_{j0} \geq 0, \qquad j = 1, \ldots, n \\
& s_{i0}^- \geq 0, \qquad i = 1, \ldots, m+p \\
& s_{r0}^+ \geq 0, \qquad r = 1, \ldots, s+t
\end{aligned}
\tag{8.6}
$$

Basically, (8.6) is the directional distance function model with $(g^-, g^+) = (x_0, y_0)$. Nevertheless, in the objective function of (6) the expression $\varepsilon \left( \Sigma_{i=1}^{m+p} s_{i0}^- + \Sigma_{r=1}^{s+t} s_{r0}^+ \right)$ is maximized, where $\varepsilon > 0$ is a non-Archimedean number. It is well-known that this technique is equivalent to using the method based on two phases for solving radial models. In other words, Halme et al. (2002) were seeking (final) Pareto-efficient targets for all the assessed units. On the other hand, (6) is not translation invariant as a consequence of the definition of the directional vector. Finally, let us point out some advantages and drawbacks of the approach proposed by Halme et al (2002). Regarding the advantages, these authors proved that the units that are initially rated as technically efficient remain efficient when they are evaluated by using (8.6). As for the drawbacks, unfortunately it is necessary to know the value of the components of the interval scale variables for all the units; information that is not always available.

There are other approaches in literature that implement a directional distance function-type model in order to deal with negative data. One of them is by Silva Portela et al. (2004). As we have shown in Sect. 8.3, the RDM introduced by these authors uses a directional vector that is translation invariant, which in turn implies that this particular directional distance function model satisfies translation invariance under VRS. However, a new approach has recently been published that criticizes the RDM with respect to its interpretation. Specifically, Kerstens and Van de Woestyne (2011) argue that a straightforward modification of the well-known proportional distance function may equally be used to accommodate negative data, with the advantage of having a simpler interpretation in terms of the percentage change that facilitates its use in a managerial context. The model that Kerstens and Van de Woestyne (2011) propose to solve is the following:

$$
\begin{aligned}
& Max \quad \beta \\
& s.t. \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} x_{ij} \leq x_{i0} - \beta \, |x_{i0}|, \quad i = 1, \ldots, m \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq y_{r0} + \beta \, |y_{r0}|, \quad r = 1, \ldots, s \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& \qquad \lambda_{j0} \geq 0, \qquad\qquad\qquad\qquad j = 1, \ldots, n
\end{aligned}
\tag{8.7}
$$

Clearly, (8.7) does not meet the property of translation invariance since the corresponding reference vector $(g^-, g^+) = (|x_0|, |y_0|)$ depends on data and is not translation invariant in contrast to the reference vector used by the RDM.

Continuing with the revision of literature, Kazemi Matin and Azizi (2011) have introduced a two-phase approach in order to set suitable targets in DEA when negative data are observed. This paper starts with a critical revision of the existing related literature. Then, the authors propose a new procedure to provide targets with non-negative values corresponding to originally negative variables (basically outputs). An

application on banking is included in the paper in order to illustrate the new procedure and its implications. Specifically, in the application, the amount of returns on equity is considered as an output with some negative observations in the sample, and any assessed unit is required to present a positive value as a target for this variable.

The model by Kazemi Matin and Azizi (2011) consists of two phases. The first one is based on the original additive model (Charnes et al. 1985) and is used for determining a set of targets that will be improved in a subsequent step of the procedure. Let $(s_0^{-*}, s_0^{+*})$ be the optimal slacks of the additive model when $DMU_0$ is rated. Then, the targets in inputs and outputs are defined for this first phase as $\hat{x}_0 = x_0 - s_0^{-*}$ and $\hat{y}_0 = y_0 + s_0^{+*}$. However, these are not the final targets since some output components could be strictly negative[2]. In particular, let us assume that the acceptable role models must have non-negative outputs. Let $R_0''$ and $R_0'$ denote the set of indexes of outputs with and without negative values in $\hat{y}_0$, respectively. Then, in the second phase of the method a modified version of a weighted additive model is solved to determine final targets with all the output components being nonnegative:

$$
\begin{aligned}
Max \quad & \sum_{r \in R_0''} \delta_{r0}^{+} \\
s.t. \quad & \\
& \sum_{j=1}^{n} \lambda_{j0} x_{ij} = \hat{x}_{i0} + \delta_{i0}^{-} \hat{x}_{i0}, \quad && i = 1, \ldots, m \\
& \sum_{j=1}^{n} \lambda_{j0} y_{rj} = \hat{y}_{r0} - \delta_{r0}^{+} \hat{y}_{r0}, \quad && r = 1, \ldots, s \\
& \sum_{j=1}^{n} \lambda_{j0} = 1, && \qquad (8.8) \\
& \delta_{i0}^{-} \leq q_i, && i = 1, \ldots, m \\
& \delta_{r0}^{+} \leq p_r, && r \in R_0' \\
& \delta_{r0}^{+} \geq 0, && r \in R_0'' \\
& \lambda_{j0} \geq 0, && j = 1, \ldots, n
\end{aligned}
$$

where $q_i$ and $p_r$ are pre-defined nonnegative parameters suggested by the decision maker. (8) is a slight modification of a weighted additive model because it is equivalent to a slacks-based model applying the following change of variables: $s_{i0}^{-} = \delta_{i0}^{-} \hat{x}_{i0}$, $i = 1, \ldots, m$ and $s_{r0}^{+} = \delta_{r0}^{+} \hat{y}_{r0}$, $r = 1, \ldots, s$. In this way, the objective function in (8) would be equivalent to $\Sigma_{r \in R_0''} \frac{s_{r0}^{+}}{\hat{y}_{r0}}$ and, therefore (see Sect. 8.3), the second phase of the procedure is not translation invariant. Anyway, by using the original additive

[2] As mentioned before, Seiford and Zhu (2002) is another previous approach where a method based on the extension of facets is proposed in order to achieve the same goal but resorting to radial models.

model, a clear weakness of this approach is that it focuses on determining suitable targets for the evaluated units neglecting the definition of a corresponding efficiency measure with good properties.

Finally, we revise a set of papers that deal with negative data modifying the traditional radial measures. Probably the most well-known among them is the approach introduced by Emrouznejad et al. (2010). In this paper, the authors propose a semi-oriented radial measure (SORM) that permits the presence of variables (inputs or outputs) with positive and negative values for several units in the sample. In particular, these authors treat each variable of this type as consisting of the combination of two new variables as follows. To put it more simply, let us consider an output variable $r'$ so that it is positive for some units and negative for others. Let us now define two new variables $y^1_{r'} \geq 0$ and $y^2_{r'} \geq 0$ as:

$$y^1_{r'j} = \begin{cases} y_{r'j}, & y_{r'j} \geq 0 \\ 0, & y_{r'j} < 0 \end{cases}, \; y^2_{r'j} = \begin{cases} 0, & y_{r'j} \geq 0 \\ -y_{r'j}, & y_{r'j} < 0 \end{cases}, j = 1, \ldots, n. \tag{8.9}$$

In this way, the value of output $y_{r'j}$ can be obtained as $y_{r'j} = y^1_{r'j} - y^2_{r'j}$.

In their paper, Emrouznejad et al. introduce several models, all of which are modifications of the usual radial models in order to accommodate them to different scenarios. In order to illustrate the approach introduced by Emrouznejad et al., we next show the case where an input-orientation is needed and negative and positive values have been observed for the $r'$-th output.

$$\begin{aligned} Min \quad & \theta_0 \\ s.t. \quad & \\ & \sum_{j=1}^{n} \lambda_{j0} x_{ij} \leq \theta_0 x_{i0}, \quad i = 1, \ldots, m \\ & \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq y_{r0}, \quad r = 1, \ldots, s; r \neq r' \\ & \sum_{j=1}^{n} \lambda_{j0} y^1_{r'j} \geq y^1_{r0}, \\ & \sum_{j=1}^{n} \lambda_{j0} y^2_{r'j} \leq y^2_{r'0}, \\ & \sum_{j=1}^{n} \lambda_{j0} = 1, \\ & \lambda_{j0} \geq 0, \quad j = 1, \ldots, n \end{aligned} \tag{8.10}$$

Note that in model (8.10) one of the defined variables has been included as an output, $y^1_{r'}$, and the other one has been included as an input, $y^2_{r'}$. Note also that the approach by Emrouznejad et al. is closely related to the ideas introduced by Halme et al. in 2002, since in both models the variable that takes positive and negative values is decomposed into two variables. The main difference between the two approaches is

that in Halme et al. the components are observable variables with managerial meaning such as income and cost, while in the case by Emrouznejad et al. the mathematical decomposition does not have an economic meaning.

Regarding the translation invariance property and SORM, we next show that this approach is not translation invariant. In fact, model (8.10) is not even translation invariant for output translations despite its resemblance to the traditional input-oriented radial model. To illustrate this point, let us introduce a simple example with one input, one output and three units: A = (1;−1), B = (3;1) and C = (3;−0.5). If we apply model (8.10) for evaluating unit C, then the optimal value equals $\theta_C^* = 2/3$ . However, if we transform the original output y into a new variable $(y+2)$ and apply (8.10) again to assess the level of technical efficiency of C, then $\theta_C^*(y+2) = 1/2 \neq 2/3$ . Consequently, as we pointed out, model (8.10), an input-oriented version of SORM, is not even translation invariant for output translations.

As for the advantages and disadvantages of SORM, we want to highlight that it is always possible to use SORM for handling negative data within DEA in contrast to other alternatives, such as the one by Halme et al. (2002), where it is necessary to know additional information (e.g., income and cost for the case of decomposing profit). This is due to the fact that it is always feasible to treat each variable as being the difference of two nonnegative new magnitudes. As for the drawbacks of SORM, we point out that this technique increases the dimensionality of the problem unnaturally as a consequence of treating negative parts of variables as new variables. This point implies that the number of technically efficient units artificially augments through the application of this method. Moreover, the possible economic meaning does not play any role and, consequently, the results could be misleading.

Kordrostami and Noveiri (2012) is a more modern version of SORM for dealing with negative data. These authors modify SORM to accommodate the scenario where 'flexible variables' are present. A flexible variable can be considered both as an input or as an output. One example for banking evaluation is the number of worthwhile customers. From a prospective view, this variable plays the role of a proxy for future investment, being then an output, but, on the other hand, it may be also considered as an "environmental" input for the analyzed branch (see also Cook and Zhu 2007).

Finally, Cheng et al. (2013) propose a variant of the (traditional) radial model, called VRM, where original values of the rated DMU are replaced with absolute values to estimate the proportion of improvement needed to reach the efficient frontier. The VRM is units invariant and preserves the interpretation of the original radial measures. Specifically, Cheng et al. (2013) introduce two versions: an input-oriented and an output oriented model. In this respect, we would like to highlight that the VRM by Cheng et al. is only a particularization of a more general model that estimates technical efficiency through changing both inputs and outputs at the same time. We are particularly referring to the model introduced by Kerstens and Van de Woestyne (2011), which is based on the directional distance function with the directional vector $(g^-, g^+) = (|x_0|, |y_0|)$. To simplify, here we show only the

input-oriented adaptation of the VRM approach.

$$
\begin{aligned}
& Max \quad \beta \\
& s.t. \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} x_{ij} \leq x_{i0} - \beta |x_{i0}|, \quad i = 1, \ldots, m \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} y_{rj} \geq y_{r0}, \qquad\qquad r = 1, \ldots, s \\
& \qquad \sum_{j=1}^{n} \lambda_{j0} = 1, \\
& \qquad \lambda_{j0} \geq 0, \qquad\qquad j = 1, \ldots, n
\end{aligned}
\tag{8.11}
$$

Note that if $x_{i0} > 0$ for all $i = 1, \ldots, m$ and we take the change of variables $\beta = 1 - \theta$, then (8.11) is equivalent at the optimal solutions to the traditional input-oriented radial model under VRS. On the other hand, it is not hard to prove that the VRM inherits the same properties regarding the translation invariance as the traditional radial models. In other words, the VRS input-oriented VRM is translation invariant to changes in outputs while the VRS output-oriented version is translation invariant to changes in inputs.

## 8.5 The Linear Loss Distance Function Model and the Property of Translation Invariance

In this section, we present a recently introduced distance function: the linear loss distance function (see Pastor and Aparicio 2010; Aparicio and Pastor 2011; Pastor et al. 2012). As we will show later on, the formulation of a linear DEA model resorting to this new distance function is a straightforward approach for an easy check of the translation invariant property.

The linear loss distance function allows us to rewrite any known linear DEA model by considering a specific subset of restrictions each time, known as the linear normalization constraints, together with a fixed structure that always includes the same set of variables, the same objective function and the same fixed subset of restrictions. Consequently, this fix structure or sub-model is common to any DEA linear model. By adding one or more specific linear normalization constraints to this sub-model, we get either the same optimal value of the multiplicative form of the corresponding DEA model or an optimal value that is related to it in a simple linear way. In other words, the DEA model that defines the linear loss distance function in each particular case has exactly the same restrictions as the corresponding multiplicative form and a linearly related objective function. The multiplicative form, which is the dual linear of the more widely used envelopment form, has a nice geometrical interpretation

since the hyperplane obtained associated to its optimal solution is a supporting hyperplane of the production possibility set, where the efficient projection of the unit under scrutiny is located (see Ali and Seiford 1993).

Besides this main structural characteristic of the linear loss distance function models there are other two features worth mentioning. First, provided some mathematical conditions are fulfilled, it is possible to derive new DEA efficiency measures from the linear loss distance function just by considering new normalization constraints (see Aparicio and Pastor 2011). Secondly, the new distance function has a simple dual relationship with respect to the profit function (see Pastor and Aparicio 2010, Proposition 12). Duality allows, among other things, profit efficiency to be decomposed into the usual two components: technical and allocative efficiency.

The linear loss distance function was inspired by Debreu's famous coefficient of resource utilization (Debreu 1951). Specifically, the notion that has most influenced this study is the 'loss function', as a precursor of Debreu's famous coefficient, and which has gone largely overlooked in literature. This concept, which was initially developed for assessing "dead loss" associated with a non-optimal allocation of resources in an economic system, is a (shadow) money metric measure of the distance from an actual allocation to a set of optimal allocations. In order to measure this loss, Debreu suggested resorting to the shadow prices associated with the convex reference technology. The minimization problem suggested by Debreu was $Min_z p_z \cdot (z_0 - z)$, with $z_0$ a vector representing the actual allocation of resources; z a vector belonging to the set of optimal allocations and $p_z$ one of the shadow price vectors of $z$. Debreu also recognized a weakness of his approach: "$p_z$ is affected by an arbitrary positive scalar". The influence of this scalar means that the objective function may be driven to zero by a down scaling of all elements of $p_z$. To avoid this problem, Debreu proposed dividing the objective function by a price index such as $p_z \cdot z_0$, reformulating the original problem as $Min_z p_z \cdot (z_0 - z)/p_z \cdot z_0$, or, equivalently, $Max_z p_z \cdot z/p_z \cdot z_0$. As pointed out by Debreu, an optimal solution to the last maximization problem is $z^* = \rho z_0$, where the scalar $\rho \leq 1$ is the so-called Debreu's coefficient of resource utilization. Nevertheless, we highlight that the influence of the arbitrary multiplicative scalar can also be consistently eliminated by adding a "normalization" constraint to the initial loss minimization problem (see Pastor et al. 2012). Indeed, Debreu's problem can be rewritten equivalently as $Min_z \{p_z \cdot (z_0 - z) : p_z \cdot z_0 = 1\}$.

Next, we introduce the notion of linear loss distance function inspired by Debreu's loss minimization problem. We need to consider a typical production context where each unit is identified through a specific set of inputs or resources that are used to produce a specific set of outputs. As usual, let us consider a unit to be rated, $(x_0, y_0) \in R_+^{m+s}$, with $m$ inputs and $s$ outputs. Let us further denote by $(x, y) \in \partial(T)$ any point that belongs to the Debreu-Farrell frontier $\partial(T)$ of the production possibility set $T$, and by $LNC(c(x, y), p(x, y))$ a set of linear normalization condition(s) on the shadow prices $c$ and $p$.

**Definition 1** Given $(x_0, y_0) \in R_+^{m+s}$ and *LNC*, a finite set of linear normalization constraint(s) on the shadow prices, the loss distance function $L(x_0, y_0; LNC)$ is

defined as the optimal value of the following minimization model.

$$\begin{aligned} L(x_0, y_0; LNC) := \quad & Min \left( \sum_{r=1}^{s} p_r y_r - \sum_{i=1}^{m} c_i x_i \right) - \left( \sum_{r=1}^{s} p_r y_{r0} - \sum_{i=1}^{m} c_i x_{i0} \right) \\ & s.t. \quad (x, y) \in \partial(T), (c(x, y), p(x, y)) \in Q(x, y) \\ & \qquad LNC(c(x, y), p(x, y)) \end{aligned} \tag{8.12}$$

where $Q(x, y)$ is the set of all shadow prices of $(x, y) \in \partial(T)$.

In a DEA framework, and assuming variable returns to scale, Pastor et al. (2012) have proved that (8.12) can be reformulated as the following linear program.

$$\begin{aligned} L(x_0, y_0; LNC) = Min \quad & \alpha - \left( \sum_{r=1}^{s} p_r y_{r0} - \sum_{i=1}^{m} c_i x_{i0} \right) \\ & s.t. \\ & \sum_{r=1}^{s} p_r y_{rj} - \sum_{i=1}^{m} c_i x_{ij} - \alpha \leq 0, \quad \forall j \\ & c \geq 0_m, p \geq 0_s \\ & LNC(c, p, \alpha) \end{aligned} \tag{8.13}$$

where $LNC(c, p, \alpha)$ denotes a finite set of linear normalization constraint(s) defined on the shadow prices of the problem and on the free variable $\alpha$[3]. In fact, $\alpha^*$ at optimum can be interpreted as shadow profit since it equals the value of the profit function at the optimal shadow prices $c^*$ and $p^*$ (recall the objective function of model (8.12)). As a consequence, the aim of (8.13) is to minimize the difference between the profit function and the profit at the assessed point $(x_0, y_0)$, evaluated through shadow prices that satisfy the corresponding linear normalization constraint(s) $LNC(c, p, \alpha)$.

Regarding the flexibility of the linear loss distance function model to encompass the most usual efficiency measures in DEA, pointed out at the beginning of this section, let us show how it is possible to generate the inefficiency part of each measure simply by considering its specific $LNC(c, p, \alpha)$. Many of the existing DEA efficiency or inefficiency measures were originally defined through optimization programs in the quantity space, what we call envelopment form. However, all of them have a dual linear; its multiplier form, which exhibits a common structure in the (shadow) price space and matches with or is closely related to the linear loss distance function model (8.13). For this reason, it is possible to derive each measure

[3] Each linear restriction is either an equality or a non-strict inequality. On the other hand, the generation of a DEA loss function program under different returns to scale is straightforward (Pastor et al. 2012). For instance, in order to get a non-increasing returns to scale program we just add the restriction $\alpha \geq 0$; if we want a constant returns to scale program we just delete $\alpha$ in model (8.13).

**Table 8.1** DEA models and their corresponding normalization constraints

| DEA Models | Normalization Constraint(s) |
|---|---|
| The input-oriented BCC | $cx_0 = 1$ |
| The output-oriented BCC | $py_0 = 1$ |
| The Additive Model | $c \geq 1_m, p \geq 1_s$ |
| The Weighted Additive Model $(w^-, w^+)$ | $c \geq w^-, p \geq w^+$ |
| The input-oriented Russell Model | $c_i \geq \frac{1}{mx_{i0}}, i = 1, \ldots, m$ |
| The output-oriented Russell Model | $p_r \geq \frac{1}{sy_{r0}}, r = 1, \ldots, s$ |
| The Enhanced Russell Graph Model (SBM) | $c_i \geq \frac{1}{mx_{i0}}, i = 1, \ldots, m$<br>$p_r \geq \frac{1}{sy_{r0}}(1 + py_0 - cx_0 - \alpha), r = 1, \ldots, s$ |
| The Directional Distance Function Model $(g^-, g^+)$ | $cg^- + pg^+ = 1$ |
| The Modified Directional Distance Function Model $(g^-, g^+)$ | $cg^- \geq 1$<br>$pg^+ \geq 1$ |

by means of the same linear program where only the normalization condition(s) have been changed according to each particular case. Next, in Table 8.1 we list the normalization constraint(s) associated to a selection of DEA models. According to our previous sections and in order to search for translation invariant models we consider only VRS technologies as model (8.13) does.

If we revise the first model in Table 8.1[4], we can see that the input-oriented BCC model has only one normalization restriction which is a linear equality. The coefficients of the variables $(c, p, \alpha)$ are $x_0$, $0_s$,0. If we now consider the VRS enhanced Russell Graph model, which is the model in Table 8.1 with the most complex set of $(m+s)$ normalization constraints, we can see that the coefficients of the variables $(c, p, \alpha)$ in the first subset of $m$ restrictions are $(0, \ldots, 0, mx_{i0}, 0, \ldots, 0), 0_s, 0$, $i = 1, \ldots, m$, while in the second and last subset of $s$ constraints they are $-x_0, (y_{01}, \ldots, y_{0r-1}, y_{0r}(1-s), y_{0r+1}, \ldots, y_s), -1$, $r = 1, \ldots, s$. As we will show later on, the nature of these coefficients is crucial for deciding if a specific model is translation invariant or not.

The relationship between the optimal value of each model and the optimal value of the loss distance function model is given in Table 8.2.

It is quite curious that all the models whose optimal value, $v^*$, correspond to a true efficiency score, verifying nice properties such as $0 \leq v^* \leq 1$ (see Cooper et al. 1999), satisfy $v^* = 1 - L(x_0, y_0, LNC)$. In Table 8.2, you can find three well-known examples. Additionally, the models that satisfy $v^* = 1 + L(x_0, y_0, LNC)$

[4] The last model of Table 8.1 is defined as follows (see Aparicio et al. 2013): $\max \left\{\beta^- + \beta^+ : \Sigma_{j=1}^n \lambda_j x_j \leq x_0 - \beta^- g^-, \Sigma_{j=1}^n \lambda_j y_j \geq y_0 + \beta^+ g^+, \Sigma_{j=1}^n \lambda_j = 1, \beta^-, \beta^+, \lambda_j \geq 0\right\}$.

**Table 8.2** Relating the optimal value of each DEA model to their corresponding loss distance function

| DEA Model, optimal value $v^*$ | Relation between $v^*$ *and* $L(x_o, y_o, LNC)$ |
|---|---|
| The input-oriented BCC | $v^* = 1 - L(x_0, y_0, LNC_{BCC-IO})$ |
| The output-oriented BCC | $v^* = 1 + L(x_0, y_0, LNC_{BCC-OO})$ |
| The Additive Model | $v^* = L(x_0, y_0, LNC_{AM})$ |
| The Weighted Additive Model $(w^-, w^+)$ | $v^* = L(x_0, y_0, LNC_{WA})$ |
| The input-oriented Russell Model | $v^* = 1 - L(x_0, y_0, LNC_{R-IO})$ |
| The output-oriented Russell Model | $v^* = 1 + L(x_0, y_0, LNC_{R-OO})$ |
| The Enhanced Russell Graph Model (SBM) | $v^* = 1 - L(x_0, y_0, LNC_{ERG})$ |
| The Directional Distance Function Model $(g^-, g^+)$ | $v^* = L(x_0, y_0, LNC_{DDF})$ |
| The Modified Directional Distance Function Model | $v^* = L(x_0, y_0, LNC_{MDDF})$ |

also generate a derived efficiency score between 0 and 1 just by taking the inverse of their optimal value.[5] In Table 8.2, you can find two examples that correspond to output-oriented models. Finally, the models that satisfy $v^* = L(x_0, y_0, LNC)$ estimate profit inefficiency and need some adjustments in their objective functions so as to get a standardized inefficiency, i.e. an inefficiency value $v^*_{St}$ between 0 and 1, or, equivalently, an efficiency score defined simply as $1 - v^*_{St}$. Well-known examples of these adjustments in connection with the weighted additive models are the RAM (Cooper et al. 1999) and the BAM (Cooper et al. 2011) measures of efficiency.

Now we are ready to present a proposition that states a condition under which the linear loss distance function is translation invariance or at least one-sided translation invariance, i.e., either in inputs or in outputs.

**Proposition 6** If any of the constraints of *LNC* do not involve the value of $\alpha$ and all the coefficients of the variables $(c, p)$ that appear in the constraints of *LNC* are translation invariant in inputs and outputs then model (8.13) is translation invariant.

*Proof.* The hypothesis of Proposition 6 guarantees that the finite subset of *LNC* restrictions is translation invariant. We propose a sequential proof with $m + s$ steps. Since a translation of $m$ inputs and $s$ outputs can be decomposed as a sequence of $m + s$ translations, each of which moves only a single variable, let us consider without loss of generality, that only the first input is translated. Specifically, we generate a new input 1, defined as $x'_{1j} = x_{1j} + h_1$, with $h_1 \in R$. Let us denote this new model (8.13) with translated data as (8.13'). Let us also consider an optimal solution of (8.13) and denote it as $(c^*, p^*, \alpha^*)$. We are going to prove that $(c^*, p^*, \alpha')$ with $\alpha' = \alpha^* - c^*_1 h_1$

[5] The known relationship between the values of the Shephard input and output distance functions (Shephard 1953) under CRS suggests the last introduced efficiency score.

is a feasible solution of (8.13'). The definition of $(c^*, p^*, \alpha^*)$ guarantees that $c^* \geq 0_m$, $p^* \geq 0_s$. The hypothesis guarantees that $LNC(c^*, p^*, \alpha') = LNC(c^*, p^*, \alpha^*)$ and that, therefore, $(c^*, p^*, \alpha')$ satisfies all the normalization constraints. To complete the proof we need to verify that $\Sigma_{r=1}^{s} p_r^* y_{rj} - c_1^* x'_{1j} - \Sigma_{i\neq 1}^{m} c_i^* x_{ij} - \alpha' \leq 0, \forall j$. It is accomplished as follows. $\Sigma_{r=1}^{s} p_r^* y_{rj} - c_1^* x'_{1j} - \Sigma_{i\neq 1}^{m} c_i^* x_{ij} - \alpha' = \Sigma_{r=1}^{s} p_r^* y_{rj} - c_1^*(x_{1j} + h_1) - \Sigma_{i\neq 1}^{m} c_i^* x_{ij} - (\alpha^* - c_1^* h_1) = \Sigma_{r=1}^{s} p_r^* y_{rj} - \Sigma_{i=1}^{m} c_i^* x_{ij} - \alpha^* \leq 0$, for all $j = 1, \ldots, n$.

Hence, $(c^*, p^*, \alpha')$ is a feasible solution of (8.13'). Regarding the corresponding value of the objective function the next chain of equalities show that it is the same as the optimal value of (8.13): $\alpha' - (\Sigma_{r=1}^{s} p_r^* y_{r0} - c_1^* x'_{10} - \Sigma_{i\neq 1}^{m} c_i^* x_{i0}) = \alpha^* - c_1^* d - (\Sigma_{r=1}^{s} p_r^* y_{r0} - c_1^*(x_{10} + d) - \Sigma_{i\neq 1}^{m} c_i^* x_{i0}) = \alpha^* - (\Sigma_{r=1}^{s} p_r^* y_{r0} - \Sigma_{i=1}^{m} c_i^* x_{i0})$.

Let us finally prove that $(c^*, p^*, \alpha')$ is an optimal solution of (8.13'), by contradiction. If it is not, there would exist another feasible solution of (8.13'), $(\hat{c}, \hat{p}, \hat{\alpha})$, such that $\hat{\alpha} - (\Sigma_{r=1}^{s} \hat{p}_r y_{r0} - \hat{c}_1 x'_{10} - \Sigma_{i\neq 1}^{m} \hat{c}_i x_{i0}) < \alpha' - (\Sigma_{r=1}^{s} p_r^* y_{r0} - c_1^* x'_{10} - \Sigma_{i\neq 1}^{m} c_i^* x_{i0})$. But then, it is not hard to prove, following the same steps as before, that $(\hat{c}, \hat{p}, \tilde{\alpha})$ with $\tilde{\alpha} = \hat{\alpha} + \hat{c}_1 h_1$ would be a feasible solution of (13). In addition, $\tilde{\alpha} - (\Sigma_{r=1}^{s} \hat{p}_r y_{r0} - \Sigma_{i=1}^{m} \hat{c}_i x_{i0}) = \hat{\alpha} + \hat{c}_1 h_1 - (\Sigma_{r=1}^{s} \hat{p}_r y_{r0} - \hat{c}_1(x'_{10} - h_1) - \Sigma_{i\neq 1}^{m} \hat{c}_i x_{i0}) = \hat{\alpha} - (\Sigma_{r=1}^{s} \hat{p}_r y_{r0} - \hat{c}_1 x'_{10} - \Sigma_{i\neq 1}^{m} \hat{c}_i x_{i0}) < \alpha' - (\Sigma_{r=1}^{s} p_r^* y_{r0} - c_1^* x'_{10} - \Sigma_{i\neq 1}^{m} c_i^* x_{i0}) = \alpha^* - (\Sigma_{r=1}^{s} p_r^* y_{r0} - \Sigma_{i=1}^{m} c_i^* x_{i0})$, which contradicts that $(c^*, p^*, \alpha^*)$is an optimal solution of (8.13). Therefore, $(c^*, p^*, \alpha')$ is an optimal solution of (8.13') and, as shown before, the optimal values of both models, (8.13) and (8.13'), coincide. This completes the proof[6].

We also note that: (1) since the optimal solution of model (8.13) always corresponds to an inefficiency measure we can talk about a "translation invariant inefficiency measure". (2) If the inefficiency measure of model (8.13) is linearly related to a true efficiency measure—see Table 8.2—we can talk about a "translation invariant efficiency measure". (3) The proof of the last proposition clearly shows why a CRS model cannot be translation invariant. In fact, CRS requires $\alpha = 0$ and any translation modifies the value of $\alpha$!

Next, we present a by-product of the above result that allows us to relate, among other things, the linear loss distance function to translation invariance when input or output orientation is assumed.

**Corollary 6.1** If any of the constraints of *LNC* do not involve the value of $\alpha$ and if all the coefficients associated to the variables $(c, p)$ that appear in the constraints of *LNC* are only translation invariant in inputs (outputs), then model (8.13) is only translation invariant in inputs (outputs).

Proof. It is a direct consequence of Proposition 6.

The findings of the last results clearly show that translation invariance in a DEA model depends only on the behavior of its linear normalization restrictions. We have created Table 8.3 to systematize these results. Looking at Table 8.1, we see

[6] If we were translating the first output instead of the first input the proof is completely similar. The new translated input would be $y'_{1j} = y_{1j} + k_1, \forall j$, which generates a derived feasible solution defined as $(c^*, p^*, \alpha')$ where $\alpha' = \alpha^* + p_1^* k_1$.

**Table 8.3** DEA models, normalization constraints and translation invariance

| DEA models | Linear Normalization equalities or inequalities | Translation invariance |
|---|---|---|
| The input-oriented BCC | $cx_0 = 1$<br>Depends on input values | *Only for outputs* |
| The output-oriented BCC | $py_0 = 1$:<br>Depends on output values | *Only for inputs* |
| The Additive Model | $c \geq 1_m, p \geq 1_s$ | *Yes* |
| The Weighted Additive Model $(w^-, w^+)$ | $c \geq w^-, p \geq w^+$<br>Depends on $(w^-, w^+)$ | *Not always* |
| The input-oriented Russell Model | $c_i \geq \frac{1}{mx_{i0}}, i = 1, ..., m$<br>Depends on input values | *Only for outputs* |
| The output-oriented Russell Model | $p_r \geq \frac{1}{sy_{r0}}, r = 1, \ldots, s$<br>Depends on output values | *Only for inputs* |
| The Enhanced Russell Graph Model (SBM) | $c_i \geq \frac{1}{mx_{i0}}, i = 1, ..., m$<br>$p_r \geq \frac{1}{sy_{r0}}(1 + py_0 - cx_0 - \alpha), r = 1, ..., s$<br>First (second) restriction depends on input (output) values | *No* |
| The Directional Distance Function Model $(g^-, g^+)$ | $cg^- + pg^+ = 1$<br>Depends on $(g^-, g^+)$ | *Not always* |
| The Modified Directional Distance Function Model $(g^-, g^+)$ | $cg^- \geq 1$<br>$pg^+ \geq 1$<br>First (second) restriction depends on g- ($g^+$) | *Not always* |

that the coefficients of the variables of the unique restriction of $LNC_{BCC-IO}$ do depend only on the input values of the unit being rated. Therefore the BCC-IO model is only translation invariant for outputs. A similar and symmetric reasoning is valid for the BCC-OO model. The Weighted Additive model is translation invariant provided the weights attached to the objective function are translation invariant. In particular, the Additive Model (Charnes et al. 1985) is a translation invariant model. The RAM and the BAM models are also translation invariant. On the other hand, the input-oriented Russell model is translation invariant in outputs, while the output-oriented Russell model is translation invariant in inputs for the same reasons as the BCC models. Moreover, the Enhanced Russell Graph model is never translation invariant. As a matter of fact, the first subset of $m$ restrictions of $LNC_{ERG}$ is only translation invariant for outputs, and the second subset of $s$ restrictions is only translation invariant for inputs. Finally, the Directional Distance Function and the Modified- Directional Distance Function models are in general not translation invariant. Both models are translation invariant if $(g^-, g^+)$ is a constant vector. But they are not if, e.g. $(g^-, g^+) = (x_0, y_0)$ in contrast to the claim by Färe and Grosskopf (2013).

## 8.6 Conclusions

We have been able to accomplish two very differentiated tasks. The first one has been to present a state of the art of translation invariant DEA models, extended to non-translation invariant DEA models that, nonetheless, are able to deal with negative data. The second one has been to present a new distance function, the linear loss distance function, which constitutes a powerful tool for revising translation invariance of any known DEA model. As far as we know, the linear loss distance function is the first distance function defined on the multiplier—or shadow price—space. Its most relevant feature is that it unifies all the existing DEA models under a similar structure which allows an easy checking of the different models. Consequently we have been able to revise the most well known DEA models and to classify them according to their translation invariant characteristics.

**Acknowledgements** We would like to thank Prof. Joe Zhu for kindly inviting us to contribute a chapter to the edition of this book. Additionally, Juan Aparicio is grateful to the Generalitat Valenciana for supporting this research with grant GV/2013/112.

## References

Ali AI, Seiford LM (1990) Translation Invariance in Data Envelopment Analysis. Oper Res Lett 9:403–405

Ali AI, Seiford LM (1993) The mathematical programming approach to efficiency analysis. In: Fried H, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency: techniques and applications. Oxford University Press, Oxford.

Aparicio J, Pastor JT (2011) A general input distance function based on opportunity costs. Adv Decis Sci Article ID 505241, 11 pages. doi: 10.1155/2011/505241

Aparicio J, Pastor JT, Ray SC (2013) An overall measure of technical inefficiency at the firm and at the industry level: the 'lost profit on outlay'. Eur J Oper Res 226:154–162

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions, and Nerlovian efficiency. J Optim Theory Appl 98(2):351–364

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Rhodes E (1979) Short communication: measuring the efficiency of decision making units. Eur J Oper Res 3(4):339

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for pareto-koopmans efficient empirical production functions. J Econom 30:91–107

Charnes A, Cooper WW, Thrall RM (1986) Classifying and characterizing efficiencies and inefficiencies in data envelopment analysis. Oper Res Lett 5(3):105–110

Charnes A, Cooper WW, Rousseau J, Semple J (1987) Data envelopment analysis and axiomatic notions of efficiency and reference sets. Res Report CCS558, Center for Cybernetic Studies, University of Texas, Austin TX, USA

Cheng G, Zervopoulos P, Qian Z (2013) A variant of radial measure capable of dealing with negative inputs and outputs in data envelopment analysis. Eur J Oper Res 225:100–105

Cook WD, Zhu J (2007) Classifying inputs and outputs in data envelopment analysis. Eur J Oper Res 80:692–699

Cooper WW, Park KS, Pastor JT (1999) RAM: a range adjusted measure of inefficiency for use with additive models, and relations to others models and measures in DEA. J Product Anal 11:5–42
Cooper WW, Pastor JT, Borras F, Aparicio J, Pastor D (2011) BAM: a bounded adjusted measure of efficiency for use with bounded additive models. J Product Anal 35:85–94
Debreu G (1951) The coefficient of resource utilization. Econometrica 19:273–292
Emrouznejad A, Anouze AL, Thanassoulis E (2010) A semi-oriented radial measure for measuring the efficiency of decision making units with negative data, using DEA. Eur J Oper Res 200:297–304
Färe R, Grosskopf S (2013) DEA, directional distance functions and positive, affine data transformation. Omega 41:28–30
Hadad MD, Hall MJB, Kenjegalieva KA, Santoso W, Simper R (2012) A new approach to dealing with negative numbers in efficiency analysis: an application to the Indonesian banking sector. Expert Syst Appl 39:8212–8219
Hadi-Vencheh A, Esmaeilzadeh A (2013) A new super-efficiency model in the presence of negative data. J Oper Res Soc 64:396–401
Halme M, Joro T, Koivu M (2002) Dealing with interval scale data in data envelopment analysis. Eur J Oper Res 137:22–27
Jahanshahloo GR, Piri M (2013) Data Envelopment Analysis (DEA) with integer and negative inputs and outputs. J Data Envel Anal Decis Sci 2013:1–15
Kazemi Matin R, Azizi R (2011) A two-phase approach for setting targets in DEA with negative data. Appl Math Modell 35:5794–5803
Kerstens K, Van de Woestyne I (2011) Negative data in DEA: a simple proportional distance function approach. J Oper Res Soc 62:1413–1419
Kordrostami S, Noveiri MJS (2012) Evaluating the efficiency of decision making units in the presence of flexible and negative data. Indian J Sci Technol 5(12):3776–3782
Lovell CAK, Pastor JT (1995) Units Invariant and Translation Invariant DEA Models. Oper Res Lett 18:147–151
Pastor JT (1994) New additive DEA models for handling zero and negative data. Working Paper, Universidad de Alicante (out of print)
Pastor JT (1996) Translation Invariance in DEA: a generalization. Ann Oper Res 66:93–102
Pastor JT, Aparicio J (2010) Distance functions and efficiency measurement. Indian Econ Rev XXXXV:193–231
Pastor JT, Ruiz JL (2007) Variables with negative values in DEA.. In: Zhu J, Cook WD (eds) Modeling data irregularities and structural complexities in data envelopment analysis. Springer, US
Pastor JT, Ruiz JL, Sirvent I (1999) An enhanced DEA Russell Graph efficiency measure. Eur J Oper Res 115:596–607
Pastor JT, Lovell CAK, Aparicio J (2012) Families of linear efficiency programs based on Debreu's loss function. J Product Anal 38:109–120
Seiford LM, Zhu J (2002) Classification invariance in data envelopment analysis. In: Misra JC (ed) Uncertainty and optimality, probability, statistics and operations research, World Scientific
Sharp JA, Meng W, Liu W (2007) A modified slacks-based measure model for data envelopment analysis with 'natural' negative outputs and inputs. J Oper Res Soc 58:1672–1677
Shephard RW (1953) Cost and production functions. Princeton University, Princeton
Silva Portela MCA, Thanassoulis E (2010) Malmquist-type indices in the presence of negative data: an application to bank branches. J Bank Financ 34:1472–1483
Silva Portela MCA, Thanassoulis E, Simpson G (2004) Negative data in DEA: a directional distance function approach applied to bank branches. J Oper Res Soc 55:1111–1121
Thanassoulis E, Silva Portela M, Despic O (2008) DEA—the mathematical programming approach to efficiency analysis. In: Fried H, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency and productivity change. Oxford University, New York, pp 251–420
Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509

# Chapter 9
# Scale Elasticity in Non-parametric DEA Approach

**Biresh K. Sahoo and Kaoru Tone**

**Abstract** This contribution is an attempt to make an exhaustive critical review of various possible estimation methods of scale economies in a non-parametric data envelopment analysis approach. Three types of technology structure—piecewise linear, piecewise log-linear, and FDH—are found to be adopted for such estimation exercise. These technology structures are built up either in input-output space or in cost-output space. The strengths and weaknesses of the uses of each of these estimation methods are discussed. The issue of which method to use in any empirical application is a matter of an examination of various issues concerning (1) whether factor inputs are indivisible, (2) whether price data are available, and if available, whether they are well measured with certainty, and (3) whether the non-convexities in the underlying production technology are present.

**Keywords** Scale elasticity · Returns to scale · Economies of scale · Data envelopment analysis

## 9.1 Introduction

Competition is a driving force behind numerous important policy changes. It puts forth downward pressure on costs, reduces slacks, provides incentives, and drives innovation forward. To analyze the performance of a firm, the concept of *productivity growth* has been widely used in the literature, and the sources of this growth are largely due to the contributions from either *economies of scale (returns to scale, RTS)* or *technical change* or both. However, empirical evidence suggests that the sources of productivity growth are due to more of increasing returns to scale and

B. K. Sahoo (✉)
Xavier Institute of Management, Xavier Square, 751013 Bhubaneswar, India
e-mail: biresh@ximb.ac.in

K. Tone
National Graduate Institute for Policy Studies,
7-22-1 Roppongi, Minato-ku, 106-8677 Tokyo, Japan
e-mail: tone@grips.ac.jp

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_9

less of technical progress (Rosenberg 1963, 1981; Sokoloff 1988; Morrison 1992; Devereux et al. 1996; Basu and Fernald 1997; Jones 2004).

One of the most important aspects in applied production analysis of firms is the measurement of RTS since its informational contents can provide important insights to firm managers making operational decisions in strengthening their competitive position. In this paper, we concentrate on this aspect within the framework of data envelopment analysis (DEA). The RTS or *scale elasticity* (SE) or *Passus Coefficient* in the terminology of Frisch (1965), is the property of a production function, and is regularly used to describe the relationship between *scale* and *efficiency*. In case of a multi-output-multi-input production technology, the RTS relates to the case where it measures the maximum proportional increase in all outputs relative to a given proportional increase in all inputs. Constant RTS are said to prevail at a point on the production frontier if an increase of all inputs by, say, 1 % leads to an increase of all outputs by 1 %. Decreasing RTS are present if outputs increase by less than 1 %, while increasing RTS exist if they increase by more than 1 %.The changes in input- and output mixes remain constant in this measure of SE. See Hanoch (1970), Starrett (1977), Panzar and Willig (1977) and Baumol et al. (1982) for the detailed discussion on this.

Economies of scale may arise from four major sources: (a) economies of scope (b) indivisibility comprising size, (c) learning by doing through cumulative experience, and (d) reduced input costs due to power over suppliers. The benefits of scale increases flow from the many diverse components of the concept of a 'firm'. The emphasis here is not only on technology but more on the entire gamut of organization, management, learning by doing, reorganization of inputs, and other capabilities of firm. These sources of scale can be easily traced in the Silberston's broader definition of scale—"economies of scale can be said to exist if an expansion in the volume of output produced results in a decrease in the unit cost of production when at each higher level of output, all possible adaptations in technology and organization have been carried through" (Silberston 1972).

Therefore, an appropriate estimation strategy for the underlying production technology structure is essential in understanding and capturing the RTS properties of a firm. We have chosen the DEA for the estimation of RTS. The non-parametric DEA approach yields both *qualitative* and *quantitative* information about the RTS. Using this approach, the RTS estimation has been researched in many studies since 1984. See, e.g., Banker et al. (1984, 1986, 1996a, b, 2004); Färe et al. (1986, 1988); Banker and Thrall (1992); Zhu and Shen (1995); Førsund (1996); Golany and Yu (1997); Sueyoshi (1997, 1999); Seiford and Zhu (1998, 1999); Sahoo et al. (1999, 2012); Zhu (2000); Tone (2001); Tone and Sahoo (2003); Førsund and Hjalmarsson (2004); Hadjicostas and Soteriou (2006); Førsund et al. (2007); Sueyoshi and Sekitani (2007a, b); Podinovski et al. (2009); Podinovski and Førsund (2010), Zelenyuk (2013), and Sahoo and Sengupta (2014), among others.

In the DEA literature, researchers involved in the empirical estimation of RTS, generally, uses either a *factor-based* technology set or a *cost-based* technology set. Either of the two technology sets is generally taken to be a satisfactory way of empirically verifying scale economies behavior without mentioning whether they

are taken to highlight the same causal factors. One can, however, argue that the shortcomings associated with RTS measure underlying the factor-based technology set are too strong enough to measure any relevant scale effects of a real-life firm. This is because in this measure the output is related to the inputs only by defining the input-mix in a special way, e.g., as a replication measure, as a size measure, or as a long-run measure of only one input such as plant and machinery or capital.

The replication measure is purely statistical in nature, often, used in the statistical theory of design of experiments. Any firm can be replicated by a new firm, but this has no economic meaning because it is not a controlled experiment. The techniques and inputs used at higher scale are very different from those used at lower scale. And, it is usually very hard to judge the economic relevance (i.e., effect) of a replicated firm of a given size (i.e., treatment) unless it is well compared with an actual firm of that size. The issue involved here is how representative a replicated firm is of an actual firm to which the investigator would like to project. It can be argued that a replicated firm may not properly represent an actual firm because of *indivisibilities* associated not only in technology, but also in unique attributes associated with geographic locations and innovative mangers.

The size measure in terms of inputs is not unequivocal. In agricultural economics, it is natural to measure size by acres, but in industrial manufacturing there is no natural measure. The measure of plant and machinery is difficult due to heterogeneity except through costs. Note that if the current input mix can be represented by a size measure, then the *size elasticity* of output is a good measure of RTS, and so is for the plant and machinery. In the light of the aforementioned problems, the cost measure seems to score well over the factor-based technology set measure to estimate scale economies.

Therefore, when inputs/outputs are heterogeneous across firms, the construction of factor-based technology set in DEA becomes problematic. As a result, the alternative cost-based technology set is very useful in estimating various productive efficiency behaviors of firms. See, e.g., Färe and Grosskopf (1985); Grosskopf et al. (1987); Grosskopf and Yaiswarng (1990); Fried et al. (1998); Sengupta (1999, 2002, 2003, 2004a, b, 2005a, b); Sahoo et al. (2007, 2012, 2014a); Sahoo (2008); Sahoo and Tone (2009a, b); Sahoo and Gstach (2011); Sahoo and Tone (2013), among others. We use this approach to discuss the SE estimation of firms. Note that in the presence of factor heterogeneity (indivisibility), the non-convex (log convex) technology set could also be used to determine the RTS properties of firms.

Note that all the standard methods of determining SE in the DEA approach proceed by examining tangential planes to the frontier at a given point. This is done either by looking at the constant term (the variable $u_0$ originally introduced in the literature by Banker et al. 1984) that represents the intercept of that plane with the plane in which all inputs are set to zero or, by observing the weights of the corner points of the facet of the frontier associated with that plane. This determination, however, may be difficult because the plane need not be unique. In this study we will, therefore, deal with both the right-hand (lower bound) and left-hand (upper bound) SEs. Note that one could also use the Golany and Yu (1997)'s envelopment DEA model to determine the right-hand and left-hand SE estimates.

The remainder of the paper proceeds as follows. Section 9.2 first deals with the evaluation of right-hand and left-hand SEs based on a factor-based technology set involving no indivisibility and indivisibility respectively, then argue for the use of the multiplicative DEA models to obtain the exact estimates of SEs. Finally, this section deals with the discussion of SE based on the cost-based technology set when inputs and outputs are considered heterogeneous. Section 9.3 concludes with some remarks.

## 9.2 Technology Specification and Scale Elasticity

### *9.2.1 Technology*

There are at least two approaches for estimating scale elasticity parameter in production economics literature: (1) production/distance function approach in the primal, and (2) support functions approach such as cost and revenue functions in the dual. Throughout we assume to deal with $n$ observed firms; each uses $m$ inputs to produce $s$ outputs. Let $x_j = (x_{1j}, \ldots, x_{mj})^T \in \mathbb{R}^m_{\geq 0}$ and $y_j = (y_{1j}, \ldots, y_{sj})^T \in \mathbb{R}^s_{\geq 0}$ be, respectively, the vector of inputs and outputs of firm $j$; $w_j = (w_{1j}, \ldots, w_{mj}) \in \mathbb{R}^m_{\geq 0}$ and $p_j = (p_{1j}, \ldots, p_{sj}) \in \mathbb{R}^s_{\geq 0}$ be, respectively, the price vectors of inputs and outputs of firm $j$; and $J$ be the index set of all the observed firms, i.e., $J = \{1, \ldots, n\}$.

The production technology that transforms an input vector $x \in \mathbb{R}^m_{\geq 0}$ to an output vector $y \in \mathbb{R}^s_{\geq 0}$, can be characterized by the technology set $T \subset \mathbb{R}^m_{\geq 0} \times \mathbb{R}^s_{\geq 0}$, defined as

$$T = \left\{(x, y) \in \mathbb{R}^{m+s}_{\geq 0} \,\middle|\, x \in \mathbb{R}^m_{\geq 0} \text{ can produce } y \in \mathbb{R}^s_{\geq 0}\right\} \tag{9.1}$$

We assume here that the set $T$ satisfies the following axioms to ensure the existence of duality between cost and production: (a) inactivity is allowed, (b) "free lunch" is not allowed, (c) free disposability of both inputs and outputs and (d) technology set is compact and convex (Färe 1988; Fare and Primont 1995). The neoclassical characterization of production function is the transformation function $\psi(x, y)$, which decreases with $y$ and increases with $x$ such that

$$\psi(x, y) \leq 0 \text{ if and only if } (x, y) \in T \tag{9.2}$$

$\psi(x, y) = 0$ represents those input–output vectors that operate on the boundary of the set $T$, and hence are technically efficient. The set $T$ can also be described by an input requirement set $L(y)$:

$$L(y) = \{x : (x, y) \in T\} \tag{9.3}$$

or by an output set $P(x)$:

$$P(x) = \{y : (x, y) \in T\} \tag{9.4}$$

or by an input distance function $D_i(x, y)$:

$$D_i(x, y) = \inf \{\alpha : \alpha x \in L(x)\} \tag{9.5}$$

orby an output distance function $D_o(x, y)$:

$$D_o(x, y) = \sup \{\beta : \beta y \in P(y)\} \tag{9.6}$$

Note that $x \in L(x)$ if and only if $D_i(x, y) \leq 1$, $y \in P(y)$ if and only if $D_o(x, y) \geq 1$. $D_i(x, y)$ and $D_o(x, y)$ are linearly homogeneous in $x$ and $y$, respectively.

### 9.2.2 *Primal Measure of Scale Elasticity*

The SE is based on a relationship that with a given proportional expansion of all inputs ($\alpha$), one can find out the maximum proportional expansion in all outputs ($\beta$) such that

$$\psi(\alpha x, \beta y) = 0 \tag{9.7}$$

On differentiation of (9.7) with respect to input scaling factor $\alpha$ yields the following measure of SE $\varepsilon(x, y)$ (Hanoch 1970; Panzar and Willig 1977):

$$\frac{\partial \beta}{\partial \alpha} = \varepsilon(x, y) = -\frac{\sum_{i=1}^{m} \frac{\partial \psi(\cdot)}{\partial x_i} x_i}{\sum_{r=1}^{s} \frac{\partial \psi(\cdot)}{\partial y_r} y_r} \tag{9.8}$$

**Proposition 1** The RTS defined at point $(x, y)$ are increasing (IRS), constant (CRS) and decreasing (DRS) if $\varepsilon(x, y) > 1$, $\varepsilon(x, y) = 1$ and $\varepsilon(x, y) < 1$, respectively.

Färe et al. (1986, 1988) redefine the transformation function $\psi(x, y)$ as $(D_i(x, y) - 1)$, and this redefinition yields the following input-oriented measure of SE:

$$\varepsilon_i(x, y) = D_i(\cdot) \Big/ \left( \sum_{r=1}^{s} \frac{\partial D_i(\cdot)}{\partial y_r} y_r \right) \tag{9.9}$$

Similarly, by redefining $\psi(x, y)$ as $(1 - D_o(x, y))$, Färe et al. (1986, 1988) derive the output-oriented measure of SE as

$$\varepsilon_o(x, y) = \left( \sum_{i=1}^{m} \frac{\partial D_o(\cdot)}{\partial x_i} x_i \right) \Big/ D_o(\cdot) \tag{9.10}$$

### 9.2.3 Dual Measure of Scale Elasticity

Following Panzar and Willig (1977) and Baumol et al. (1982), the dual measure of SE, called *cost elasticity* $\varepsilon_c(y,w)$, is defined as

$$\varepsilon_c(y,w) = C(y;w) \Big/ \left( \sum_{r=1}^{s} \frac{\partial C(y;w)}{\partial y_r} \right) \tag{9.11}$$

Where $C(y;w) = \min_x \{w.x : x \in L(y)\}$ is the minimum cost of producing the output vector $y$ at the input price vector $w$. However, the duality relationship between cost function $C(y;w)$ and input distance function $D_i(x,y)$ suggests that the scale elasticity and the cost elasticity are same, i.e.,

$$\varepsilon_c(y,w) = C(y;w) \Big/ \left( \sum_{r=1}^{s} \frac{\partial C(y;w)}{\partial y_r} \right) = \varepsilon_i(x,y) = D_i(\cdot) \Big/ \left( \sum_{r=1}^{s} \frac{\partial D_i(\cdot)}{\partial y_r} y_r \right) \tag{9.12}$$

Similarly, the duality relationship between revenue function $R(x;p)$ and the output distance function $D_o(x,y)$ implies that the revenue elasticity $\varepsilon_r(x,p)$ and the scale elasticity $\varepsilon_o(x,y)$ are the same, i.e.,

$$\varepsilon_r(x,p) = \left( \sum_{i=1}^{m} \frac{\partial R(x;p)}{\partial x_i} x_i \right) \Big/ R(x;p) = \varepsilon_o(x,y) = \left( \sum_{i=1}^{m} \frac{\partial D_o(\cdot)}{\partial x_i} x_i \right) \Big/ D_o(\cdot) \tag{9.13}$$

where $R(x;p) = \max_y \{p.y : y \in P(x)\}$ is the maximum revenue obtained from the input vector x at the given the output price vector $p$. See Fare and Primont (1995, pp. 44–54) for the proof.

### 9.2.4 Scale Elasticity in DEA Models

#### 9.2.4.1 Scale Elasticity in Production DEA Models

The DEA estimator of the true but unknown technology set that allows variable returns to scale (VRS), is the BCC technology (Banker et al. 1984) given by

$$T_{VRS}^{DEA} = \left\{ (x,y) : \sum_{j \in J} x_{ij}\lambda_j \le x_i (\forall i), \sum_{j \in J} y_{rj}\lambda_j \ge y_r (\forall r), \sum_{j \in J} \lambda_j = 1, \lambda_j \ge 0 (\forall j) \right\} \tag{9.14}$$

Now we consider the evaluation of input-oriented SE for any firm $o$ ($o \in J$) in the $T_{VRS}^{DEA}$. The input-oriented technical efficiency of firm $o$ can be obtained from the following linear programming (LP) problem:

$$D_i(x_o, y_o) = \alpha(\beta) = \min \left\{ \alpha : (\alpha x, \beta y) \in T_{VRS}^{DEA}; \beta = 1 \right\} \tag{9.15}$$

Alternatively, the primal *envelopment*-form based LP program (9.15) can be set up in its dual *multiplier* form as

$$D_i(x_o, y_o) = \alpha(1) = \max \sum_{r=1}^{s} u_r y_{ro} - u_o \tag{9.16}$$

$$\text{s.t. } \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - u_o \leq 0,$$

$$\sum_{i=1}^{m} v_i x_{io} = 1,$$

$$u_r, v_i \geq 0(\forall i, r); u_o : free.$$

For any firm $o$ ($o \in J$), the following transformation function

$$\psi(\alpha(1)x_o, y_o) \equiv \sum_{r=1}^{s} u_r y_{ro} - \sum_{i=1}^{m} v_i(\alpha(1)x_{io}) - u_o = 0 \tag{9.17}$$

Using the formula (9.8), the input-oriented SE of firm $o$ can be obtained as

$$\varepsilon_i(x_o, y_o) = -\frac{\sum_{i=1}^{m} \frac{\partial \psi(\alpha(1)x_o, y_o)}{\partial x_{io}} x_{io}}{\sum_{r=1}^{s} \frac{\partial \psi(\alpha(1)x_o, y_o)}{\partial y_{ro}} y_{ro}} = \frac{\alpha(1)}{\alpha(1) + u_o} = \frac{1}{1 + u_o / D_i(x_o, y_o)} \tag{9.18}$$

Now let us turn to the SE evaluation of firm $o$ in an output-oriented BCC model given by

$$D_o(x_o, y_o) = \beta(\alpha) = \max \left\{ \beta : (\alpha x, \beta y) \in T_{VRS}^{DEA}; \alpha = 1 \right\} \tag{9.19}$$

Alternatively, the envelopment-form based LP problem (9.19) can be set up in its dual multiplier form as

$$D_o(x_o, y_o) = \beta(1) = \min \sum_{i=1}^{m} v_i x_{io} + v_o \tag{9.20}$$

$$\text{s.t. } -\sum_{r=1}^{s} u_r y_{rj} + \sum_{i=1}^{m} v_i x_{ij} + v_o \geq 0,$$

$$\sum_{r=1}^{s} u_r y_{ro} = 1,$$

$$u_r, v_i \geq 0(\forall r, i); v_o : free.$$

For any firm $o$ ($o \in J$), the following transformation function

$$\psi(x_o, \beta y_o) \equiv \sum_{r=1}^{s} u_r(\beta(1)y_{ro}) - \sum_{i=1}^{m} v_i x_{io} - v_o = 0 \tag{9.21}$$

Using the formula (9.8), the output-oriented SE of firm $o$ can now be obtained as

$$\varepsilon_o(x_o, y_o) = -\frac{\sum_{i=1}^{m} \frac{\partial \psi(x_o, \beta(1) y_o)}{\partial x_{io}} x_{io}}{\sum_{r=1}^{s} \frac{\partial \psi(x_o, \beta(1) y_o)}{\partial y_{ro}} y_{ro}} = \frac{\beta(1) - v_o}{\beta(1)} = 1 - \frac{v_o}{D_o(x_o, y_o)} \tag{9.22}$$

where $\beta(1) = D_o(x_o, y_o)$ is the reciprocal measure of output technical efficiency.

It is well known that the DEA technologies are not differentiable at extreme efficient points due to multiple optimal solutions for $u_o(v_o)$. Following Banker and Thrall (1992), we, therefore set up the following LP problems to find out the maximum and minimum values of $u_o$ for firm $o$ as:

$$\begin{aligned} & u_o^+(u_o^-) = \max(\min) u_o \qquad (9.23) \\ & \text{s.t.} \sum_{r=1}^{s} u_r y_{ro} - u_o = D_i(x_o, y_o), \sum_{i=1}^{m} v_i x_{io} = 1, \\ & \sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} - u_o \le 0 (\forall j \ne o), v_i, u_r \ge 0 (\forall i, r), u_o : \text{free} \end{aligned}$$

Based on the results of (9.23), one can determine, respectively, the input-oriented right-hand SE ($\varepsilon_i^+(\cdot)$) and left-hand SE ($\varepsilon_i^-(\cdot)$) for firm $o$ as

$$\varepsilon_i^+(x_o, y_o) = \frac{1}{1 + u_o^+ / D_i(x_o, y_o)} \text{ and } \varepsilon_i^-(x_o, y_o) = \frac{1}{1 + u_o^- / D_i(x_o, y_o)} \tag{9.24}$$

We have now our Proposition 2.

**Proposition 2** *Assuming alternate optima in* $u_o$, the firm $o$ in *the* $T_{VRS}^{DEA}$ *exhibits* (*input-oriented*) *IRS* ($\varepsilon_i^+(\cdot) > 1$) *if* $u_o^+ < 0$, (*input-oriented*) *CRS* ($\varepsilon_i^+(\cdot) \le 1 \le \varepsilon_i^-(\cdot)$) *if* $u_o^+ \ge 0 \ge u_o^-$ *and* (*input-oriented*) *DRS* ($\varepsilon_i^-(\cdot) < 1$) *if* $u_o^- > 0$.

Let us now diagrammatically illustrate in Fig. 9.1 the input-oriented measure of right-hand and left-hand SEs. For sake of simplicity, we assume a single-input-single-output technology that comprises of six firms labelled as A, B, C, D, E and F. Firms—A, B, C and D—form the BCC efficiency frontier. Consider, e.g., the SE evaluation of two firms: one efficient firm, say, A and one inefficient firm, say E, whose input and output bundles are, respectively, (2,1) and (5,1). The running of LP program (9.23) yields the following results: $D_i(2,1) = 1$, $u_A^+ = -0.5$ and $u_A^- = -1$ (for A) and $D_i(5,1) = 2/5 = 0.4$, $u_E^+ = -0.2$ and $u_E^- = -0.4$ (for E). Using (9.24), the right-hand and left-hand SEs of efficient firm A can be computed, respectively, as $\varepsilon_i^+(2,1) = 1/[1+(-0.5)/1] = 2$ and $\varepsilon_i^-(2,1) = 1/[1+(-1)/1] = \infty$. And, the right-hand and left-hand SEs of inefficient firm E can be computed at its projected point A as $\varepsilon_i^+(5,1) = 1/[1 + (-0.2)/0.4] = 2$ and $\varepsilon_i^-(5,1) = 1/[1 + (-0.4)/0.4] = \infty$ respectively. The right-hand and left-hand SEs of other firms can be computed in an analogous manner.

In order to compute the output-oriented right-hand and left-hand SEs for firm $o$, we first find out the maximum and minimum values of $v_o$ for firm $o$ from the following LP programs:

$$v_o^+(v_o^-) = \max(\min) v_o \tag{9.25}$$

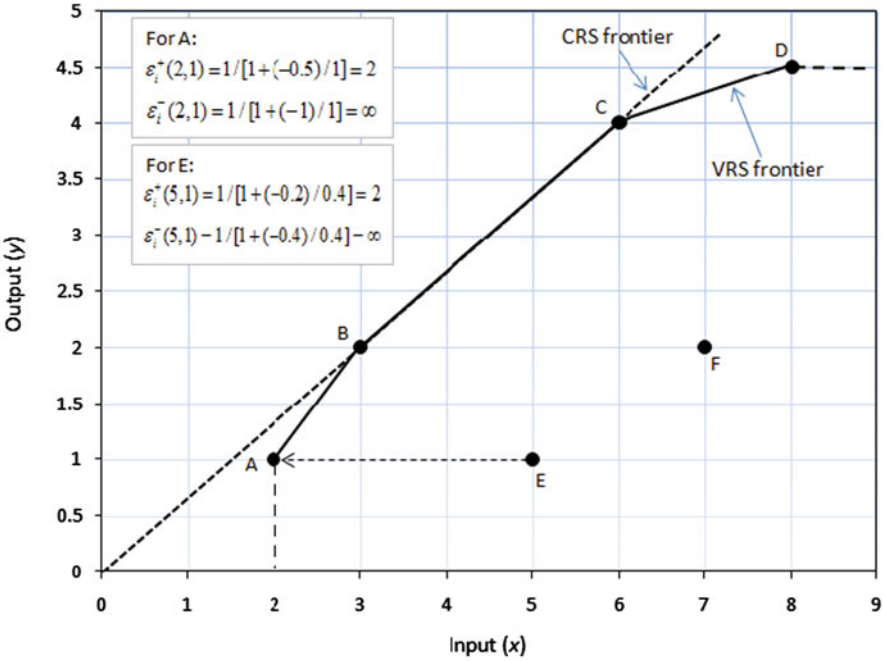



**Fig. 9.1** The diagrammatic illustration of SE measure

$$\text{s.t.} \sum_{i=1}^{m} v_i x_{io} - v_o = D_o(x_o, y_o), \sum_{r=1}^{s} u_r y_{ro} = 1,$$
$$-\sum_{r=1}^{s} u_r y_{rj} + \sum_{i=1}^{m} v_i x_{ij} + u_o \geq 0 (\forall j \neq o), v_i, u_r \geq 0 (\forall i, r), v_o : \text{free}$$

Based on the results of (9.25), the output-oriented right-hand SE $\varepsilon_o^+(x_o, y_o)$ and left-hand SE $\varepsilon_o^-(x_o, y_o)$ for firm $o$ can be computed as

$$\varepsilon_o^+(x_o, y_o) = 1 - \frac{v_o^+}{D_o(x_o, y_o)} \text{ and } \varepsilon_o^-(x_o, y_o) = 1 - \frac{v_o^-}{D_o(x_o, y_o)} \tag{9.26}$$

We have now our Proposition 3.

**Proposition 3** *Assuming alternate optima in* $v_o$, firm $o$ in *the* $T_{VRS}^{DEA}$ *exhibits* (*output-oriented*) *IRS* ($\varepsilon_o^+(\cdot) > 1$) *if* $v_o^+ < 0$, (*output-oriented*) *CRS* ($\varepsilon_o^+(\cdot) \leq 1 \leq \varepsilon_o^-(\cdot)$) *if* $v_o^+ \geq 0 \geq v_o^-$ *and* (*output-oriented*) *DRS* ($\varepsilon_o^-(\cdot) < 1$) *if* $v_o^- > 0$.

Note that Banker et al. (1984) are the first to show that the intercept $u_o(v_o)$ in the multiplier form of BCC model can be used to estimate the SE. Several contributions exist, at the extreme points, on the evaluation of right-hand (lower bound) and left-hand SE (upper bound) measures based on production model. See, e.g., Banker and Thrall (1992); Fukuyama (2000, 2001, 2003); Førsund and Hjalmarsson (2004);

Tone and Sahoo (2004); Hadjicostas and Soteriou (2006); Førsund et al. (2007), Podinovski et al. (2009); Podinovski and Førsund (2010), Zelenyuk (2013), and Sahoo and Sengupta (2014), among others.

Note that instead of using the multiplier DEA models—(9.16) and (9.20); one could also use the envelopment models—(9.15) and (9.19)—to calculate the input-oriented right-hand SE and the output-oriented left-hand SE. To start with, we assume that firm $o$ is input-oriented technically efficient, i.e., $\alpha(1) = 1$ in (9.15). In order to compute its input-oriented right-hand SE of firm $o$ $\varepsilon_i^+(x_o, y_o)$, we set up the following model:

$$D_i(x_o, y_o) = \alpha(\beta) = \min\left\{\alpha : (\alpha x, \beta y) \in T_{VRS}^{DEA}; \beta = 1 + \delta, \delta > 0\right\} \tag{9.27}$$

If $1 + \delta > \alpha(\cdot) > 1$, then IRS prevail to the right of firm $o$'s $(x_o, y_o)$; if $1 + \delta = \alpha(\cdot)$, then CRS prevail to the right of firm $o$'s $(x_o, y_o)$; and if $1 + \delta < \alpha(\cdot)$, then DRS prevail to the right of firm $o$'s $(x_o, y_o)$; and if there is no feasible solution to (9.27), then there are no data to determine RTS to the right of $(x_o, y_o)$ (Golany and Yu 1997, p. 32).

Based on the solution of program (9.27), the input-oriented right-hand SE for firm $o$ $\varepsilon_i^+(x_o, y_o)$ can be computed as

$$\left[\varepsilon_i^+(x_o, y_o)\right]^{-1} = \lim_{\beta \to 1^+} \frac{\alpha(\beta) - 1}{\beta - 1} = \frac{\alpha(1 + \delta) - 1}{\delta} \tag{9.28}$$

Now let us turn to show the computation of output-oriented left-hand SE for firm $o$ $\varepsilon_o^-(x_o, y_o)$ for which we assume firm $o$ to be output-oriented technically efficient, i.e., $\beta(1) = 1$ in (9.19). Consider the following LP program:

$$D_o(x_o, y_o) = \beta(\alpha) = \max\left\{\beta : (\alpha x, \beta y) \in T_{VRS}^{DEA}; \alpha = 1 - \delta, \delta > 0\right\} \tag{9.29}$$

If $1 > \beta(\cdot) > 1 - \delta$, then DRS prevail to the left of firm $o$'s $(x_o, y_o)$; if $\beta(\cdot) = 1 - \delta$, then CRS prevail to the left of firm $o$'s $(x_o, y_o)$; and if $\beta(\cdot) < 1 + \delta$, then IRS prevail to the left of firm $o$'s $(x_o, y_o)$; and if there is no feasible solution to (9.27), then there are no data to determine RTS to the left of $(x_o, y_o)$ (Golany and Yu, 1997, p. 32).

From the optimal solutions of program (9.29), the output-oriented left-hand SE for firm $o$ $\varepsilon_o^-(x_o, y_o)$ can be computed as

$$\varepsilon_o^-(x_o, y_o) = \lim_{\alpha \to 1^-} \frac{\beta(\alpha) - 1}{\alpha - 1} = \frac{1 - \beta(1 - \delta)}{\delta} \tag{9.30}$$

#### 9.2.4.2 Scale Elasticity in Multiplicative DEA Model

The problem underlying DEA models based on piece-wise linear technology set (9.14) is that the resulting production frontier is concave, i.e., the marginal products of factor inputs are non-increasing, which is not consistent with neoclassical production theory that posits an S-shape for the production function obeying the Frisch (1965)'s Regular Ultra Passum (RUP) Law. See Olesen and Petersen (2013)

for a discussion on how the piece-wise linear technology can be consistent with neoclassical production technologies obeying the RUP Law—the marginal product first increases but diminishing returns eventually set in. Therefore, another class of models—commonly known as multiplicative models—is developed to allow the production frontier to be concave in some region and non-concave elsewhere so as to reveal increasing, constant and decreasing marginal productivities along the frontier. Another distinct advantage of using this technology structure is that the exact estimates of SEs can be obtained. For the evaluation of SE, we selected the multiplicative model of Banker and Maindiratta (1986) that is based on the piece-wise log-linear technology $T_{VRS}^{L-DEA}$ given by

$$T_{VRS}^{L-DEA} = \left\{ \begin{array}{l} (\log x, \log y) : \sum\limits_{j \in J} \lambda_j \log x_{ij} \leq \log x_i (\forall i), \\ \sum\limits_{j \in J} \lambda_j \log y_{rj} \geq \log y_r (\forall r), \sum\limits_{j \in J} \lambda_j = 1, \lambda_j \geq 0 (\forall j) \end{array} \right\} \tag{9.31}$$

The input-oriented TE of firm $o$ can be evaluated against $T_{VRS}^{L-DEA}$ as:

$$\min \left\{ \log \delta : (\log x_o + \log \delta, \log y_o) \in T_{VRS}^{L-DEA} \right\} \tag{9.32}$$

The dual of (9.32) can be set up as

$$\begin{aligned} \max & \sum_{r=1}^{s} u_r \log y_{ro} - \sum_{i=1}^{m} v_i \log x_{io} - \hat{u}_o \\ \text{s.t.} & \sum_{r=1}^{s} u_r \log y_{rj} - \sum_{i=1}^{m} v_i \log x_{ij} - \hat{u}_o \leq 0, \\ & \sum_{i=1}^{m} v_i = 1, \\ & u_r, v_i \geq 0 (\forall r, i); \hat{u}_o : free. \end{aligned} \tag{9.33}$$

For any technically efficient firm $o$ ($o \in J$), the following transformation function

$$\psi(x_o, y_o) \equiv \sum_{r=1}^{s} u_r \log y_{ro} - \sum_{i=1}^{m} v_i \log x_{io} - \hat{u}_o = 0 \tag{9.34}$$

Using the formula (9.8) the input-oriented SE of firm $o$ $\varepsilon_i^m(x_o, y_o)$ can be obtained as

$$\varepsilon_i^m(x_o, y_o) = -\frac{\sum_{i=1}^{m} \frac{\partial \psi(x_o, y_o)}{\partial x_{io}} x_{io}}{\sum_{r=1}^{s} \frac{\partial \psi(x_o, y_o)}{\partial y_{ro}} y_{ro}} = \frac{\sum_{i=1}^{m} v_i}{\sum_{r=1}^{s} u_r} = \frac{1}{\sum_{r=1}^{s} u_r} \tag{9.35}$$

**Proposition 4** Firm $o$ in the $T_{VRS}^{L-DEA}$ exhibits (input-oriented) increasing RTS if $\sum_{r=1}^{s} u_r < 1$, (input-oriented) constant RTS if $\sum_{r=1}^{s} u_r = 1$, and (input-oriented) decreasing RTS if $\sum_{r=1}^{s} u_r < 1$.

One can also use the other multiplicative models (Banker et al. 2004; Zarepisheh et al. 2010; Mehdiloozad et al. 2014) to determine SE.

#### 9.2.4.3 Scale Elasticity in Production DEA Models with Indivisibilities

The DEA models based on the piece-wise linear technology (9.14) are based on the maintained hypothesis that the technology set is *convex*. Convexity, as argued by Farrell (1959), assumes away some important technological features such as *indivisible* production activities, *economies of scale* and *economies of specialization*, which all, in fact, result from *concavities* in production. Barring a few authors like Thrall (1999), the recent literature favoring the dropping of convexity axiom include, among others, Scarf (1981a, b, 1986, 1994); Tulkens (1993); Tulkens and Vanden Eeckaut (1995); Bouhnik et al. (2001); Tone and Sahoo (2003), Kuosmanen (2003); Briec et al. (2004); and Briec and Liang (2011). And the literature on some exciting economic analysis arising from the violation of convexity include Yang and Ng (1993); Yang (1994); Yang and Rice (1994); Borland and Yang (1995) and Shi and Yang (1995). The only argument favoring convexity postulate is that there is possible reduction of small sample errors, which, but, comes at the cost of possible specification error that is likely to be negligible in large samples.

A closer look at the DEA-related economic literature on production analysis (Afriat 1972; Hanoch and Rothschild 1972; Varian 1984) reveals that the convexity properties are motivated from the perspective of economic objectives, but not as inherent feature of technology. However, looking at the structures of real-life technologies leads one to suspect the harmless character of the convexity postulate. For the details, see Cherchye et al. (2000, 2001) and Kuosmanen (2003) who have provided empirical as well as theoretical arguments, less in favor of, but, mostly against the convexity postulate.

In the presence of indivisibilities, the technology set is no longer convex; as a result, the linear technology in (9.14) fails to enable us to correctly determine the local RTS possibilities along the frontier. Tone and Sahoo (2003) have shown that in the presence of technological indivisibilities, the use of convex BCC technology produces erroneous inferences concerning the local RTS possibilities of firms. Kerstens and Vanden Eeckaut (1999) and later on Briec et al. (2000) proposed a more general method by considering variations in RTS on the existing free disposal hull (FDH) technology that is suitable for all reference technologies - $FDH^{CRS}$, $FDH^{NIRS}$ and $FDH^{NDRS}$, where the superscripts CRS, NIRS and NDRS represent, respectively, CRS, non-increasing RTS and non-decreasing RTS. These technologies are given by

$$FDH^{CRS} = \left\{ (x,y) : \sum_{j\in J} x_j z_j \leq x, \sum_{j\in J} y_j z_j \geq y, \sum_{j\in J} \lambda_j \right.$$
$$\left. = 1, \lambda_j \in \{0,1\}, z_j = \delta\lambda_j, \delta \geq 0 \right\} \tag{9.36}$$

$$FDH^{NIRS} = \left\{ (x,y) : \sum_{j \in J} x_j z_j \leq x, \sum_{j \in J} y_j z_j \geq y, \sum_{j \in J} \lambda_j = 1, \lambda_j \in \{0,1\}, z_j = \delta \lambda_j, 0 \leq \delta \leq 1 \right\} \tag{9.37}$$

$$FDH^{NDRS} = \left\{ (x,y) : \sum_{j \in J} x_j z_j \leq x, \sum_{j \in J} y_j z_j \geq y, \sum_{j \in J} \lambda_j = 1, \lambda_j \in \{0,1\}, z_j = \delta \lambda_j, \delta \geq 1 \right\} \tag{9.38}$$

Here, $\lambda$ is the only activity operating subject to a non-convexity constraint and the re-scaled activity, $z$ allows for any scaling of the observations spanning the production frontier. The technical efficiency of any firm $o$ can be computed against these technologies as follows:

$$\rho^{FDH-CRS} = \min \left\{ \rho : (\rho x_o, y_o) \in FDH^{CRS} \right\} \tag{9.39}$$

$$\rho^{FDH-NIRS} = \min \left\{ \rho : (\rho x_o, y_o) \in FDH^{NIRS} \right\} \tag{9.40}$$

$$\rho^{FDH-NDRS} = \min \left\{ \rho : (\rho x_o, y_o) \in FDH^{NDRS} \right\} \tag{9.41}$$

**Proposition 5** The input-oriented RTS for any firm $o$ ($o \in J$) can be characterized as follows:

$$CRS \Leftrightarrow \rho^{FDH-CRS} = \max \left\{ \rho^{FDH-CRS}, \rho^{FDH-NIRS}, \rho^{FDH-NDRS} \right\}$$

$$IRS \Leftrightarrow \rho^{FDH-NDRS} = \max \left\{ \rho^{FDH-CRS}, \rho^{FDH-NIRS}, \rho^{FDH-NDRS} \right\}$$

$$DRS \Leftrightarrow \rho^{FDH-NIRS} = \max \left\{ \rho^{FDH-CRS}, \rho^{FDH-NIRS}, \rho^{FDH-NDRS} \right\}$$

Bouhnik et al. (2001) also proposed two new non-convex DEA models, referred to as Fixed-Charge models where a lower limit on permissible divisibility is established. In these models the observed firms are scaled for the construction of composite firm where the scaled firm must be at least as large as a pre-defined lower bound imposed on the envelopment intensities to ensure fair comparisons without sacrificing the required discriminating power inherent in the CRS and VRS DEA models. For the details, refer to Bouhnik et al. (2001).

#### 9.2.4.4 Scale Elasticity in Cost DEA Models

Sueyoshi (1997) is the first who measured scale elasticity using the cost efficiency DEA model of Färe et al. (1985) given by

$$C(y_o; w_o) = \min_{x,\lambda} \sum_{i=1}^{m} w_{io} x_i \tag{9.42}$$

$$\text{s.t.} \sum_{j \in J} x_{ij} \lambda_j \leq x_i (\forall i), \sum_{j \in J} y_{rj} \lambda_j \geq y_{ro} (\forall r), \sum_{j \in J} \lambda_j = 1, \lambda_j \geq 0.$$

Here, the cost efficiency of firm $o$ ($CE_o$) is defined as the ratio of *minimum* cost $C(y_o; w_o)$ over *actual* cost $c_o$, i.e.,

$$CE_o = C(y_o; w_o)/c_o = \sum\nolimits_{i=1}^{m} w_{io} x_i^* \Big/ \sum\nolimits_{i=1}^{m} w_{io} x_{io} \tag{9.43}$$

where $x_i^*$ is the optimal solution to (9.42). The following dual LP program of (9.42) can be used to compute the SE of firm $o$, $\varepsilon_c(y, w)$ as follows:

$$C(y_o; w_o) = \max_{u,\omega} \sum_{r=1}^{s} u_r y_{ro} - \omega_o$$

$$s.t. \sum_{r=1}^{s} u_r y_{rj} - \omega_o \leq c_j, (\forall j), u_r \geq 0 \ (\forall r), \omega_o : \text{free} \tag{9.44}$$

If firm $o$ is efficient, the minimum and actual costs are both the same, and it also holds that

$$c_o = C(y_o; w_o) = \sum_{r=1}^{s} u_r^* y_{ro} - \omega_o^* \tag{9.45}$$

Following Panzar and Willig (1977), $\varepsilon_c(y; w)$ can then be obtained as

$$\varepsilon_c(y; w) = \frac{C(y_o; w_o)}{\sum_{r=1}^{s} y_{ro} \frac{\partial C(y_o; w_o)}{\partial y_{ro}}} = \frac{C(y_o; w_o)}{\sum_{r=1}^{s} u_r^* y_{ro}} = \frac{1}{1 + [\omega_o^* / C(y_o; w_o)]} \tag{9.46}$$

**Proposition 5** Firm $o$ in *the* $T_{VRS}^{DEA}$ exhibits increasing RTS if $\omega_o^* < 0$ in all optimal solutions, constant RTS if $\omega_o^* = 0$ in an optimal solution, and decreasing RTS if $\omega_o^* > 0$ in all optimal solutions.

Since there are multiple optima in $\omega_o^*$, following Banker and Thrall (1992), we set up the following LP problems to find out the max. and min. values of $\omega_o$ for firm $o$ as:

$$\omega_o^+(\omega_o^-) = \max(\min)\ \omega_o \tag{9.47}$$

$$\text{s.t.} \quad \sum_{r=1}^{s} u_r y_{ro} - \omega_o = C(y_o; w_o),$$

$$\sum_{r=1}^{s} u_r y_{rj} - u_o \leq c_j (\forall j \neq o), v_i, u_r \geq 0\ (\forall i, r), u_o : \text{free}$$

Using the results of (9.47), one can determine the input-oriented right-hand SE ($\varepsilon_c^+(\cdot)$) and left-hand SE ($\varepsilon_c^-(\cdot)$) for firm $o$ as

$$\varepsilon_c^+(y_o; w_o) = \frac{1}{1 + \omega_o^+ / C(y_o; w_o)} \text{and } \varepsilon_c^-(y_o; w_o) = \frac{1}{1 + \omega_o^- / C(y_o; w_o)} \tag{9.48}$$

**Proposition 6** *Assuming alternate optima in* $\omega_o$, firm $o$ in *the* $T_{VRS}^{DEA}$ *exhibits IRS* ($\varepsilon_c^+(\cdot) > 1$) *if* $\omega_o^+ < 0$, *CRS* ($\varepsilon_c^+(\cdot) \leq 1 \leq \varepsilon_c^-(\cdot)$) *if* $\omega_o^+ \geq 0 \geq \omega_o^-$ *and DRS* ($\varepsilon_c^-(\cdot) < 1$) *if* $\omega_o^- > 0$.

Note that the CE model (9.42), which is based on a *factor-based technology set* (9.14), can be of limited use in actual applications when market imperfections exist (Camanho and Dyson 2008; Sahoo and Tone 2013; Sahoo et al. 2014a). This is because this model is based on a number of simplifying assumptions, which hardly hold in practice. First, the inputs are assumed to be homogeneous across firms; and their prices are also assumed to be exogenously given, and are, measured and known, with full certainty by firms. In real-life applications, however, when production is expanded, firms experience *changes* in the organization of their processes or in the characteristics of their inputs that are economically more attractive than the replicated alternatives of those already in use. That is, the techniques and inputs used at higher scale are very different from those used at lower scale. Hence, the inputs are heterogeneous across firms. Since the inputs available at the firm levels vary in their quality, the construction of *factor-based* linear *technology set* (9.14) becomes problematic.

Furthermore, even if the inputs are homogenous across firms, in many cases their prices cannot be measured accurately enough to make use of CE measurement. This is because accounting data can give a poor approximation for economic prices, i.e., marginal opportunity costs due to debatable valuation and depreciation schemes. The input prices are not exogenous, but they vary according to the actions by firms (Chamberlin 1933; Robinson 1933; Engel and Rogers 1996). Also, firms often face *ex ante* price uncertainty while making their production decisions (McCall 1967; Sandmo 1971; Camanho and Dyson 2005). Economic theory suggests that the firms enjoying some degree of monopoly power should charge different prices if there is heterogeneity in productivity of their inputs. This is empirically valid since most firms are observed facing upward slopping supply curve in their input purchase decisions. This observation also suggests that the assumption of facing common unit prices by firms, i.e., the law of one price, which has long been maintained as a necessary and

sufficient condition for Pareto efficiency in competitive markets (Kuosmanen et al. 2006), is not at all justified in revealing the proper CE and SE behavior of firms.

The CE model (9.42) can also be of limited value in actual applications even when the inputs are homogeneous across firms. This is because as pointed out by Camanho and Dyson (2008), the CE measure, as defined in (9.43), reflects only input inefficiencies (technical inefficiency and/or allocative efficiency) but not market (price) inefficiencies (deviation from fully competitive setting leading to price differences between firms). Therefore, as a remedy, they suggested using a more comprehensive scheme to measure CE that can be attributed to both the input inefficiencies and the market (price) inefficiencies. Note that this problem was also alternatively addressed by Cross and Färe (2008) who attributed the value-based technical efficiency (technical efficiency obtained from inputs measured in value terms) to three sources: (a) quantity-based technical efficiency (technical efficiency obtained from the inputs measured in physical quantities), (b) technology effect, and (c) firm effect.

While it is true that the measure of scale economies, as defined in (9.46), involves the cost effects of output expansion with input prices held constant in perfectly competitive market structure, at the same one cannot deny the possibility of linking scale economies with further cost reduction due to other sources, i.e., pecuniary economies. Hence, it can be argued that the above measure of scale economies is not comprehensive in actual applications when imperfections exist. Therefore, when the inputs are heterogeneous, in order to account for situation where the input prices vary between firms as a result of negotiations or to reflect the qualitative differences in the resources, the alternative value-based CE model of Tone (2002) should be followed by setting up the technology set in cost-output space. This alternative CE model was further extended to directional DEA of Chambers et al. (1996, 1998) by Fukuyama and Weber (2004), Färe and Grosskopf (2006) and Sahoo et al. (2014a) to develop the directional value-based measures of technical inefficiencies.

#### 9.2.4.5 Scale Elasticity in Alternative Cost DEA Model

Let us now turn to discuss the alternative value-based CE model of Tone (2002) in estimating the scale economies behavior of firms. This alternative CE model is based on a value-based technology set in the cost-output space given by

$$T_{VRS}^{C-DEA} = \left\{ (c, y) : \sum_{j \in J} c_j \lambda_j \leq c, \sum_{j \in J} y_{rj} \lambda_j \geq y_r (\forall r), \sum_{j \in J} \lambda_j = 1, \lambda_j \geq 0 (\forall j) \right\}, \tag{9.49}$$

where $c_j = \sum_{i=1}^{m} w_{ij} x_{ij}$. The value-based technical efficiency of firm $o$ can be obtained from the following LP program:

$$\theta^* = \min \left\{ \theta : (\theta c_o, y_o) \in T_{VRS}^{C-DEA} \right\} \tag{9.50}$$

In order to compute the SE of firm $o$, we consider the dual of (9.50):

$$\theta^* = \max \sum_{r=1}^{s} u_r y_{ro} - \bar{\omega}_o \tag{9.51}$$

$$\text{s.t.} \sum_{r=1}^{s} u_r y_{rj} - v c_j - \bar{\omega}_o \leq 0,$$

$$v c_o = 1,$$

$$v \geq 0, u_r \geq 0 \ (\forall r); \bar{\omega}_o : free.$$

For any firm $o$ ($o \in J$), the following transformation function

$$\psi(\theta^* c_o, y_o) \equiv \sum_{r=1}^{s} u_r y_{ro} - v(\theta^* c_o) - \bar{\omega}_o = 0 \tag{9.52}$$

Using the SE formula (9.8) the input-oriented SE of firm $o$ can be obtained as

$$\varepsilon_i(c_o, y_o) = -\frac{\frac{\partial \psi(\theta^* c_o, y_o)}{\partial c_o} c_o}{\sum_{r=1}^{s} \frac{\partial \psi(\theta^* c_o, y_o)}{\partial y_{ro}} y_{ro}} = \frac{1}{1 + \bar{\omega}_o / \theta^*} \tag{9.53}$$

**Proposition 7** Firm $o$ exhibit increasing RTS ($\varepsilon_i(c_o, y_o) > 1$) if $\bar{\omega}_o < 0$ in all optimal solutions; constant RTS ($\varepsilon_i(c_o, y_o) = 1$) if $\bar{\omega}_o = 0$ in an optimal solution; and decreasing RTS ($\varepsilon_i(c_o, y_o) < 1$) if $\bar{\omega}_o > 0$ in all optimal solutions.

Assuming multiple optima in $\bar{\omega}_o$, we set up the following LP problems to find out the maximum and minimum values of $\bar{\omega}_o$ for firm $o$ as:

$$\bar{\omega}_o^+ (\bar{\omega}_o^-) = \max(\min) \bar{\omega}_o \tag{9.54}$$

$$\text{s.t.} \sum_{r=1}^{s} u_r y_{ro} - \bar{\omega}_o = \theta^*$$

$$\sum_{r=1}^{s} u_r y_{rj} - v c_j - \bar{\omega}_o \leq 0 \ (\forall j \neq o),$$

$$v c_o = 1,$$

$$v \geq 0, u_r \geq 0 \ (\forall r).$$

Using the results of (9.54), the input-oriented right-handand left-hand SEs of firm $o$ can be obtained as

$$\varepsilon_i^+(c_o, y_o) = \frac{1}{1 + \bar{\omega}_o^+ / \theta^*} \quad \text{and} \quad \varepsilon_i^-(c_o, y_o) = \frac{1}{1 + \bar{\omega}_o^- / \theta^*} \tag{9.55}$$

**Proposition 8** *Assuming alternate optima in* $\bar{\omega}_o$, the firm $o$ in *the* $T_{VRS}^{C-DEA}$ *exhibits* (*input-oriented*) *IRS* $(\varepsilon_i^+(c_o, y_o) > 1)$ *if* $\bar{\omega}_o^+ < 0$, (*input-oriented*) *CRS* $(\varepsilon_i^+(c_o, y_o) \leq 1 \leq \varepsilon_i^-(c_o, y_o))$ *if* $\bar{\omega}_o^+ \geq 0 \geq \bar{\omega}_o^-$ *and* (*input-oriented*) *DRS* $(\varepsilon_i^-(c_o, y_o) < 1)$ *if* $\bar{\omega}_o^- > 0$.

See, Tone and Sahoo (2005, 2006); Sengupta and Sahoo (2006); Sahoo et al. (2007); Sahoo and Gstach (2011); Sahoo et al. (2012) and Sahoo and Tone (2013) on an elaborate discussion of the estimation of RTS based on the alternative CE model.

## 9.3 Concluding Remarks

Various models for the empirical evaluation of SE are critically reviewed in non-parametric DEA approach. This nonparametric approach is classified into two: production approach and cost approach. In the former three types of technology structure—piece-wise linear, piece-wise log-linear and FDH—are employed. In the latter, the piece-wise linear technology is employed in two environment—one in input-output space and the other in cost-output space.The SE estimates based on the piece-wise linear technology in production environment can be biased upward or downward when the technology involves some indivisibilities. Furthermore, the SE estimates of firms are not unique in this technology. In this scenario the SE estimates based on the piece-wise log-linear and FDH technologies are argued to be preferred. Between the two cost models, the cost model defined in cost-output space is to be preferred to the one defined in input-output space on two grounds: (1) the price data are often not available, or, if available, are not well measured due to debatable valuation and depreciation schemes, and (2) the factor inputs are heterogeneous across real-life firms.

The DEA models discussed in this study for the evaluation of SE treat production technology as a black box. It is, however, possible that the idea underlying these models could be used to compute the SE in network technologies that will enable researchers in locating the sources of increasing returns of a network firm in the sub-technologies. We consider this as avenue for future research, which we have addressed elsewhere in Sahoo et al. (2014b).

## References

Afriat S (1972) Efficiency estimation of production functions. Int Econ Rev 13:568–598

Banker RD, Maindiratta A (1986) Piecewise loglinear estimation of efficient production surfaces. Manage Sci 32:126–135

Banker RD, Thrall RM (1992) Estimation of returns to scale using data envelopment analysis. Eur J Oper Res 62:74–84

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manage Sci 30:1078–1092

Banker RD, Conrad RF, Straus RP (1986) A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production. Manage Sci 32:30–44

Banker RD, Bardhan I, Cooper WW (1996a) A note on returns to scale in DEA. Eur J Oper Res 88:583–585
Banker RD, Chang H, Cooper WW (1996b) Equivalence and implementation of alternative methods for determining returns to scale in data envelopment analysis. Eur J Oper Res 89:473–481
Banker RD, Cooper WW, Seiford LM, Thrall RM, Zhu J (2004) Returns to scale in different DEA models. Eur J Oper Res 154:345–362
Basu S, Fernald JG (1997) Returns to scale in us production: estimates and implications. J Political Econ 105:249–283
Baumol WJ, Panzar JC, Willig RD (1982) Contestable markets and theory of industrial structure. Harcourt Brace Jovanovich, New York
Borland J, Yang X (1995) Specialization, product development, evolution of the institution of the firm, and economic growth. J Evolut Econ 5:19–42
Bouhnik S, Golany B, Passy S, Hackman ST, Vlatsa DA (2001) Lower bound restrictions on intensities in data envelopment analysis. J Prod Anal 16:241–261
Briec W, Liang QB (2011) On some semilattice structures for production technologies. Eur J Oper Res 215:740–749
Briec W, Kerstens K, Leleu H, Vanden Eeckaut P (2000) Returns to scale on nonparametric deterministic technologies: simplifying goodness-of-fit methods using operationson technologies. J Prod Anal 14:267–274
Briec W, Kerstens K, Vanden Eeckaut P (2004) Non-convex technologies and cost functions: definitions, duality and nonparametric tests of convexity. J Econ 81:155–192
Camanho AS, Dyson RG (2005) Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. Eur J Oper Res 161:432–446
Camanho AS, Dyson RG (2008) A generalization of the Farrell cost efficiency measure applicable to non-fully competitive settings. Omega 36:147–162
Chamberlin EH (1933) The theory of monopolistic competition: a reorientation of the theory of value. Harvard University Press, Cambridge
Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. J Econ Theory 70:407–419
Chambers RG, Chung Y, Färe R (1998) Profit, directional distance functions, and Nerlovian efficiency. J Optim Theory Appl 98:351–364
Cherchye L, Kuosmanen T, Post T (2000) "Why convexify? An assessment of convexity axioms in DEA", Helsinki School of Economics and Business Administration Working Paper No. W-270
Cherchye L, Kuosmanen T, Post T (2001) What is the economic meaning of FDH? A reply to Thrall. J Prod Anal 13:259–263
Cross R, Färe R (2008) Farrell efficiency under value and quantity data. J Prod Anal 29:193–199
Devereux MB, Head AC, Lapham BJ (1996) Monopolistic competition, increasing returns, and the effects of government spending. J Money Credit Bank 28:233–254
Engel C, Rogers JH (1996) How wide is the better? Am Econ Rev 86:1112–1125
Färe R (1988) Fundamentals of production theory, lecture notes in economics and mathematical systems, vol 311. Springer, Berlin
Färe R, Grosskopf S (1985) A non-parametric cost approach to scale efficiency. Scand J Econ 87:594–604
Färe R, Grosskopf S (2006) Resolving a strange case of efficiency. J Oper Res Soc 57:1366–1368
Färe R, Primont D (1995) Multi-output production and duality: theory and application. Kluwer Academic Press, Boston
Färe R, Grosskopf S, Lovell CAK (1985) The measurement of efficiency of production. Kluwer-Nijhoff Publishing, Boston
Färe R, Grosskopf S, Lovell CAK (1986) Scale economies and duality. J Econ 46:175–182
Färe R, Grosskopf S, Lovell CAK (1988) Scale elasticity and scale efficiency. J Inst Theor Econ 144:721–729
Farrell MJ (1959) Convexity assumptions in theory of competitive markets. J Polit Econ 67:377–391
Førsund FR (1996) On the calculation of scale elasticity in DEA models. J Prod Anal 7:283–302

Førsund FR, Hjalmarsson L (2004) Calculating scale elasticity in DEA Models. J Oper Res Soc 55:1023–1038
Førsund FR, Hjalmarsson L, Krivonozhko V, Utkin OB (2007) Calculation of scale elasticities in DEA models: direct and indirect approaches. J Prod Anal 28:45–56
Fried HO, Schmidt SS, Yaisawarng S (1998) Productive, scale and scope efficiencies in U.S. hospital-based nursing homes. INFOR 36:103–119
Frisch R (1965) Theory of production. D. Reidel Publishing Company, Dordrecht
Fukuyama H (2000) Returns to scale and scale elasticity in data envelopment analysis. Eur J Oper Res 125:93–112
Fukuyama H (2001) Returns to scale and scale elasticity in Farrell, Russell and additive models. J Prod Anal 16:225–239
Fukuyama H (2003) Scale characterizations in a DEA directional technology distance function framework. Eur J Oper Res 144:108–127
Fukuyama H, Weber WL (2004) Economic inefficiency measurement of input spending when decision-making units face different input prices. J Oper Res Soc 55:1102–1110
Golany B, Yu G (1997) Estimating returns to scale in DEA. Eur J Oper Res 103:28–37
Grosskopf S, Yaisawarng S (1990) Economies of scope in the provision of local public services. Nat Tax J 43:61–74
Grosskopf S, Hayes K, Yaisawarng S (1987) Measuring economies of scope in farming: two alternative approaches, Southern Illinois University at Carbondale Discussion Paper, No. 87-14
Hadjicostas P, Soteriou AC (2006) One-sided elasticities and technical efficiency in multi-output production: a theoretical framework. Eur J Oper Res 168:425–449
Hanoch G (1970) Homotheticity in joint production. J Econ Theory 2:423–426
Hanoch G, Rothschild M (1972) Testing assumptions in production theory: a nonparametric approach. J Polit Econ 80:256–275
Jones CI (2004) Growth and ideas. National Bureau of Economic Research, Cambridge
Kerstens K, Vanden Eeckaut P (1999) Estimating returns to sale using non-parametric technologies: a new method based on goodness-of-fit. Eur J Oper Res 113:206–214
Kuosmanen T (2003) Duality theory of non-convex technologies. J Prod Anal 20:273–304
Kuosmanen T, Cherchye L, Sipilainen T (2006) The law of one price in data envelopment analysis: restricting weight flexibility across firms. Eur J Oper Res 170:735–757
McCall JJ (1967) Competitive production for constant risk utility functions. Rev Econ Stud 34:417–420
Mehdiloozad M, Sahoo BK, Roshdi I (2014) A generalized multiplicative directional distance function for efficiency measurement in DEA. Eur J Oper Res 232(3):679–688
Morrison CJ (1992) Unraveling the productivity growth slowdown in the United States, Canada and Japan: the effects of subequilibrium, scale economies and markups. Rev Econ Statist 74:381–393
Olesen OB, Petersen NC (2013) Imposing the regular Ultra Passum Law in DEA models. Omega 41:16–27
Panzar JC, Willig RD (1977) Economies of scale in multi-output production. Q J Econ XLI:481–493
Podinovski VV, Førsund FR (2010) Differential characteristics of efficient frontiers in data envelopment analysis. Oper Res 58:1743–1754
Podinovski VV, Førsund FR, Krivonozhko VE (2009) A simple derivation of scale elasticity in data envelopment analysis. Eur J Oper Res 197:149–153
Robinson J (1933) The economics of imperfect competition. Macmillan, London
Rosenberg N (1963) Technological change in the machine tool industry, 1840–1910. J Econ Hist 23:414–443
Rosenberg N (1981) Why in America? In: Mayr O, Post RC (eds) Yankee Enterprise, the rise of the American system of manufactures. Smithsonian Institution, Washington DC
Sahoo BK (2008) A non-parametric approach to measuring short-run expansion path. J Quant Econ 6:137–150
Sahoo BK, Gstach D (2011) Scale economies in Indian commercial banking sector: evidence from DEA and translog estimates. Int J Inf Syst Soc Change 2:13–30

Sahoo BK, Tone K (2009a) Decomposing capacity utilization in data envelopment analysis: an application to banks in India. Eur J Oper Res 195:575–594
Sahoo BK, Tone K (2009b) Radial and non-radial decompositions of profit change: with an application to Indian banking. Eur J Oper Res 196:1130–1146
Sahoo BK, Tone K (2013) Non-parametric measurement of economies of scale and scope in non-competitive environment with price uncertainty. Omega 41:97–111
Sahoo BK, Mohapatra PKJ, Trivedi ML (1999) A comparative application of data envelopment analysis and frontier translog production function for estimating returns to scale and efficiencies. Int J Syst Sci 30:379–394
Sahoo BK, Sengupta JK, Mandal A (2007) Productive performance evaluation of the banking sector in India using data envelopment analysis. Int J Oper Res 4:1–17
Sahoo BK, Kerstens K, Tone K (2012) Returns to growth in a non-parametric DEA approach. Int Trans Oper Res 19:463–486
Sahoo BK, Sengupta JK (2014) Neoclassical characterization of returns to scale in nonparametric production analysis. J Quant Econ 12(1):77–85
Sahoo BK, Mehdiloozad M, Tone K (2014a) Cost, revenue and profit efficiency measurement in DEA: a directional distance function approach. Eur J Oper Res. doi:10.1016/j.ejor.2014.02.017
Sahoo BK, Zhu J, Tone K, Klemen BM (2014b) Decomposing technical efficiency and scale elasticity in two-stage network DEA. Eur J Oper Res 233(3):584–594
Sandmo A (1971) On the theory of the competitive firm under price uncertainty. Am Econ Rev 61:65–73
Scarf HE (1981a) Production sets with indivisibilities part I: generalities. Econometrica 49:1–32
Scarf HE (1981b) Production sets with indivisibilities part II: the case of two activities. Econometrica 49:395–423
Scarf HE (1986) Neighborhood systems for production sets with indivisibilities. Econometrica 54:507–532
Scarf HE (1994) The allocation of resources in the presence of indivisibilities. J Econ Perspect 8:111–128
Seiford LM, Zhu J (1998) On alternative optimal solutions in the estimation of returns to scale in DEA. Eur J Oper Res 108:149–152
Seiford LM, Zhu J (1999) An investigation of returns to scale under data envelopment analysis. Omega 27:1–11
Sengupta JK (1999) A dynamic efficiency model using data envelopment analysis. Int J Prod Econ 62:209–218
Sengupta JK (2002) Economics of efficiency measurement by the DEA approach. Appl Econ 34:1133–1139
Sengupta JK (2003) New efficiency theory: with application of data envelopment analysis. Springer, New York
Sengupta JK (2004a) Estimating technical change by nonparametric methods. Appl Econ 36:413–420
Sengupta JK (2004b) The survivor technique and the cost frontier: a nonparametric approach. Int J Product Econ 87:185–193
Sengupta JK (2005a) Nonparametric efficiency analysis under uncertainty using data envelopment analysis. Int J Prod Econ 95:39–49
Sengupta JK (2005b) Data envelopment analysis with heterogeneous data: an application. J Oper Res Soc 56:676–686
Sengupta JK, Sahoo BK (2006) Efficiency models in data envelopment analysis: techniques of evaluation of productivity of firms in a growing economy. Palgrave Macmillan, London
Shi H, Yang X (1995) A new theory of industrialization. J Compar Econ 20:171–189
Silberston ZA (1972) Economies of scale in theory and practice. Econ J 82:369–391
Sokoloff KL (1988) Inventive activity in early industrial America: evidence from patent records. J Econ Hist 48:813–850

Starrett DA (1977) Measuring returns to scale in the aggregate and the scale effect of public goods. Econometrica 45:1439–1455
Sueyoshi T (1997) Measuring efficiencies and returns to scale in nippon telegraph and telephone in production and cost analysis. Manage Sci 43:779–796
Sueyoshi T (1999) DEA duality on Returns to Scale (RTS) in production and cost analyses: an occurrence of multiple solutions and differences between production-based and cost-based RTS estimates. Manage Sci 45:1593–1608
Sueyoshi T, Sekitani K (2007a) Measurement of returns to scale using non-radial DEA model: a range-adjusted measure approach. Eur J Oper Res 176:1918–1946
Sueyoshi T, Sekitani K (2007b) The measurement of returns to scale under a simultaneous occurrence of multiple solutions in a reference set and a supporting hyperplane. Eur J Oper Res 181:549–570
Thrall RM (1999) What is the economic meaning of FDH? J Prod Anal 11:243–250
Tone K (2001) On returns to scale under weight restrictions in data envelopment analysis. J Productiv Anal 16:31–47
Tone K (2002) A strange case of cost and allocative efficiencies in DEA. J Oper Res Soc 53:1225–1231
Tone K, Sahoo BK (2003) Scale, indivisibilities and production function in data envelopment analysis. Int J Prod Econ 84:165–192
Tone K, Sahoo BK (2004) Degree of scale economies and congestion: a unified DEA approach. Eur J Oper Res 158:755–772
Tone K, Sahoo BK (2005) Evaluating cost efficiency and returns to scale in the life insurance corporation of india using data envelopment analysis. Socio-Econ Plan Sci 39:261–285
Tone K, Sahoo BK (2006) Re-examining scale elasticity in DEA. Ann Oper Res 145:69–87
Tulkens H (1993) On FDH analysis: some methodological issues and applications to retail banking, courts and urban transit. J Prod Anal 4:183–210
Tulkens H, Vanden Eeckaut P (1995) Non-parametric efficiency, progress and regress measures for panel data: methodological aspects. Eur J Oper Res 80:474–499
Varian HR (1984) The nonparametric approach to production analysis. Econometrica 52:279–297
Yang X (1994) Endogeneous vs. exogeneous comparative advantage and economics of specialization vs. economies of scale. J Econ 60:29–54
Yang X, Ng YK (1993) Specialization and economic organization. Elsevier Science, Amsterdam
Yang X, Rice R (1994) An equilibrium model endogenizing the emergence of a dual structure between the urban and rural sectors. J Urban Econ 35:346–368
Zarepisheh M, Khorram E, Jahanshahloo GR (2010) Returns to scale in multiplicative models in data envelopment analysis. Ann Oper Res 173:195–206
Zelenyuk V (2013) A scale elasticity measures for directional distance function and its dual: theory and DEA estimation. Eur J Oper Res 228(3):592–600
Zhu J (2000) Setting scale efficient targets in DEA via returns to scale estimation methods. J Oper Res Soc 51:376–378
Zhu J, Shen Z-H (1995) A discussion of testing DMUs' returns to scale. Eur J Oper Res 81:590–596

# Chapter 10
# DEA Based Benchmarking Models

**Joe Zhu**

**Abstract** Data envelopment analysis (DEA) is a methodology for identifying the efficient or best-practice frontier of decision making units (DMUs). It is required that all DMUs under consideration be evaluated against each other in a same pool. Adding or deleting an inefficient DMU does not alter the efficient frontier and the efficiencies of the existing DMUs. The inefficiency scores change only if the efficient frontier is altered. Benchmarking is the process of comparing a DMU's performance to the best practices formed by a set of DMUs. DEA is also called "balanced benchmarking", because DEA considers multiple performance metrics in a single model. Under such a notion, the best practices are the benchmarks identified by DEA. However, in a more general sense, best practices do not have to be identified by DEA—they can be existing "standards". This chapter presents two DEA-based benchmarking approaches where one set of DMUs is compared (or benchmarked) against another. One approach is called "context-dependent" DEA where a set of DMUs is evaluated against a particular evaluation context. Each evaluation context represents an efficient frontier composed by DMUs in a specific performance level. The context-dependent DEA measures the attractiveness and the progress when DMUs exhibiting poorer and better performance are chosen as the evaluation context, respectively. The other approach consists of a fixed benchmark model and a variable benchmark model where each (new) DMU is evaluated against a set of given benchmarks (standards).

**Keywords** Data Envelopment Analysis (DEA) · Attractiveness · Progress · Best practice · Context-dependent · Benchmarking

## 10.1 Introduction

Data envelopment analysis (DEA) uses the linear programming technique to evaluate the relative efficiency of decision making units (DMUs) with multiple performance metrics. These performance metrics are classified as DEA outputs and inputs. DEA classifies a set of DMUs into a set of efficient DMUs which form a best-practice

J. Zhu (✉)
School of Business, Worcester Polytechnic Institute, 01609 Worcester, MA, USA
e-mail: jzhu@wpi.edu

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_10

frontier and a set of inefficient DMUs. Adding or deleting an inefficient DMU does not alter the efficient frontier and the efficiencies of the existing DMUs. The inefficiency scores change only if the efficient frontier is altered. The performance of DMUs depends only on the identified efficient frontier characterized by the DMUs with a unity efficiency score.

If the performance of inefficient DMUs deteriorates or improves, the efficient DMUs still may have a unity efficiency score. Although the performance of inefficient DMUs depends on the efficient DMUs, efficient DMUs are only characterized by a unity efficiency score. The performance of efficient DMUs is not influenced by the presence of inefficient DMUs, once the DEA frontier is identified.

In this sense, all DMUs under consideration are being benchmarked against the "identified" DEA efficient frontier or best practice. Note that the best practices are part of the DMUs under evaluation. In other words, DEA simultaneously identifies the best practices and measures the performance of under-performing DMUs. As such, DEA is called "balanced benchmarking" where multiple performance metrics are integrated in a single model (Sherman and Zhu 2013).

However, benchmarking can refer to a situation where a set of DMUs is compared to a set of given standards or DMUs. The setup in the conventional DEA does not allow such benchmarking to be performed using DEA. There are two DEA-based approaches that benchmark DMUs against a given set of standards represented by a set of DMUs.

One approach is called "context-dependent" DEA (Seifrod and Zhu 2003) where a set of DMUs is evaluated against a particular evaluation context. Each evaluation context represents an efficient frontier composed by DMUs in a specific performance level. The context-dependent DEA measures the attractiveness and the progress when DMUs exhibiting poorer and better performance are chosen as the evaluation context, respectively.

The other approach consists of a fixed benchmark model and a variable benchmark model where each (new) DMU is evaluated against a set of given benchmarks (standards) (Cook et al. 2004).

## 10.2 Context-Dependent Data Envelopment Analysis

Performance evaluation is often influenced by the context. A DMU's performance will appear more attractive against a background of less attractive alternatives and less attractive when compared to more attractive alternatives. Researchers of the consumer choice theory point out that consumer choice is often influenced by the context. e.g., a circle appears large when surrounded by small circles and small when surrounded by larger ones. Similarly, a product may appear attractive against a background of less attractive alternatives and unattractive when compared to more attractive alternatives (Tversky and Simonson 1993).

Considering this influence within the framework of DEA, one could ask "what is the relative attractiveness of a particular DMU when compared to others?" As in

Tversky and Simonson (1993), one agrees that the relative attractiveness of DMU$x$ compared to DMU$y$ depends on the presence or absence of a third option, say DMU$z$ (or a group of DMUs). Relative attractiveness depends on the evaluation context constructed from alternative options (or DMUs).

In fact, a set of DMUs can be divided into different levels of efficient frontiers. If we remove the (original) efficient frontier, then the remaining (inefficient) DMUs will form a new second-level efficient frontier. If we remove this new second-level efficient frontier, a third-level efficient frontier is formed, and so on, until no DMU is left. Each such efficient frontier provides an evaluation context for measuring the relative attractiveness. e.g., the second-level efficient frontier serves as the evaluation context for measuring the relative attractiveness of the DMUs located on the first-level (original) efficient frontier. On the other hand, we can measure the performance of DMUs on the third-level efficient frontier with respect to the first or second level efficient frontier.

The context-dependent DEA (Seiford and Zhu 2003) is introduced to measure the relative attractiveness of a particular DMU when compared to others. Relative attractiveness depends on the evaluation context constructed from a set of different DMUs.

The context-dependent DEA is a significant extension to the original DEA approach. The original DEA approach evaluates each DMU against a set of efficient DMUs and cannot identify which efficient DMU is a better option with respect to the inefficient DMU. This is because all efficient DMUs have an efficiency score of one. Although one can use the super-efficiency DEA model (Andersen and Petersen 1993; Seiford and Zhu 1999b) to rank the performance of efficient DMUs, the evaluation context changes in the evaluation of each efficient DMU, and the efficient DMUs are not evaluated against the same reference set.

In the context-dependent DEA, the evaluation contexts are obtained by partitioning a set of DMUs into several levels of efficient frontiers. Each efficient frontier provides an evaluation context for measuring relative attractiveness and progress. When DMUs in a specific level are viewed as having equal performance, the attractiveness measure allows us to differentiate the "equal performance" based upon the same specific evaluation context. A combined use of attractiveness and progress measures can further characterize the performance of DMUs.

Context-dependent DEA has been used for the ranking and benchmarking of the Asian Games achievements (Wu et al. 2013). Lu and Lo (2012) construct the China regions' benchmark-learning ladders for those inefficient regions to improve progressively and to identify real benchmark for those efficient regions to rank ascendant by incorporating the stratification DEA method, attractiveness measure, and progress measure.

Chiu and Wu (2010) adopt the context-dependent DEA model to analyze the operating efficiencies of 49 international tourism hotels in Taiwan from 2004 through 2006. Ulucan and Atici (2010) evaluate the efficiency of a World Bank supported Social Risk Mitigation Project in Turkey through context-dependent DEA. Yang et al. (2007) use context-dependent DEA to explore the operating efficiency and

the benchmark-learning roadmap of military retail stores for Taiwan's General Welfare Service Ministry. Chen et al. (2005) also provide an illustrative application to measuring the performance of Tokyo public libraries.

Context-dependent DEA has been extended to use with cross efficiency (Lim 2012). Lu and Hung (2008) propose an alternative context-dependent DEA technique to explore the managerial performance and the benchmarks of 24 global leading telecom operators. Tsang and Chen (2013) present a revised context-dependent DEA model to identify multilevel strategic groups in the case of International Tourist Hotels in Taiwan. Brissimis and Zervopoulos (2012) develop a step-by-step effectiveness assessment model for customer-oriented service organizations based upon the context-dependent DEA.

### *10.2.1 Stratification DEA Model*

The first step in the context-dependent DEA is to identify the performance levels or contexts. Assume that there are $n$ DMUs which have $s$ outputs and $m$ inputs. We define the set of all DMUs as $J^1$ and the set of efficient DMUs in $J^1$ as $E^1$. Then the sequences of $J^l$ and $E^l$ are defined interactively as $J^{l+1} = J^l - E^l$. The set of $E^l$ can be found as the DMUs with optimal value $\phi_k^l$ of 1 to the following linear programming problem:

$$\begin{aligned} \underset{\lambda,\theta}{\text{minimize}} \quad & \theta_k^l = \theta \\ \text{subject to} \quad & \sum_{j \in J^l} \lambda_j x_{ij} \le \theta x_{ik}, i = 1, \ldots, m \\ & \sum_{j \in J^l} \lambda_j y_{rj} \ge y_{rk}, r = 1, \ldots, s \\ & \lambda_j \ge 0, j \in J^l \end{aligned} \tag{10.1}$$

where $x_{ij}$ and $y_{rj}$ are $i$-th input and $r$-th output of DMU$j$ respectively. When $l = 1$, model (10.1) becomes the original input-oriented CCR model (Charnes et al. 1978) and $E^1$ consists of all the radially efficient DMUs. A radially efficient DMU may have non-zero input/output slack values. The DMUs in set $E^1$ define the first-level efficient frontier. When $l = 2$, model (10.1) gives the second-level efficient frontier after the exclusion of the first-level efficient DMUs. In this manner, we identify several levels of efficient frontiers. We call $E^l$ the $l$-th level efficient frontier. The following algorithm accomplishes the identification of these efficient frontiers by model (10.1).

*Step 1* Set $l = 1$. Evaluate the entire set of DMUs, $J^1$, by model (10.1) to obtain the first-level efficient DMUs, set $E^1$ (the first-level efficient frontier).

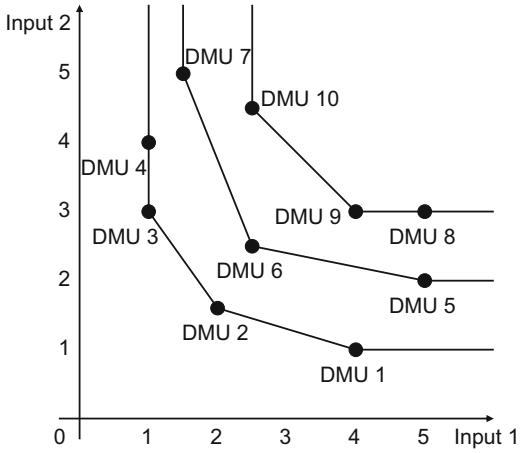



**Fig. 10.1** Efficient Frontiers in Different Levels

**Table 10.1** Numerical example

| DMU | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Input 1 | 4 | 2 | 1 | 1 | 5 | 2.5 | 1.5 | 5 | 4 | 2.5 |
| Input 2 | 1 | 1.5 | 3 | 4 | 2 | 2.5 | 5 | 3 | 3 | 4.5 |

*Step 2* Let $J^{l+1} = J^l - E^l$ to exclude the efficient DMUs from future DEA runs. If $J^{l+1} = \emptyset$ then stop.

*Step 3* Evaluate the new subset of "inefficient" DMUs, $J^{l+1}$, by model (10.1) to obtain a new set of efficient DMUs $E^{l+1}$ (the new efficient frontier).

*Step 4* Let $l = l + 1$. Go to step 2.

*Stopping rule* If $J^{l+1} = \emptyset$, the algorithm stops.

Model (10.1) yields a stratification of the whole set of DMUs, which partitions the DMUs into different subgroups of efficiency levels characterized by $E^l$. It is easy to show that these sets of DMUs have the following properties:

1. $J^1 = \bigcup E^l$ and $E^l \cap E^{l'} = \phi$ for $l \neq l'$;
2. The DMUs in $E^{l'}$ are dominated by the DMUs in $E^l$ if $l' > l$;
3. Each DMU in set $E^l$ is efficient with respect to the DMUs in set $J^{l'}$ for all $l' > l$.

Figure 10.1 plots the three levels of efficient frontiers of 10 DMUs with two inputs and one single output as shown in Table 10.1.

### 10.2.2 Attractiveness and Progress

Based upon the evaluation context $E^l$, the context-dependent DEA measures the relative attractiveness of DMUs. Consider a specific DMU$q$ in $E^l$. The following model is used to characterize the attractiveness with respect to levels exhibiting poorer performance in $E^{l'}$ for $l' > l$.

$$\begin{aligned} \underset{\lambda,\theta}{\text{minimize}} \quad & \theta_q^{l'} = \theta \\ \text{subject to} \quad & \sum_{j \in J^{l'}} \lambda_j x_{ij} \le \theta x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in J^{l'}} \lambda_j y_{rj} \ge y_{rq}, r = 1, \ldots, s \\ & \lambda_j \ge 0, j \in J^{l'} \end{aligned} \tag{10.2}$$

It is easy to show that $\theta_q^{l'} > 1$ for $l' > l$, and $\theta_q^{l_1} > \theta_q^{l_2}$ for $l_1 > l_2$. Then $\theta_q^{l'}$ is called the input-oriented $d$-degree attractiveness of DMU$q$ from a specific level $E^l$, where $d = l' - l$.

In model (10.2), each efficient frontier represents an evaluation context for evaluating the relative attractiveness of DMUs in $E^l$. Note that the bigger the value of $\theta_q^{l'} > 1$, the more attractive DMU$q$ is, because DMU$q$ makes itself more distinctive from the evaluation context $E^{l'}$. We are able to rank the DMUs in $E^l$ based upon their attractiveness scores and identify the best one.

To obtain the progress measure for a specific DMU$q$ in $E^l$, we use the following context-dependent DEA, which is used to characterize the progress with respect to levels exhibiting better performance in $E^{l'}$ for $l' < l$.

$$\begin{aligned} \underset{\lambda,\varphi}{\text{minimize}} \quad & \varphi_q^{l'} = \varphi \\ \text{subject to} \quad & \sum_{j \in J^{l'}} \lambda_j x_{ij} \le \varphi x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in J^{l'}} \lambda_j y_{rj} \ge y_{rq}, r = 1, \ldots, s \\ & \lambda_j \ge 0, j \in J^{l'} \end{aligned} \tag{10.3}$$

We have that $\varphi_q^{l'} < 1$ for $l' < l$, and $\varphi_q^{l_1} < \varphi_q^{l_2}$ for $l_1 > l_2$. Then $\varphi_q^{l'}$ is called the input-oriented $g$-degree progress of DMU$q$ from a specific level $E^l$, where $g = l - l'$.

### 10.2.3 Output Oriented Context-Dependent DEA Model

Here we provide the output-oriented context-dependent DEA model. Consider the following linear programming problem for DMU$q$ in specific level $E^l$ based upon

the evaluation context $E^{l'}$ for $l' > l$.

$$\begin{aligned} \underset{\lambda, H}{\text{maximize}} \quad & H_q^{l'} = H \\ \text{subject to} \quad & \sum_{j \in J^{l'}} \lambda_j x_{ij} \leq x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in J^{l'}} \lambda_j y_{rj} \geq H y_{rq}, r = 1, \ldots, s \\ & \lambda_j \geq 0, j \in J^{l'} \end{aligned} \tag{10.4}$$

This problem is used to characterize the attractiveness with respect to levels exhibiting poorer performance in $E^{l'}$. Note that dividing each side of the constraint of (10.4) by $H$ gives

$$\sum_{j \in J^{l'}} \tilde{\lambda}_j x_{ij} \leq \frac{1}{H} x_{iq}$$

$$\sum_{j \in J^{l'}} \tilde{\lambda}_j y_{rj} \geq y_{rq}$$

$$\tilde{\lambda}_j = \frac{\lambda_j}{H} \geq 0, j \in J^{l'}$$

Therefore, (10.4) is equivalent to (10.2), and we have that $H_q^{l'} < 1$ for $l' > l$ and $H_q^{l'} = 1/\theta_q^{l'}$. Then $H_q^{l'}$ is called the output-oriented $d$-degree attractiveness of DMU$q$ from a specific level $E^l$, where $d = l' - l$. The smaller the value of $H_q^{l'}$ is, the more attractive DMU$q$ is. Model (10.4) determines the relative attractiveness score for DMU$q$ when inputs are fixed at their current levels.

To obtain the progress measure for DMU$q$ in $E^l$, we develop the following linear programming problem, which is used to characterize the progress with respect to levels exhibiting better performance in $E^{l'}$ for $l' < l$.

$$\begin{aligned} \underset{\lambda, G}{\text{maximize}} \quad & G_q^{l'} = G \\ \text{subject to} \quad & \sum_{j \in J^{l'}} \lambda_j x_{ij} \leq x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in J^{l'}} \lambda_j y_{rj} \geq G y_{rq}, r = 1, \ldots, s \\ & \lambda_j \geq 0, j \in J^{l'} \end{aligned} \tag{10.5}$$

We have that $G_q^{l'} > 1$ for $l' < l$ and $G_q^{l'} = 1/\varphi_q^{l'}$. Then $G_q^{l'}$ is called the output-oriented $g$-degree progress of DMU$q$ from a specific level $E^l$, where $g = l - l'$.

To improve the performance of inefficient DMU, the target of improvement should be given among the efficient DMUs. The reference set suggests the target of improvement for the inefficient DMUs. Actually, when $l = 1$, model (10.1) gives the reference set of DMUs from the efficient DMUs for inefficient DMUs. It may be a final goal of improvement; however, for some inefficient DMUs, this goal may be quite different from the current performance and difficult to achieve. Therefore, it is not appropriate to set a benchmark target for improvement from the efficient DMUs directly. Step-by-step improvement is a useful way to improve the performance, and the benchmark target at each step is provided based on the evaluation context at each level of efficient frontier.

### *10.2.4 Context-Dependent DEA With Value Judgment*

Both attractiveness and progress are measured radially with respect to different *levels* of efficient frontiers. The measurement does not require *a priori* information on the importance of the attributes (input/output) that feature in the performance of DMUs. However different attributes play different roles in the evaluation of a DMU's overall performance. Therefore, we introduce value judgment into the context-dependent DEA.

In order to incorporate such *a priori* information into our measures of attractiveness and progress, we first specify a set of weights related to the $m$ inputs, $v_i, i = 1, \ldots, m$ such that $\sum_{i=1}^{m} v_i = 1$. Based upon Zhu (1996), we develop the following linear programming problem for DMU$q$ in $E^l$.

$$\begin{aligned}
\underset{\lambda_j, \Theta_{iq}}{\text{Maximize}} \quad & \Theta_q^{l'*} = \sum_{r=1}^{s} v_i \Theta_{iq} \\
\text{subject to} \quad & \sum_{j \in E^{l'}} \lambda_j x_{ij} \leq \Theta_{iq} x_{iq}, i = 1, \ldots, m \\
& \sum_{j \in E^{l'}} \lambda_j y_{rj} \geq y_{rq}, r = 1, \ldots, s \\
& \Theta_{iq} \geq 1, i = 1, \ldots, m \\
& \lambda_j \geq 0, j \in E^{l'}
\end{aligned} \tag{10.6}$$

$\Theta_q^{l'*}$ is called the input-oriented value judgment *d-degree* attractiveness of DMU$q$ from a specific level $E^l$, where $d = l' - l$. Obviously, $\Theta_q^{l'*} > 1$. The larger the $\Theta_q^{l'*}$ is, the more attractive the DMU$q$ appears under the weights $v_i, i = 1, \ldots, m$. We now can rank DMUs in the same level by their attractiveness scores with value judgment which are incorporated with the preferences over outputs.

If one wishes to prioritize the options (DMUs) with higher values of the $i_o$-th input, then one can increase the value of the corresponding weight $v_{i_o}$. These user-specified weights reflect the relative degree of desirability of the corresponding outputs. For example, if one prefers a printer with faster printing speed to one with higher print quality, then one may specify a larger weight for the speed. The constraints of $\Theta_{iq} \geq 1, i = 1, \ldots, m$ ensure that in an attempt to make itself as distinctive as possible, DMU$q$ is not allowed to decrease some of its outputs to achieve higher levels of other preferred outputs.

Note that $\Theta_q^{l'^*}$ is an overall attractiveness of DMU$q$ in terms of inputs while keeping the outputs at their current levels. On the other hand, each individual optimal value of $\Theta_{iq}, i = 1, \ldots, m$ measures the attractiveness of DMU$q$ in terms of each input dimension. $\Theta_{iq}^*$ is called the input-oriented value judgment input-specific attractiveness measure for DMU$q$.

With the input-specific attractiveness measures, one can further identify which inputs play important roles in distinguishing a DMU's performance. On the other hand, if $\Theta_{i_oq}^* = 1$, then other DMUs in $E^{l'}$ or their combinations can also produce the same amount as the $i_o$-th input of DMU$q$, i.e., DMU$q$ does not exhibit better performance with respect to this specific input dimension. Therefore, DMU$q$ should improve its performance on the $i_o$-th input to distinguish itself in the future.

Similar to the development in the previous section, we can define the input-oriented value judgment progress measure:

$$\begin{aligned} \underset{\lambda_j, \Phi_{iq}}{\text{Maximize}} \quad & \Phi_q^{l'^*} = \sum_{r=1}^{s} v_i \Phi_{iq} \\ \text{subject to} \quad & \sum_{j \in E^{l'}} \lambda_j x_{ij} \leq \Phi_{iq} x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in E^{l'}} \lambda_j y_{rj} \geq y_{rq}, r = 1, \ldots, s \\ & \Phi_{iq} \leq 1, i = 1, \ldots, m \\ & \lambda_j \geq 0, j \in E^{l'} \end{aligned} \tag{10.7}$$

The optimal value $\Theta_q^{l'^*}$ is called the input-oriented value judgment *g-degree* progress DMU$q$ from a specific level $E^l$, where $g = l - l'$. The larger $\Theta_q^{l'^*}$ is, the greater the amount of progress is expected for DMU$q$. Here the user-specified weights reflect the relative degree of desirability of improvement on the individual output levels. Let $\Phi_{iq}^*, i = 1, \ldots, m$, represent the optimal value of (10.7) for a specific level $l$. By Zhu (1996), we know that $\sum_{j \in E^{l'}} \lambda_j^* x_{ij} = \Phi_{iq}^* x_{iq}$ holds at optimality for each

$i = 1, \ldots, m$. Consider the following linear programming problem:

$$\begin{aligned} \text{Maximize} \quad & \sum_{r=1}^{s} s_r^+ \\ \text{subject to} \quad & \sum_{j \in E^{l'}} \lambda_j x_{ij} = \Phi^*{}_{iq} x_{iq}, i = 1, \ldots, m \\ & \sum_{j \in E^{l'}} \lambda_j y_{rj} - s_r^+ = y_{rq}, r = 1, \ldots, s \\ & s_r^+ \geq 0, r = 1, \ldots, s \\ & \lambda_j \geq 0, j \in E^{l'} \end{aligned} \tag{10.8}$$

The following point

$$\begin{cases} \hat{x}_{iq} = \Phi_{iq}^* x_{iq}, i = 1, \ldots, m \\ \hat{y}_{rq} = y_{rq} + s_r^{+*}, r = 1, \ldots, s \end{cases}$$

is called a *preferred global efficient target* for DMU$q$ in level $E^l$ for $l' = l - 1$; otherwise, if $l' < l - 1$, it represents a *preferred local efficient target*, where $\Phi_{iq}^*$ is the optimal value in (10.7), and $s_r^{+*}$ represent the optimal values in (10.8).

## 10.3 Variable and Fixed Benchmarking Models

Cook et al. (2004) develop DEA-based models for use in benchmarking where multiple performance measures are needed to examine the performance and productivity changes. The standard data envelopment analysis method is extended to incorporate benchmarks through (i) a variable-benchmark model where a unit under benchmarking selects a portion of benchmark such that the performance is characterized in the most favorable light, and (ii) a fixed-benchmark model where a unit is benchmarked against a fixed set of benchmarks. Cook et al. (2004) apply these models to a large Canadian bank where some branches' services are automated to reduce costs and increase the service speed, and ultimately to improve productivity. Their empirical investigation indicates that although the performance appears to be improved at the beginning, productivity gain has not been discovered. The models can facilitate the bank in examining its business options and further point to weaknesses and strengths in branch operations. The current chapter presents the benchmarking models developed by Cook et al. (2004).

### 10.3.1 Variable-Benchmark Model

Let $E^*$ represent the set of benchmarks or the best-practice identified by the DEA. Based upon the input-oriented Constant Returns to Scale (CRS) DEA model, we have

$$\begin{aligned} &\min \delta^{CRS} \\ &\text{subject to} \\ &\sum_{j\in E^*} \lambda_j x_{ij} \le \delta^{CRS} x_i^{new} \\ &\sum_{j\in E^*} \lambda_j y_{rj} \ge y_r^{new} \\ &\lambda_j \ge 0, j \in E^* \end{aligned} \tag{10.9}$$

where a new observation is represented by $DMU^{new}$ with inputs $x_i^{new}(i = 1, \ldots, m)$ and outputs $y_r^{new}(r = 1, \ldots, s)$. The superscript of CRS indicates that the benchmark frontier composed by benchmark DMUs in set $E^*$ exhibits CRS.

Model (10.9) measures the performance of $DMU^{new}$ with respect to benchmark DMUs in set $E^*$ when outputs are fixed at their current levels. Similarly, based upon the output-oriented CRS envelopment model, we can have a model that measures the performance of $DMU^{new}$ in terms of outputs when inputs are fixed at their current levels.

$$\begin{aligned} &\max \tau^{CRS} \\ &\text{subject to} \\ &\sum_{j\in E^*} \lambda_j x_{ij} \le x_i^{new} \\ &\sum_{j\in E^*} \lambda_j y_{rj} \ge \tau^{CRS} y_r^{new} \\ &\lambda_j \ge 0, j \in E^* \end{aligned} \tag{10.10}$$

Note that $\delta^{CRS*} = 1/\tau^{CRS*}$, where $\delta^{CRS*}$ is the optimal value to model (10.9) and $\tau_o^{CRS*}$ is the optimal value to model (10.10).

Model (10.9) or (10.10) yields a benchmark for $DMU^{new}$. The $i$th input and the $r$th output for the benchmark can be expressed as

$$\begin{cases} \sum_{j\in E^*} \lambda_j^* x_{ij} \quad (ith \quad input) \\ \sum_{j\in E^*} \lambda_j^* y_{ij} \quad (rth \quad output) \end{cases} \tag{10.11}$$

Note also that although the DMUs associated with set $E^*$ are given, the resulting benchmark may be different for each new DMU under evaluation. For each new DMU

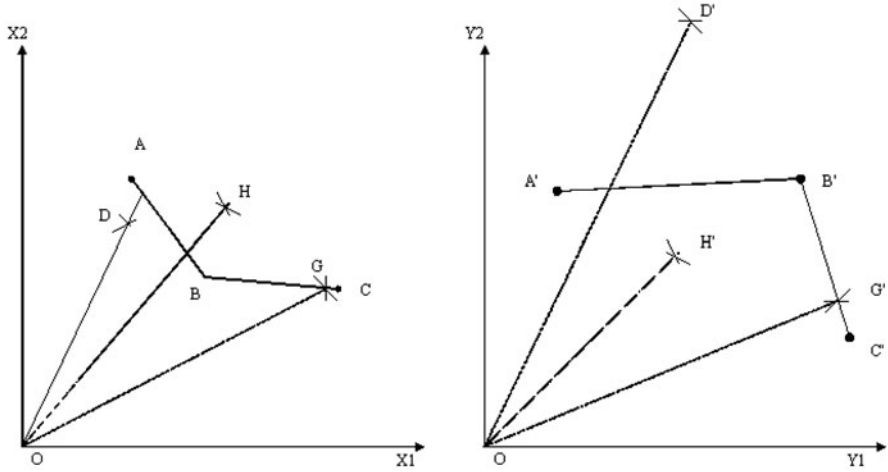



**Fig. 10.2** Variable-benchmark Model

under evaluation, (10.11) may represent a different combination of DMUs associated with set $E^*$. Thus, models (10.9) and (10.10) represent a variable-benchmark scenario.

We have

1. $\delta^{CRS^*} < 1$ or $\tau^{CRS^*} > 1$ indicates that the performance of $DMU_o^{new}$ is dominated by the benchmark in (10.11).
2. $\delta^{CRS^*} = 1$ or $\tau^{CRS^*} = 1$ indicates that $DMU^{new}$ achieves the same performance level as the benchmark in (10.11).
3. $\delta^{CRS^*} > 1$ or $\tau^{CRS^*} < 1$ indicates that input savings or output surpluses exist in $DMU_o^{new}$ when compared to the benchmark in (10.11).

Figure 10.2 illustrates the three cases. ABC (A'B'C') represents the input (output) benchmark frontier. D, H and G (or D', H', and G') represent the new DMUs to be benchmarked against ABC (or A'B'C'). We have $\delta_D^{CRS^*} > 1$ for DMU D ( $\tau_{D'}^{CRS^*} < 1$ for DMU D') indicating that DMU D can increase its input values by $\delta_D^{CRS^*}$ while producing the same amount of outputs generated by the benchmark (DMU D' can decrease its output levels while using the same amount of input levels consumed by the benchmark). Thus, $\delta_D^{CRS^*} > 1$ is a measure of input savings achieved by DMU D and $\tau_{D'}^{CRS^*} < 1$ is a measure of output surpluses achieved by DMU D'.

For DMU G and DMU G', we have $\delta_G^{CRS^*} = 1$ and $\tau_{G'}^{CRS^*} = 1$ indicating that they achieve the same performance level of the benchmark and no input savings or output surpluses exist. For DMU H and DMU H', we have $\delta_H^{CRS^*} < 1$ and $\tau_{H'}^{CRS^*} > 1$ indicating that inefficiency exists in the performance of these two DMUs.

Note that for example, in Fig. 10.2, a convex combination of DMU A and DMU B is used as the benchmark for DMU D while a convex combination of DMU B and

DMU C is used as the benchmark for DMU G. Thus, models (10.9) and (10.10) are called variable-benchmark models.

We can define $\delta^{CRS^*} - 1$ or $1 - \tau^{CRS^*}$ as the performance gap between $DMU^{new}$ and the benchmark. Based upon $\delta^{CRS^*}$ or $\tau^{CRS^*}$, a ranking of the benchmarking performance can be obtained.

It is likely that scale inefficiency may be allowed in the benchmarking. We therefore modify models (10.9) and (10.10) to incorporate scale inefficiency by assuming variable returns to scale (VRS).

$$
\begin{aligned}
&\min \delta^{VRS} \\
&\text{subject to} \\
&\sum_{j \in E^*} \lambda_j x_{ij} \le \delta^{VRS} x_i^{new} \\
&\sum_{j \in E^*} \lambda_j y_{rj} \ge y_r^{new} \\
&\sum_{j \in E^*} \lambda_j = 1 \\
&\lambda_j \ge 0, j \in E^*
\end{aligned}
\tag{10.12}
$$

$$
\begin{aligned}
&\max \tau^{VRS} \\
&\text{subject to} \\
&\sum_{j \in E^*} \lambda_j x_{ij} \le x_i^{new} \\
&\sum_{j \in E^*} \lambda_j y_{rj} \ge \tau^{VRS} y_r^{new} \\
&\sum_{j \in E^*} \lambda_j = 1 \\
&\lambda_j \ge 0, j \in E^*
\end{aligned}
\tag{10.13}
$$

We have

1. $\delta^{VRS^*} < 1$ or $\tau^{VRS^*} > 1$ indicates that the performance of $DMU^{new}$ is dominated by the benchmark in (10.11).
2. $\delta^{VRS^*} = 1$ or $\tau^{VRS^*} = 1$ indicates that $DMU^{new}$ achieves the same performance level as the benchmark in (10.11).
3. $\delta^{VRS^*} > 1$ or $\tau^{VRS^*} < 1$ indicates that input savings or output surpluses exist in $DMU^{new}$ when compared to the benchmark in (10.11).

Note that model (10.10) is always feasible, and model (10.9) is infeasible only if certain patterns of zero data are present (Zhu 1996b). Thus, if we assume that all

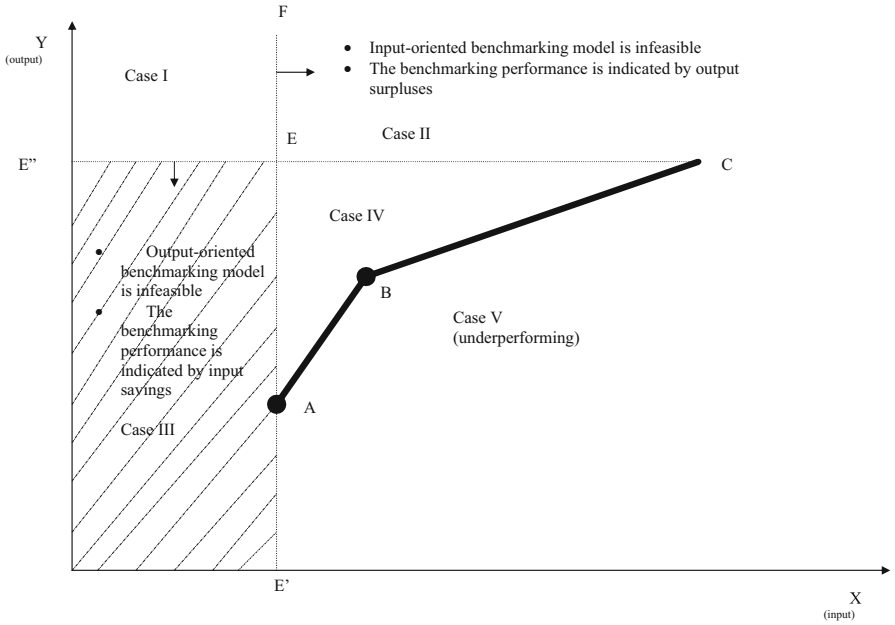



**Fig. 10.3** Infeasibility of VRS Variable-benchmark Model

the data are positive, (10.9) is always feasible. However, unlike models (10.9) and (10.10), models (10.12) and (10.13) may be infeasible.

We have

1. If model (10.12) is infeasible, then the output vector of $DMU^{new}$ dominates the output vector of the benchmark in (10.11).
2. If model (10.13) is infeasible, then the input vector of $DMU^{new}$ dominates the input vector of the benchmark in (10.11).

The implication of the infeasibility associated with models (10.12) and (10.13) needs to be carefully examined. Consider Fig. 10.3 where ABC represents the benchmark frontier. Models (10.12) and (10.13) yield finite optimal values for any $DMU^{new}$ located below EC and to the right of EA. Model (10.12) is infeasible for $DMU^{new}$ located above ray E"C and model (10.13) is infeasible for $DMU^{new}$ located to the left of ray E'E.

Both models (10.12) and (10.13) are infeasible for $DMU^{new}$ located above E"E and to the left of ray EF. Note that if $DMU^{new}$ is located above E"C, its output value is greater than the output value of any convex combinations of A, B and C.

Note also that if $DMU^{new}$ is located to the left of E'F, its input value is less than the input value of any convex combinations of A, B and C.

Based upon Fig. 10.3, we have four cases:

Case I: When both models (10.12) and (10.13) are infeasible, this indicates that $DMU^{new}$ has the smallest input level and the largest output level compared to the benchmark. Thus, both input savings and output surpluses exist in $DMU^{new}$.

Case II: When model (10.12) is infeasible and model (10.13) is feasible, the infeasibility of model (10.12) is caused by the fact that $DMU^{new}$ has the largest output level compared to the benchmark. Thus, we use model (10.13) to characterize the output surpluses.

Case III: When model (10.13) is infeasible and model (10.12) is feasible, the infeasibility of model (10.13) is caused by the fact that $DMU^{new}$ has the smallest input level compared to the benchmark. Thus, we use model (10.12) to characterize the input savings.

Case IV: When both models (10.12) and (10.13) are feasible, we use both of them to determine whether input savings and output surpluses exist.

### *10.3.2 Fixed-Benchmark Model*

Although the benchmark frontier is given in the variable-benchmark models, a $DMU^{new}$ under benchmarking has the freedom to choose a subset of benchmarks so that the performance of $DMU^{new}$ can be characterized in the most favorable light. Situations when the same benchmark should be fixed are likely to occur. For example, the management may indicate that DMUs A and B in Fig. 10.2 should be used as the fixed benchmark. i.e., DMU C in Fig. 10.2 may not be used in constructing the benchmark.

To couple with this situation, Cook et al. (2004) turn to the multiplier DEA models. For example, the input-oriented CRS multiplier DEA model determines a set of referent best-practice DMUs represented by a set of binding constraints in optimality. Let set $\mathbf{B} = \left\{DMU_j : j \in \mathbf{I}_B\right\}$ be the selected subset of benchmark set $E^*$. i.e., $\mathbf{I}_B \subset E^*$ Based upon the input-oriented CRS multiplier model, we have

$$
\begin{aligned}
&\tilde{\sigma}^{\mathrm{CRS}^*} = \max \sum_{r=1}^{s} \mu_r y_r^{new} \\
&\text{subject to} \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} = 0 \quad j \in \mathbf{I}_B \\
&\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0 \quad j \notin \mathbf{I}_B \\
&\sum_{i=1}^{m} v_i x_i^{new} = 1 \\
&\mu_r, v_i \ge 0.
\end{aligned}
\tag{10.14}
$$

**Table 10.2** Fixed-benchmark Models

| Frontier type | Input-oriented | Output-oriented |
|---|---|---|
| | $\max \sum_{r=1}^{s} \mu_r y_r^{new} + \mu$ | $\min \sum_{i=1}^{m} v_i x_i^{new} + v$ |
| | subject to | subject to |
| | $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + \mu = 0 \quad j \in \mathbf{I}_B$ | $\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + v = 0 \; j \in \mathbf{I}_B$ |
| | $\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} + \mu \le 0 \quad j \notin \mathbf{I}_B$ | $\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} + v \ge 0 \; j \notin \mathbf{I}_B$ |
| | $\sum_{i=1}^{m} v_i x_i^{new} = 1$ | $\sum_{r=1}^{s} \mu_r y_r^{new} = 1$ |
| | $\mu_r, v_i \ge 0$ | $\mu_r, v_i \ge 0$ |
| CRS | Where $\mu = 0$ | Where $v = 0$ |
| VRS | Where $\mu$ free | Where $v$ free |

By applying equalities in the constraints associated with benchmark DMUs, model (10.14) measures $DMU^{new}$'s performance against the benchmark constructed by set $\boldsymbol{B}$. At optimality, some DMU$_j$ $j \notin \mathbf{I}_B$, may join the fixed-benchmark set if the associated constraints are binding.

Note that model (10.14) may be infeasible. For example, the DMUs in set $\boldsymbol{B}$ may not be fit into the same facet when they number greater than m + s-1, where $m$ is the number of inputs and $s$ is the number of outputs. In this case, we need to adjust the set $\boldsymbol{B}$.

Three possible cases are associated with model (10.14). $\tilde{\sigma}^{CRS^*} > 1$ indicating that $DMU^{new}$ outperforms the benchmark. $\tilde{\sigma}^{CRS^*} = 1$ indicating that $DMU^{new}$ achieves the same performance level of the benchmark. $\tilde{\sigma}^{CRS^*} < 1$ indicating that the benchmark outperforms $DMU^{new}$.

By applying returns to scale (RTS) frontier type and model orientation, we obtain the fixed-benchmark models in Table 10.2.

A commonly used measure of efficiency is the ratio of output to input. For example, profit per employee measures the labor productivity. When multiple inputs and outputs are present, we may define the following efficiency ratio

$$\frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}$$

where $v_i$ and $u_r$ represent the input and output weights, respectively.

DEA calculates the ratio efficiency without the information on the weights. In fact, the multiplier DEA models can be transformed into linear fractional programming problems. For example, if we define $v_i = \mathrm{t}\, v_i$ and $\mu_r = \mathrm{t}\, u_r$, where $\mathrm{t} = 1/\sum v_i x_{io}$,

the input-oriented CRS multiplier model can be transformed into

$$\begin{aligned} &\max \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \\ &\text{subject to} \\ &\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 j = 1,2,\ldots,n \\ &u_r, v_i \geq 0 \quad \forall r, i \end{aligned} \tag{10.15}$$

The objective function in (10.15) represents the efficiency ratio of a DMU under evaluation. Because of the constraints in (10.15), the (maximum) efficiency cannot exceed one. Consequently, a DMU with an efficiency score of one is on the frontier. It can be seen that no additional information on the weights or tradeoffs are incorporated into the model (10.15).

If we apply the input-oriented CRS fixed-benchmark model to (10.15), we obtain

$$\begin{aligned} &\max \frac{\sum_{r=1}^{s} u_r y_r^{new}}{\sum_{i=1}^{m} v_i x_i^{new}} \\ &\text{subject to} \\ &\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} = 1 j \in \mathbf{I}_{\mathrm{B}} \\ &\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 j \notin \mathbf{I}_{\mathrm{B}} \\ &u_r, v_i \geq 0 \quad \forall r, i \end{aligned} \tag{10.16}$$

It can be seen from (10.16) that the fixed benchmarks incorporate implicit tradeoff information into the efficiency evaluation. i.e., the constraints associated with $\mathbf{I}_{\mathrm{B}}$ can be viewed as the incorporation of tradeoffs or weight restrictions in DEA. Model (10.16) yields the (maximum) efficiency under the implicit tradeoff information represented by the benchmarks.

As more DMUs are selected as fixed benchmarks, more complete information on the weights becomes available.

## 10.4 Concluding Remarks

This chapter presents the context-dependent DEA and benchmarking DEA approaches. Morita et al. (2005) show that non-zero slacks can be incorporated into the context-dependent DEA. Zhu (2014) provides spreadsheet models for calculating the presented DEA models. The benchmarking models developed by Cook et al. (2004) provide tools needed to monitor the performance change and further facilitates the development of the best strategic option for the organization with regard to DMU makeup. The interested reader is referred to Cook et al. (2004).

## References

Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. Manag Sci 39(10):1261–1264

Brissimis SN, Zervopoulos PD (2012) Developing a step-by-step effectiveness assessment model for customer-oriented service organizations. Eur J Oper Res 223:226–233

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Chen Y, Morita H, Zhu J (2005) Context-dependent DEA with an application to Tokyo public libraries. Int J Info Technol Decis Mak 4(3):385–394

Chiu Y-H, Wu M-F (2010) Performance evaluation of international tourism hotels in Taiwan—application of context-dependent DEA. INFOR 48(3):155–170

Cook WD, Seiford LM, Zhu J (2004) Models for performance benchmarking: measuring the effect of e-business activities on banking performance. Omega 32(4):313–322

Lim S (2012) Context-dependent data envelopment analysis with cross-efficiency evaluation. J Oper Res Soc 63(1):38–46

Lu W-M, Hung S-W (2008) Benchmarking the operating efficiency of global telecommunication firms. Int J Info Technol Decis Mak 7(4):737–750

Lu W-M, Lo S-F (2012) Constructing stratifications for regions in China with sustainable development concerns. Qual Quant 46(6):1807–1823

Morita H, Hirokawa K, Zhu J (2005) A slack-based measure of efficiency in context-dependent data envelopment analysis. Omega 33:357–362

Seiford LM, Zhu J (1999b) Infeasibility of super-efficiency DEA models. INFOR 37(2):174–187

Seiford LM, Zhu J (2003) Context-dependent data envelopment analysis: measuring attractiveness and progress. Omega 31(5) 397–408

Sherman HD, Zhu J (2013) Analyzing performance in service organizations. Sloan Manag Rev 54(4):37–42

Tsang S-S, Chen Y-F (2013) Facilitating benchmarking with strategic grouping and data envelopment analysis: the case of international tourist hotels in Taiwan. Asia Pac J Tour Res 18(5):518–533

Tversky A, Simonson I (1993) Context-dependent preferences. Manag Sci 39:1179–1189

Ulucan A, Atici KB (2010) Efficiency evaluations with context-dependent and measure-specific data envelopment approaches: An application in a World Bank supported project. OMEGA 38(1–2):68–83

Wu H, Chen B, Xia Q, Zhou H (2013) Ranking and benchmarking of the Asian games achievements based on DEA: the case of Guangzhou 2010. Asia-Pac J Oper Res 30(6)

Yang C, Wang T-C, Lu W-M (2007) Performance measurement in military provisions: the case of retail stores of Taiwan's General Welfare Service Ministry. Asia-Pac J Oper Res 24(3):313–332

Zhu J (1996) Data envelopment analysis with preference structure. J Oper Res Soc 47(1):136–150

Zhu J (1996b) Robustness of the efficient DMUs in data envelopment analysis. Eur J Oper Res 90(3):451–460

Zhu J (2014) Quantitative models for performance evaluation and benchmarking—data envelopment analysis with spreadsheets (3rd edn). Springer, New York

# Chapter 11
# Data Envelopment Analysis with Non-Homogeneous DMUs

**Wade D. Cook, Julie Harrison, Raha Imanirad, Paul Rouse and Joe Zhu**

**Abstract** Data envelopment analysis (DEA), as originally proposed is a methodology for evaluating the relative efficiencies of a set of *homogeneous* decision making units (DMUs) in the sense that each uses the same input and output measures (in varying amounts from one DMU to another). In some situations, however, the assumption of homogeneity among DMUs may not apply. As an example, consider the case where the DMUs are plants in the same industry which may not all produce the same products. Evaluating efficiencies in the absence of homogeneity gives rise to the issue of how to fairly compare a DMU to other units, some of which may not be exactly in the same 'business'. A related problem, and one that has been examined extensively in the literature, is the *missing data* problem; a DMU produces a certain output, but its value is not known. One approach taken to address this problem is to 'create' a value for the missing output (e.g. substituting zero, or by taking the average of known values), and use it to fill in the gaps. In the present setting, however, the issue isn't that the data for the output is missing for certain DMUs, but rather that the output isn't produced. We argue herein that if a DMU has chosen not to produce a certain output, or for any reason cannot produce that output, and therefore does not put the resources in place to do so, then it would be inappropriate to artificially assign that DMU a zero value or some 'average' value for the nonexistent factor. Specifically, the desire is to fairly evaluate a DMU for what it does, rather than

W. D. Cook (✉) · R. Imanirad
Schulich School of Business, York University, Toronto, ON M3J 1P3, Canada
e-mail: wcook@schulich.yorku.ca

R. Imanirad
e-mail: rimanirad09@schulich.yorku.ca

J. Harrison · P. Rouse
Department of Accounting & Finance, University of Auckland, Auckland, New Zealand
e-mail: j.harrison@auckland.ac.nz

P. Rouse
e-mail: p.rouse@auckland.ac.nz

J. Zhu
School of Business, Worcester Polytechnic Institute, Worcester, MA 01609, USA
e-mail: jzhu@wpi.edu

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_11

penalize or credit it for what it doesn't do. In the current chapter we present DEA-based models for evaluating the relative efficiencies of a set of DMUs where the requirement of homogeneity is relaxed. We then use these models to examine the efficiencies of a set of manufacturing plants.

**Keywords** Nonhomogeneous DMUs · Missing outputs · Subgroups · Assurance regions

## 11.1 Introduction

Data envelopment analysis (DEA), as originally proposed by Charnes et al. (1978), is a methodology for evaluating the relative efficiencies of a set of *homogeneous* decision making units (DMUs) belonging to the same technology in the sense that each uses the same inputs and outputs, measured the same way (in varying amounts from one DMU to another). In some situations, however, the assumption of homogeneity among DMUs may not apply, even though they use the same technology. As an example, consider the case where the DMUs are plants in the same industry which may not all produce the same products, and therefore are not homogeneous. Another example is that where the DMUs are a set of universities, where not all have the same departments. In the current chapter we present DEA-based models for evaluating the relative efficiencies of a set of DMUs that belong to the same technology, but where the requirement of homogeneity is relaxed. We then use these models to examine the efficiencies of a set of manufacturing plants.

Evaluating efficiencies in the absence of homogeneity gives rise to the issue of how to fairly compare a DMU to other units, some of which may not be exactly in the same 'business'. A related problem, and one that has been examined extensively in the literature, is the *missing data* problem; a DMU produces a certain output, but its value is not known. One approach taken to address that problem is to 'create' a value for the missing output (e.g. by taking the average of known values), and use it to fill in the gaps. For outputs, using zero as a dummy for blank entries is another prescribed solution. The question of blank output entries is thus closely related to the treatment of zeros in the data matrices (see e.g. Thompson et al. (1993) for discussion).

In the present setting, however, the issue isn't that the data for the output is missing for certain DMUs, but rather that the output isn't produced. In the case of the universities acting as the DMUs, those without engineering departments cannot be directly compared to those that do have such departments, and substituting a value such as zero for this 'missing' data is not appropriate. We argue herein that if a DMU has chosen not to produce a certain output (e.g. the missing engineering department), or for any reason cannot produce that output, and therefore does not put the resources in place to do so, then it would be inappropriate to artificially assign that DMU a zero value or some 'average' value for the nonexistent factor. Specifically, the desire is to fairly evaluate a DMU for what it does, rather than penalize or credit it for what it doesn't do.

Potentially, the non-homogeneous DMU issue could be handled by breaking the set of DMUs into multiple groups, with all members of any group producing the same outputs, and then doing a separate DEA analysis for each group. In this way, a DMU is evaluated against only *true peers*, specifically those whose output profiles are identical to its own. No attempt would be made to compare a DMU to other 'partial peers', namely those whose output profiles overlap with, but are not identical to those of the said DMU. There are at least two problems with this approach. One is a small sample issue in that there may be, in some cases, very few (if any) actual peers. Specifically, in some situations this would require the set of DMUs to be split into multiple small sets to reflect the permutations. The greater the number of splits required, the more difficult it is to estimate meaningful efficiency. It would commonly mean that efficiency scores would be artificially inflated. Another problem is that true best practices for a DMU may in fact be those practices adopted by the partial peers, and excluding consideration of the latter may result in a failure to identify such best practices. This being the case, we wish, wherever possible, to include all DMUs in the comparison set.

Section 2 describes a problem setting involving the evaluation of a set of manufacturing plants, where identifiable groups of DMUs produce only proper subsets of the full set of outputs. Section 3 is devoted to the development of a DEA-type model for handling the general missing output situation. Generally, this is brought about by viewing the DMU as consisting of mutually exclusive subgroups of outputs. One important extension of the DEA concept that has been discussed extensively in the literature is that involving the imposition of multiplier restrictions, in particular those based upon assurance regions (AR). In a setting where there is lack of homogeneity among DMUs, such AR constraints can be problematic in that multiple and often inconsistent sets of restrictions may materialize out of the above-mentioned output subgroups. Section 4 extends the new DEA methodology to allow for consideration of such conflicting AR constraints. Section 5 looks into other issues that may arise relating to non-homogeneous DMUs, and suggests ways of handling such issues. Section 6 applies the new methodology to data for a set of 47 plants relating to the steel fabrication industry, as discussed above. Conclusions appear in Sect. 7.

## 11.2 Manufacturing Plants with Variable Output Sets

To demonstrate the problem of non-homogeneity of DMUs in DEA, a set of 47 steel fabrication plants is considered. The main product lines manufactured by the plants consist of:

1. Sheet steel products (ladders, guards, bumpers and conveyors);
2. Flat bar products used mainly in building construction (brackets, base plates, headers and posts);
3. Pipes and cylinders (storm drains, plumbing products, etc);
4. Furnace and air conditioning ducts;
5. Structural steel (e.g., joists and support beams);
6. Tanks (residential and industrial).

| | Outputs | | | | | |
|---|---|---|---|---|---|---|
| **Group** | Sheet Steel(1) | Flat Bar(2) | Pipes/ Cylinders(3) | Cylindrical. Bearings(4) | Structural Steel (5) | Storage Tanks (6) |
| $N_1$ | X | X | X | | X | |
| $N_2$ | | X | X | X | X | X |
| $N_3$ | | | X | | X | X |
| $N_4$ | X | | X | | X | |

**Fig. 11.1** Product lines by DMU Group

In addition, resources employed by all plants are comprised of: (1) Plant labor; (2) Shears and saws; (3) Presses and rolling equipment; and (4) Cutting torches and welding equipment.

In this particular industry some plants choose not to manufacture certain products. As shown in Fig. 11.1, plants with similar product lines have been grouped together into $P$ DMU groups $N_p$ $p = 1, .., P$, where in our particular case $P = 4$. Observe, for example, that plants in $N_1$ manufacture products 1,2,3,5; those in $N_2$ make products 2,3,4,5,6; etc. Part of the reason for the variability of products across a business (DMU) has to do with the focus on industrial versus residential clientele. Some companies also may cater more to sectors such as automotive than is true of others.

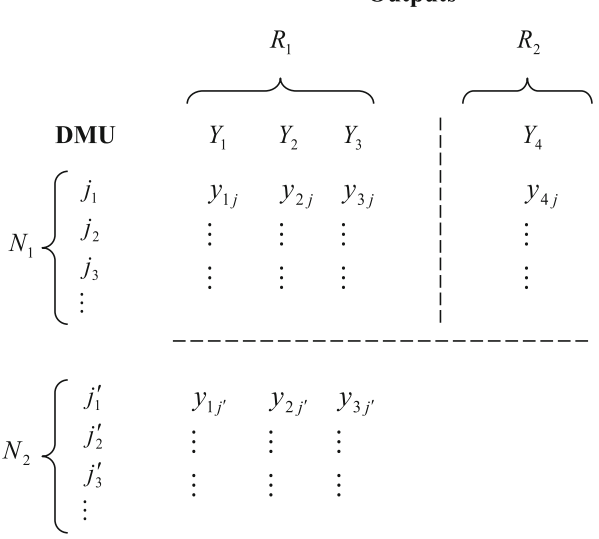
In the following section we develop a DEA based methodology for dealing with non-homogeneous settings such as that represented by Fig. 11.1.

## 11.3 A DEA Model for DMUs with Variable Output Sets

In an earlier paper (Cook et al. (2012)) a simple case where DMUs appeared in a 2-group setting was explored, and affords a convenient and transparent backdrop and introduction for demonstrating the methodology to be developed herein. For completeness, we summarize some of the elements of that earlier development. Specifically, consider the situation where $n$ DMUs are organized into two subgroups $N_1$ and $N_2$, with those in $N_1$ producing 4 outputs $y_1, y_2, y_3, y_4$, while those in $N_2$ produce only 3 outputs $y_1, y_2, y_3$, with both subgroups using the same inputs. Figure 11.2 demonstrates the split of DMUs across subgroups $N_1$ and $N_2$.

Hence, when we want to evaluate a DMU, say in the first group $N_1$, we argue that the evaluation may reasonably be undertaken by carrying out a separate DEA analysis on each of that DMUs 2 output subgroups $R_1 = \{y_1, y_2, y_3\}$, and $R_2 = \{y_4\}$. We further argue that for DMUs in $N_1$, one may think of each input as being split between the production of the subset of outputs in $R_1$ and those in $R_2$. (The situation where some inputs are not separable is discussed later in Sect. 6). If we knew what the

**Fig. 11.2** A two-group setting




proportional split of inputs was between these two output groups, we could proceed in three stages as follows:

**Stage 1** In this stage we decide on a split of the inputs across the output subgroups. For the moment and to facilitate transparency, let us assume it is known that $\alpha_{1N_1} = 90\,\%$ of each input for any DMU in $N_1$ goes toward the production of $R_1$, and that the remaining $\alpha_{2N_1} = 10\,\%$ goes toward the production of outputs in $R_2$. We generalize this idea below.

**Stage 2** In this stage we derive, for each DMU, efficiency scores for the individual subgroups making up that DMU. Specifically, take 90 % of each of the inputs held by each DMU in $N_1$ and carry out a standard DEA analysis of *all n* DMUs (using outputs in $R_1$). Here, whenever we are looking at a DMU $j$ in $N_1$, we need to remember we have replaced the original amounts of its inputs $x_{ij}$ by the proportional amounts of these $\tilde{x}^1_{ij} = \alpha_{1N_1} x_{ij}$ (that have been assigned to the outputs in $R_1$). Note as well, that in this simple case, the inputs held by DMUs in $N_2$ do not need to be split up, as there is only one relevant subgroup of outputs ($R_1$). Hence, in this case, $\alpha_{1N_2} = 1$ and $\alpha_{2N_2} = 0$. Carry out a standard DEA analysis of each of the *members* in $N_1$ using the output in $R_2$ and with inputs $\tilde{x}^2_{ij} = \alpha_{2N_1} x_{ij}$. Recall that we do not include the members from $N_2$ in this analysis, as these DMUs do not produce the output in $R_2$.

**Stage 3** For DMUs in $N_2$, the DEA scores arrived at in stage 2 are the final scores. For DMUs in $N_1$ we combine the scores from steps 1 and 2 by taking a weighted average (discussed below).

We point out that this idea of splitting inputs across various subsets of outputs is similar in nature to the methodology developed for uncovering multiple variable proportionality (MVP), as described in Cook and Zhu (2011)

It is reasonable to argue at this point that rather than following the above procedure, one might instead simply assign to DMUs in $N_2$ a value of zero for the missing output $y_4$ and proceed with a conventional DEA analysis. Perhaps the best counter argument to this is that the DMUs in $N_2$ are, under a conventional DEA analysis, at liberty to assign a zero weight to those outputs ($y_4$ in this case), that are at a zero level, thereby inferring in a mathematical sense that the DMUs in $N_1$ are in the same 'business' as those in $N_2$. The problem with this is that DMUs in $N_2$, with their limited product line, would commonly use less resources than is the case for their *full service* peers in $N_1$ ... less labor, less machine time, less inventory carrying cost, etc. Hence, DMUs in $N_2$ are accorded an unfair advantage over their $N_1$ peers. To illustrate, consider the simple example where two DMUs have the following profiles:

| DMU # | $y_1$ | $y_2$ | $x$ |
|---|---|---|---|
| 1 | 100 | 100 | 20 |
| 2 | 100 | | 20 |

Here DMU #1 is producing 100 units of each of two products, utilizing 20 units of a single input, while DMU #2 uses the same amount of input but produces only 100 units of output 1. Clearly under a conventional DEA analysis both DMUs will be deemed efficient given that DMU #2 can assign a zero multiplier to the second output. Suppose, however that we knew that approximately 80 % of DMU #1's input went toward the production of product 1 and the remaining 20 % was used to produce product 2. Let us present the above data in a more exact manner, replacing DMU #1 by two sub-DMUs which we call DMU #1(a) and DMU #1(b).

| DMU# | $y_1$ | $y_2$ | $x$ |
|---|---|---|---|
| 1(a) | 100 | | 16 |
| 1(b) | | 100 | 4 |
| 2 | 100 | | 20 |

Now, following the above notation, $R_1$ is the output set consisting of $y_1$ and $R_2$ contains the output $y_2$. DMU #2 is now evaluated properly against DMU #1(a), and since the latter uses less input than the former, the input-oriented efficiency score for DMU #2 is only 0.8.

Hence, our argument is that in evaluating the efficiency of DMU #2 (or in general those in $N_2$), the comparison to DMU #1 (or those in $N_1$), should be against only that part of DMU #1's business that it has in common with DMU #2.

**The General Case** Let us now examine the general setting, and as a working example, consider the situation portrayed in Fig. 11.1 where a set of manufacturing

plants produces a certain set of products, but not all products are produced in all plants. Suppose the plants fall into $P$ mutually exclusive (M.E.) groups, as described in Sect. 2, which we denote by $\{N_p\}_{p=1}^{P}$ Here $P = 4$.

Now form M.E. output subgroups $R_k, k = 1, \ldots K$, where $R_k$ denotes the subset of outputs with the property that all of its members appear as the outputs of exactly the same set of DMUs (same DMU 'profile'). Specifically, if outputs $r_1, r_2 \in R_k$, then the DMU profiles of these two outputs are identical. Hence, if $r_1$ is an output for DMU groups 1,2,4, then $r_2$ is an output for exactly the same DMU groups. Also, each $R_k$ is *maximal* in the sense that there is no output $r \notin R_k$ that has the same DMU profile as members of $R_k$. It can be shown that for the above DMU profiles, the $K$ output sets are:

$$R_1 = \{1\}, R_2 = \{2\}, R_3 = \{3,5\}, R_4 = \{4\}, R_5 = \{6\}$$

A general algorithm for deriving the maximal output groupings is found in Appendix 1.

**Theorem 3.1** The generated set of maximal output subgroups is unique.

*Proof* Let us assume that the set of maximal output subgroups is not unique. In that case there must exist at least two different sets of output subgroups $S_1$ and $S_2$. It can then be implied that there must be at least one $R_k$ in $S_1$ that is different from $R_k$ in $S_2$. Consequently, there must exist at least one output $r \in R_k$ in $S_1$ such that $r \notin R_k$ in $S_2$. This proves that $R_k$ in $S_2$ is not maximal because there exists output $r \notin R_k$ that has the same DMU profile as members of $R_k$. Hence, it can be concluded that for each $k$ there exists only one maximal $R_k$ and as a result there can only be one set of maximal output subgroups. This completes the proof.

**Definition 3.1** Let $L_{N_p}$ denote those $R_k$ forming the full output set for any DMU in $N_p$.

In the steel plant setting, $L_{N_1} = \{R_1, R_2, R_3\}$, $L_{N_2} = \{R_2, R_3, R_4, R_5\}$, $L_{N_3} = \{R_3, R_5\}$, $L_{N_4} = \{R_1, R_3\}$.

To evaluate the efficiency of a given DMU, we need to proceed in three stages. In stage 1 we decide (for the DMU under evaluation, say $j_o \in N_{p^o}$), what portion of each input $i$ will be allocated to each of the output subgroups $R_k \in L_{N_{p^o}}$; we denote this proportion by $\alpha_{i R_k p^o}$. In stage 2 we evaluate the efficiency of the DMU in terms of each of its subgroups $R_k$, and in stage 3 we take a weighted average of these subgroup scores to get the overall efficiency of the DMU.

**Stage 1: Deriving the Split of Inputs** Let us formalize the ideas for the situation where we *do not* know the precise split of resources as was assumed above. Let the decision variable $\alpha_{i R_k p}$ denote the proportion of input $i$ to be allocated to outputs in subgroup $R_k$ of $L_{N_p}$. We argue that the best way to divide up the resources, hence determining the most appropriate alpha variables, is to do so in a manner that results in the best overall or aggregate score for the DMU, across all of its business subunits. Further, we argue that the overall efficiency of a DMU $j_o \in N_{p^o}$ can reasonably be represented as a weighted average (convex combination) of the $R_k$-subgroup

efficiencies (across all output subgroups in $N_{p^o}$). We point out that this argument is essentially that the DMU is the sum of its parts, and therefore assumes there are no economies or dis-economies of scope. In cases where it is believed such economies (dis-economies) of scope exist, our approach may not accurately capture efficiency at the aggregate level.

Given that it is aggregate efficiency of the DMU that we wish to derive, and that this aggregate will be represented as a convex combination of the $R_k$-subgroup efficiencies, we set out to determine the $\alpha$-split of inputs with the objective of maximizing this aggregate efficiency. With this in mind, consider the following input-oriented radial projection model (11.1) for a DMU $j_o \in N_{p^o}$. It is noted that the development in this section is, in the spirit of Charnes et al. (1978), presented from the perspective of the constant returns to scale (CRS) technology. As demonstrated in a later section, however, the concepts are equally valid for a variable returns to scale (VRS) technology.

$$e_o = \max \sum_{R_k \in L_{N_{p^o}}} W_{R_k j_o} \Big[ \sum_{r \in R_k} u_r y_{r j_o} \Big/ \sum_i v_i \alpha_{i R_k p^o} x_{i j_o} \Big] \tag{11.1a}$$

*subject to*

$$\sum_{R_k \in L_{Np}} W_{R_k j} \Big[ \sum_{r \in R_k} u_r y_{rj} \Big/ \sum_i v_i \alpha_{i R_k p} x_{ij} \Big] \le 1 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.1b}$$

$$\sum_{r \in R_k} u_r y_{rj} - \sum_i v_i \alpha_{i R_k p} x_{ij} \le 0 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.1c}$$

$$\sum_{R_k \in L_{Np}} \alpha_{i R_k p} = 1 \quad \forall i,\ p = 1 \ldots P \tag{11.1d}$$

$$a_{i R_k p} \le \alpha_{i R_k p} \le b_{i R_k p} \quad \forall i,\ R_k,\ p = 1, \ldots P \tag{11.1e}$$

$$u_r, v_i, \alpha_{i R_k p} \ge 0, \quad \forall i, R_k, p \tag{11.1f}$$

We point out that while in the above example it was assumed that the same values of alpha applied to all DMUs, in the general case here, the model makes provision for a different set of alpha variables for each DMU $j$. The basic idea of this model is to represent the overall efficiency of a DMU as a convex combination $\left( \sum_{R_k \in L_{N_{p^o}}} W_{R_k j_o} = 1 \right)$ of the efficiencies $\sum_{r \in R_k} u_r y_{r j_o} / \sum_i v_i \alpha_{i R_k p^o} x_{i j_o}$ of the individual subgroups $R_k$. While the weights $W_{R_k j_o}$ may be any set of values that represent the importance to be attached to the relevant subgroups, there would appear to be at least two reasonable and obvious choices. From an accounting perspective, it is appropriate and reasonable to let the proportion of inputs assigned to (or consumed by) a subgroup, dictate the importance of that subgroup to the overall DMU; the subgroup assigned the largest share of resources would be given the highest weight.

An equally valid definition of importance of a subgroup would be to base it upon the proportion of the aggregate output for the DMU generated by that subgroup; the subgroup that creates the greatest value for the DMU would be weighted the highest. One might also adopt a net contribution or profit criterion to select weights. As a convenience in the case of the input oriented model adopted herein, we select the first of these two approaches, namely we base the weights for the subgroup ratios on the proportions of the aggregate inputs consumed by those subgroups. Thus, we define the weight $W_{R_k j_o}$ to be assigned to subgroup $R_k$ as:

$$W_{R_k j_o} = \sum_i v_i \alpha_{i R_k p^o} x_{i j_o} \Big/ \sum_{R_k \in L_{N_{p^o}}} \Big[ \sum_i v_i \alpha_{i R_k p^o} x_{i j_o} \Big] \tag{11.2}$$

Constraints (11.1b) require that the multipliers chosen for a DMU $j_o$ satisfy the condition that when they are applied to any other DMU, the corresponding ratio (of outputs to inputs) does not exceed unity. At the same time, and in anticipation of the second stage, we impose the requirement that the ratio of outputs to inputs at the *subgroup level* also not exceed unity. Specifically, constraints (11.1c) specify that the resource splitting variables $\alpha_{i R_k p}$ be selected in a manner that allows the efficiency ratio corresponding to the subset of outputs in $R_k$ to assume a value that does not exceed unity for some values of the multipliers $u_r, v_i$. We note that in the presence of (11.1c), constraints (11.1b) are redundant and may be dropped from the model.

Constraints (11.1d) specify that the $\alpha$ values assigned to the subgroups of outputs corresponding to any set $p$ sum to unity for each $i$. Finally, constraints (3.1e) place lower and upper limits on the sizes of the $\alpha$ variables. It is worth noting that in a situation wherein a particular input may not in fact impact certain outputs or output subgroups, the corresponding $\alpha_{i R_k p}$ can of course be set to zero.

*The Equivalent Linear Formulation* Problem (11.1) in its current form is nonlinear. To facilitate linearization, first note that by virtue of the definition we choose to use for the $W_{R_k j_o}$ as given by (11.2), the objective function (11.1a) becomes

$$e_o = \max \Big[ \sum_{R_k \in L_{N_{p^o}}} \sum_{r \in R_k} u_r y_{r j_o} \Big/ \sum_i v_i x_{i j_o} \Big] \tag{11.1a'}$$

Specifically, maximizing the weighted average of subgroup ratios is equivalent to maximizing the overall efficiency ratio of the DMU.

Now make the change of variables $z_{i R_k p} = v_i \alpha_{i R_k p}$, and note that

$$\sum_{R_k \in L_{N_p}} \alpha_{i R_k p} = 1 \Rightarrow v_i \sum_{R_k \in L_{N_p}} \alpha_{i R_k p} = v_i \Rightarrow \sum_{R_k \in L_{N_p}} z_{i R_k p} = v_i$$

Using the usual transformation $t = 1/\sum_i v_i x_{i j_o}$ (see Charnes et al. (1978)), and defining $\mu_r = t u_r, v_i = t v_i, \gamma_{i R_k p} = t z_{i R_k p}$, problem (11.1) becomes:

$$e_o = \max \sum_{R_k \in L_{N_{p^o}}} \sum_{r \in R_k} \mu_r y_{r j_o} \tag{11.3a}$$

*subject to*

$$\sum_{R_k \in L_{N_{p^o}}} \left( \sum_i \gamma_{i R_k p^o} x_{ij^\circ} \right) = 1 \tag{11.3b}$$

$$\sum_{r \in R_k} \mu_r y_{rj} - \sum_i \gamma_{i R_k p} x_{ij} \leq 0 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.3c}$$

$$\sum_{R_k \in L_{N_p}} \gamma_{i R_k p} = \upsilon_i \quad \forall i, \ p = 1 \ldots P \tag{11.3d}$$

$$\upsilon_i a_{i R_k p} \leq \gamma_{i R_k p} \leq \upsilon_i b_{i R_k p} \quad \forall i, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.3e}$$

$$\mu_r, \upsilon_i, \gamma_{i R_k p} \geq \varepsilon, \quad \forall r, i, R_k, p = 1, \ldots P \tag{11.3f}$$

**Stage 2: Deriving the Subgroup Efficiency Scores** Note that the purpose of stage 1 is to derive, for each DMU $j_o$ in $N_{p^o}$, the 'optimal' proportions of inputs $\hat{\alpha}_{i R_k p^o}$ to be assigned to output subgroups $R_k$. These are given by $\hat{\alpha}_{i R_k p^o} = \hat{\gamma}_{i R_k p^o} / \hat{\upsilon}_i$. When these proportions are available (from the solution to model (11.3)), one can then allocate to subgroup $R_k$ the appropriate amount of input $x_{i j_o}$, namely $\tilde{x}^k_{i j_o} = \hat{\alpha}_{i R_k p^o} x_{i j_o}$. The conventional CCR DEA model (see Charnes et al. (1978)) can then be applied to each of the subgroups $R_k$ of $j_o$. Specifically, determine $M_{R_k}$, the set of all DMU groups that have $R_k$ as a member, that is

$$M_{R_k} = \{N_p \ such \ that \ R_k \in L_{N_p}\}. \tag{11.4}$$

Note, for example, in the six-output steel fabrication application described above, $M_{R_1} = \{N_1, N_4\}$, $M_{R_2} = \{N_1, N_2\}$, ... etc.

Now, for each DMU $j_o$, and each subgroup $R_{k^\circ}$ corresponding to the set $N_{p^o}$ that contains $j_o$ as a member, solve the DEA model:

$$e_{R_{k^\circ} j_o} = \max \sum_{r \in R_{k^\circ}} \mu_r y_{r j_o}$$

*subject to*

$$\begin{aligned} &\sum_i \upsilon_i \tilde{x}^{k^\circ}_{i j_o} = 1 \\ &\sum_{r \in R_{k^\circ}} \mu_r y_{rj} - \sum_i \upsilon_i \tilde{x}^{k^\circ}_{ij} \leq 0, \ j \in N_p, \ for \ N_p \in M_{R_{k^\circ}} \\ &\mu_r, \upsilon_i \geq \varepsilon \end{aligned} \tag{11.5}$$

**Stage 3: Deriving the Aggregate Efficiencies** The overall efficiency score of the DMU $j_o$ is now derived by taking a weighted average of the subgroup scores obtained in stage 2, using the $W_{R_k j_o}$ defined in (11.2). It should be pointed out that in computing

$W_{R_k j_o}$ an appropriate set of input multipliers $v_i$ needs to be chosen. Furthermore, the multipliers need to be computed in an environment where all subunits are being compared simultaneously. The aggregate model (11.3) provides such an environment. That is, in (11.3) when DMU $j_o$ is being evaluated, the input portion of expression (11.3c), namely $\sum_i \gamma_{i R_k p^o} x_{i j_o}$ (for $j = j_o$), represents the *value* of that DMU's resources that are assigned to subgroup $R_k$. The *total value* of all resources consumed by DMU $j_o$ is given by $\sum_i \upsilon_{i j_o} x_{i j_o}$, which is scaled to unity as per constraint (11.3b). Hence, the weights $W_{R_k j_o}$ reduce to $W_{R_k j_o} = \sum_i \gamma_{i R_k p^o} x_{i j_o}$. Note again that this set of weights is dependent on the particular DMU $j_o$ under investigation, to reflect the fact that the proportion of inputs allocated to the $k$th subunit is DMU-specific.

The model developed in this section permits one to evaluate efficiencies of a set of DMUs where output profiles are not homogeneous across those units. The proposed approach portrays a DMU's performance as a convex combination of its component parts (subgroups). It is important to point out that in the above structure we do not consider restrictions that might be imposed on the multipliers $\mu, \upsilon$ (referring to model (11.3)), other than those that restrict efficiency ratios to not exceed unity. This is raised here because such restrictions may lead to infeasibilities that would normally not occur in the conventional DEA setting. The following section investigates the role that multiplier restrictions play in this more general environment. First, however, we point to related literature.

**Relation to Previous Work** The methodology developed above is related to two strands of previous research. First, network DEA as originated by Fare and Grosskopf (1996), sets out to evaluate DMU performance by examining the internal sub-processes that make up the DMU. While one can define performance in many ways, if one concentrates on technical efficiency, network DEA provides for both sub-process efficiency scores as well as an overall score for the DMU itself. Thus, the approach herein is a form of network DEA analysis in that the sub-processes are the subunits as we have described above. Arguably, one difference between our methodology and that characterizing network DEA is that our definition of the overall performance of the DMU is that it is a weighted average of the subunit efficiencies. What is normally done in network DEA is to use a conventional DEA model to describe overall efficiency in terms of all inputs entering the DMU versus all outputs leaving the DMU. As well, sub-process shares of inputs would normally be known in advance (except in allocative efficiency settings), as opposed to those shares being derived as part of the optimization procedure, as is the situation herein. Furthermore, there is no clear direct connection in network DEA between the efficiency score for the overall DMU and the scores of the sub-processes. We provide that connection in the methodology presented here.

Other related research carried out by Cook and Hababou (2001) and by Cook et al. (2000) is closer still to work done herein. In that former work, the DMUs are bank branches which are viewed as consisting of two components or subunits, namely sales and service. Those authors develop an overall efficiency score for the branch using a model analogous to (11.1) above. Their model sets out to optimize the ratio of total weighted outputs to total weighted inputs for the overall branch.

Component efficiencies (sales and service) are then simply taken as the ratio of weighted outputs to inputs for those components, $\sum_{r\in R_c} \mu_r y_{rj} / \sum_{i\in I} \gamma_{ik} x_{ij}$ similar to expression (11.3c). Here $R_c$ denotes the output bundles for either the sales or service component. The problem with using this ratio to capture component efficiency is that it doesn't properly capture the component's performance. Specifically, since it is overall branch performance that is being maximized, there is no internal mechanism for insuring that at the same time component scores are appropriately set, consistent with DEA constructs. Our model herein takes the next important step (step 2 above) of using the resource split across the subunits to find the maximal efficiencies for each of those subunits, and then taking the weighted average of those maximal scores (step 3) to arrive at the score for the DMU. In addition, our methodology herein identifies the subunits into which to decompose the DMU, whereas the earlier research pertained only to those applications where components or subunits are well defined in advance. Finally, the earlier work did not consider the issue of conflicting AR constraints as we do herein. This topic is covered in the next section.

## 11.4 AR Restrictions on Pairs of Input Variables

Many different forms of multiplier restrictions in DEA analyses have been discussed in the literature, but none more than those that take the form of assurance regions (AR). AR constraints, as first discussed by Thompson et al. (1990), involve the placing of bounds on the ratios of pairs of multipliers. The resulting DEA-AR model has been employed extensively in numerous performance measurement settings. In this section, we address the problem of non-homogenous DMUs in the presence of such AR restrictions on input multipliers, and the inherent problems that can arise thereon. It should be noted that while the discussion focuses on input multipliers, the concepts apply equally to the output side.

Let us assume, in reference to model (11.3), that for each output subgroup $R_k$ AR constraints of the form $c^k_{iL} v_{i_2} \le v_{i_1} \le c^k_{iU} v_{i_2}, k = 1, \ldots, K$, have been specified. As indicated above, such constraints identify the relative magnitudes of pairs of input multipliers $v_{i_1}$ to $v_{i_2}$ within output subgroup $R_k$. It can be argued that all such constraints across all output subgroups $R_k$ can be expressed in the form $c^k_{iL} \le v_i/v_n \le c^k_{iU}$, such that $v_n$ is the designated *numeraire*, the multiplier for one of the inputs against which all other multipliers are compared (see Thompson et al. (1990)). Given that subgroups of outputs are in some senses a signal that multiple business units are operating under one umbrella, it is often the case that multiple sets of AR constraints on any given pair of multipliers can emerge simultaneously. Moreover, such multiple sets can result in infeasibility. For example, let us assume that the following constraints $3 \le v_2/v_1 \le 5$ and $6 \le v_2/v_1 \le 8$ have been specified for output subgroups $R_{k_1}$ and $R_{k_2}$, respectively. If these two sets of restrictions were to be imposed simultaneously on model (11.3), infeasibility would obviously result. This being the case, there is reason to look for a mechanism that will permit one to fold such multiple sets of constraints involving any multiplier $v_i$ into a single set, thereby insuring that model (11.3) is feasible.

Assume AR constraints $c_{iL}^{k_1} \leq \upsilon_i/\upsilon_n \leq c_{iU}^{k_1}$ and $c_{iL}^{k_2} \leq \upsilon_i/\upsilon_n \leq c_{iU}^{k_2}$ have been imposed within $R_{k_1}$ and $R_{k_2}$ respectively. To reduce this pair of AR restrictions to a single expression, we propose focusing attention on one of the bounds, say the lower bound. It is observed that by expressing

$$c_{iL}^{k_2} \leq \upsilon_i/\upsilon_n \leq c_{iU}^{k_2} \tag{11.6}$$

in the form $(c_{iL}^{k_1}/c_{iL}^{k_2})c_{iL}^{k_2} \leq (c_{iL}^{k_1}/c_{iL}^{k_2})\upsilon_i/\upsilon_n \leq (c_{iL}^{k_1}/c_{iL}^{k_2})c_{iU}^{k_2}$, and by subsequently making the following transformation $\upsilon_i' = (c_{iL}^{k_1}/c_{iL}^{k_2})\upsilon_i$, (11.6) can be converted to $c_{iL}^{k_1} \leq \upsilon_i'/\upsilon_n \leq (c_{iL}^{k_1}/c_{iL}^{k_2})c_{iU}^{k_2}$. Consequently, for each DMU $j_o \in M_{R_{k_2}}$, $\upsilon_i x_{ij}$ can be replaced by $\frac{c_{iL}^{k_1}}{c_{iL}^{k_2}}\upsilon_i x_{ij}\frac{c_{iL}^{k_2}}{c_{iL}^{k_1}}$ or $\upsilon_i' x_{ij}\frac{c_{iL}^{k_2}}{c_{iL}^{k_1}}$, meaning that by scaling the multiplier $\upsilon_i$ by a factor $\frac{c_{iL}^{k_1}}{c_{iL}^{k_2}}$, we can scale the data for $x_i$ in $M_{R_{k_2}}$ by the reciprocal of that factor.

To illustrate, refer again to the above example of constraints $3 \leq \upsilon_2/\upsilon_1 \leq 5$ and $6 \leq \upsilon_2/\upsilon_1 \leq 8$ in subgroups $R_{k_1}$ and $R_{k_2}$ respectively, and assume that input 1 is used as the numeraire. In order to arrive at a single set of constraints involving $\upsilon_2$ we first replace the constraint $6 \leq \upsilon_2/\upsilon_1 \leq 8$ with $\frac{3}{6}6 \leq \frac{3}{6}\upsilon_2/\upsilon_1 \leq \frac{3}{6}8$. By making the transformation $\upsilon_2' = \frac{3}{6}\upsilon_2$ we can then replace $\upsilon_2 x_2$ in $M_{R_{k_2}}$ with $\frac{3}{6}\upsilon_2(\frac{6}{3}x_2)$ or $\upsilon_2'(\frac{6}{3}x_2)$.

Specifically, by scaling $\upsilon_2$ down by a factor $\frac{3}{6}$ we can scale up the data for $x_2$ in subgroup $M_{R_{k_2}}$ by a factor $\frac{6}{3}$. Along the same lines, the upper bound on $\upsilon_i/\upsilon_n$ is replaced by $\bar{c}_{iU}^{k_2} = (\frac{c_{iL}^{k_1}}{c_{iL}^{k_2}})c_{iU}^{k_2}$.

This exercise is then repeated for all other output subgroups that have AR constraints involving multipliers $\upsilon_i$ and $\upsilon_n$. Let us define

$$\bar{c}_{iL} = c_{iL}^{k_1} \tag{11.7}$$

$$\bar{c}_{iU} = \min\{\bar{c}_{iU}^{k_1}, \bar{c}_{iU}^{k_2}, \ldots..\}, \tag{11.8}$$

where it is understood that the minimum in (11.8) is taken over all $R_k$ that contain an AR constraint involving the two variables $\upsilon_i$ and $\upsilon_n$. Expression (11.6) can now be replaced by

$$\bar{c}_{iL} \leq \frac{\upsilon_i}{\upsilon_n} \leq \bar{c}_{iU} \tag{11.9}$$

To repeat, assume a set of AR restrictions on a pair of input variables $(\upsilon_i, \upsilon_n)$ has been imposed within various output subgroups $R_k$. That is, the AR restrictions can vary by output subgroup. Let one of these subgroups, $R_{\hat{k}}$ be the base against which all other sets will be compared. As a result of the adjustments made to reduce these multiple restrictions to a single AR constraint, the corresponding inverse adjustments must be made to variable $x_i$ within each of the $R_k$ subgroups. (We are assuming that $\upsilon_n$ is the designated numeraire for this pair of variables). Let us now denote the adjusted input data by $x_{ikj}$, that is

$$x_{ikj} = (c_{iL}^{k}/c_{iL}^{\hat{k}})x_{ij} \tag{11.10}$$

With these adjustments having been made to the input data, model (11.3) for a given DMU $j_o$ in DMU group $N_{p^o}$ now takes the form

$$e_o = \max \sum_{R_k \in L_{N_{p^o}}} \sum_{r \in R_k} \mu_r y_{r j_o} \tag{11.11a}$$

*subject to*

$$\sum_{R_k \in L_{N_{p^o}}} \left( \sum_i \gamma_{i R_k p^o} x_{i k j^o} \right) = 1 \tag{11.11b}$$

$$\sum_{r \in R_k} \mu_r y_{rj} - \sum_i \gamma_{i R_k p} x_{ikj} \leq 0 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.11c}$$

$$\sum_{R_k \in L_{N_p}} \gamma_{i R_k p} = \upsilon_i \quad \forall i, \ p = 1 \ldots P \tag{11.11d}$$

$$\upsilon_i a_{i R_k p} \leq \gamma_{i R_k p} \leq \upsilon_i b_{i R_k p} \quad \forall i, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.11e}$$

$$\upsilon_n \bar{c}_{iL} \leq \upsilon_i \leq \upsilon_n \bar{c}_{iU} \quad \forall i, i \neq n \tag{11.11f}$$

$$\mu_r, \upsilon_i, \gamma_{i R_k p} \geq \varepsilon, \quad \forall r, i, R_k, p = 1, \ldots P \tag{11.11g}$$

Note that (11.11f) represents the final constraints resulting from the amalgamation of the multiple AR restrictions corresponding to the various $R_k$ subgroups.

In the case of the stage 2 subgroup optimization, where the efficiency of subgroup $R_{k^o}$ is to be determined, the AR-equivalent of model (11.5) is given by (11.12). Here, it is important to note that only AR restrictions relevant to this particular output subgroup, and no AR restrictions outside this subgroup are invoked. This being the case, input data requires AR-adjustment only in cases where multiple AR restrictions on a pair of variables arise. This latter can happen when constraints (on a given pair of variables) are invoked in one of the output subgroups $R_k$ that are different from those invoked in another subgroup. We use the notation $\hat{x}_{ij}^{k^o}$ to denote the alpha-adjusted, and AR- adjusted version of input $x_i$ consumed by output subgroup $R_{k^o}$. Constraints (11.12d) reflect the imposed AR constraints. The following model is now solved for each of the output subgroups, $R_{k^o}$ or each DMU $j_o$.

$$e_{R_{k^o} j_o} = \max \sum_{r \in R_{k^o}} \mu_r y_{r j_o} \tag{11.12a}$$

*subject to*

$$\sum_i \upsilon_i \hat{x}_{i j_o}^{k^o} = 1 \tag{11.12b}$$

$$\sum_{r \in R_{k^o}} \mu_r y_{rj} - \sum_i \upsilon_i \hat{x}_{ij}^{k^o} \leq 0, \ j \in N_p, \ for \ N_p \in M_{R_{k^\circ}} \tag{11.12c}$$

$$\upsilon_n c_{iL}^{k^\circ} \leq \upsilon_i \leq \upsilon_n c_{iU}^{k^\circ} \quad \forall i, i \neq n \tag{11.12d}$$

$$\mu_r, \upsilon_i \geq \varepsilon$$

## 11.5 Other Considerations

**Non-Separable Inputs** In many instances there can be inputs that do not lend themselves to subdivision in the manner described above. If, for example, in the analysis of the steel fabrication plants, one wished to include as an input a quality measure pertaining to supplier reliability, it would appear to be unreasonable to suggest subdividing this factor, and assigning portions of it across the various subunits; such a factor, in its entirety, would affect the outputs in each subunit $k$. Generalizing, let us use the notation $I_s, I_{ns}$ to denote the sets of separable and non-separable inputs respectively. In the discussion thus far, all inputs have been assumed to belong to $I_s$. The efficiency ratio for a given subgroup $R_k$ within DMU $j_o$ can now be expressed in the form $\sum_{r\in R_k} u_r y_{r j_o} / (\sum_{i\in I_s} v_i \alpha_{i R_k} x_{i j_o} + \sum_{i\in I_{ns}} v_i^{R_k} x_{i j_o})$ where $v_i^{R_k}$ is the worth or weight assigned to the non-separable input $x_{i j_o}$, $i \in I_{ns}$, and represents the impact of that input on the outputs in $R_k$. Note that we are permitting this weight to be different from one subgroup to another. Following the logic of (11.2) we define the weight attached to the efficiency ratio for subgroup $R_k$ by

$$W_{R_k j_o} = \Big[\sum_{i\in I_s} v_i \alpha_{i R_k p^o} x_{i j_o} + \sum_{i\in I_{ns}} v_i^{R_k} x_{i j_o}\Big] \Big/ \sum_{R_k \in L_{N_{p^o}}} \Big[\sum_{i\in I_s} v_i \alpha_{i R_k p^o} x_{i j_o} + \sum_{i\in I_{ns}} v_i^{R_k} x_{i j_o}\Big] \tag{11.13}$$

The optimization model for this more general case would be identical in form to (11.3) with the exception that constraint (11.3b) and (11.3c) are replaced by

$$\sum_{i\in I_s} v_i x_{i j_o} + \sum_{R_k \in N_{p^o}} \Big(\sum_{i\in I_{ns}} v_i^{R_k}\Big) x_{i j_o} = 1 \tag{11.3b'}$$

and

$$\sum_{r\in R_k} \mu_r y_{rj} - \sum_{i\in I_s} \gamma_{ki} x_{ij} - \sum_{i\in I_{ns}} v_i^{R_k} x_{ij} \le 0,\ \forall j \in N_p, \tag{11.13c'}$$

respectively, to account for the two types of inputs. Furthermore, constraints (11.4d) and (11.4e) apply only to separable inputs $i$. Here, $v_i^{R_k} = t v_i^{R_k}$ under the usual transformation as discussed above.

**Variable Returns to Scale** The development above is based on a CRS technology. In the situation where a VRS technology is deemed to be more appropriate, it is sufficient to replace terms such as $\sum_{r\in R_k} u_r y_{rj}$ by $\sum_{r\in R_k} u_r y_{rj} - u^o$. An advantage of the VRS formulation is that the sign of $u^o$ is subgroup-dependent, and signals whether the projected version of that (sub) DMU will be experiencing increasing, constant or decreasing returns to scale. This can provide useful information to management regarding the returns to scale orientation of various parts of the business, and may aid in deciding how to redistribute resources, given that common resources ($I_s$) are shared among the subgroups.

**Output Orientation** The development throughout has assumed that efficiency is to be viewed from the perspective of an input orientation. If the organization intends to improve efficiency by pursuing output expansion rather than input reduction, then an output orientation would be an appropriate model structure to use. Specifically, model (3.1) would be replaced by:

$$e_o = \min \sum_{R_k \in L_{N_{p^o}}} W_{R_k j_o} \Big[ \sum_i v_i \alpha_{i R_k p^o} x_{i j_o} / \sum_{r \in R_k} u_r y_{r j_o} \Big] \tag{11.13a}$$

*subject to*

$$\sum_{R_k \in L_{N_p}} W_{R_k j} \Big[ \sum_i v_i \alpha_{i R_k p} x_{ij} / \sum_{r \in R_k} u_r y_{rj} \Big] \geq 1 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.13b}$$

$$\sum_i v_i \alpha_{i R_k p} x_{ij} - \sum_{r \in R_k} u_r y_{rj} \geq 0 \quad \forall j \in N_p, R_k \in L_{N_p}, p = 1, \ldots P \tag{11.13c}$$

$$\sum_{R_k \in L_{Np}} \alpha_{i R_k p} = 1 \quad \forall i, \ p = 1 \ldots P \tag{11.13d}$$

$$a_{i R_k p} \leq \alpha_{i R_k p} \leq b_{i R_k p} \quad \forall i, \ R_k, \ p = 1, \ldots P \tag{11.13e}$$

$$u_r, v_i, \alpha_{i R_k p} \geq 0, \quad \forall i, R_k, p \tag{11.13f}$$

As discussed earlier, it appears equally valid to base the definition of weights $W_{R_k j_o}$ on either inputs consumed or outputs generated. In the case of the output oriented model it is therefore reasonable to weight the efficiency ratios according to the latter (outputs generated). Specifically, if the weight on subgroup $R_k$ is chosen as

$$W_{R_k j_o} = \sum_{r \in R_k} u_r y_{r j_o} / \sum_{R_k \in L_{N_{p^o}}} \Big[ \sum_{r \in R_k} u_r y_{r j_o} \Big] \tag{11.14}$$

then the above aggregate objective function (11.13a) becomes the ratio of overall DMU input to overall DMU output, namely,

$$e_o = \min \sum_{R_k \in L_{N_{p^o}}} \Big[ \sum_i v_i x_{i j_o} / \sum_{r \in R_k} u_r y_{r j_o} \Big] \tag{11.13a'}$$

We now apply the above methodology to the derivation of efficiencies of a set of steel fabrication plants.

## 11.6 Application

To demonstrate the application of the models developed in the earlier sections, data on a set of 47 plants, as per Appendix 3 Tables 11.8 and 11.9, were considered. These plants are grouped into four DMU subgroups $N_1$ to $N_4$ such that plants

**Table 11.1** Input cost rates per machine per quarter

| | Quarterly costs thousands of dollars | | | | |
|---|---|---|---|---|---|
| Input | k = 1 | k = 2 | k = 3 | k = 4 | k = 5 |
| Labor | $5–$7.5 | $5–$7.5 | $5–$7.5 | $5–$7.5 | $5–$7.5 |
| Shears | $7–$10 | $14–$19 | $14–$19 | $7–$10 | $9–$12 |
| Presses | $6–$9.6 | $16–$21 | $12–$15 | $6–$9.6 | $5–$9 |
| Lathes | $7–$11 | $7–$11 | $7–$11 | $7–$11 | $7–$11 |

belonging to any DMU group $N_p$ produce identical products. The profiles of the four groups of DMUs $N_p$ are as described in Sect. 2. For example, DMUs in $N_1$ produce outputs 1,2,3 and 5. Note from Appendix 3 Table 11.8, that $N_1$ consists of DMUs 1,3,4,6,9,10,11,19,32,35,39,and 46.

In the application considered herein, AR restrictions play an important role, particularly on the input side. They provide a way to bring resource tradeoffs into the picture. Input multipliers effectively mimic resource costs. Hence, while it is the case in many real world settings that the development of such restrictions can be problematic, in manufacturing situations resource costs often provide the appropriate route to deriving the desired restrictions. To that end data was collected relating to per unit costs for each of the four inputs. The data provided in Table 11.1 represent per unit costs incurred during the last quarter of 2010. For example, the range of cost estimates specified for labor ($x_1$) for the last quarter of 2010 is from $5000 to $7500 per plant employee (wages and benefits). While there is no implied variation in labor costs across the five bundles, $k = 1,2,3,4,5$, wage rates can differ from plant to plant and over time due to the mix of full time and part time labor used. For this reason a range is given for this input.

For the other three inputs, machine 'rates' were assumed to be the estimated quarterly costs of depreciation, routine maintenance and unforeseen breakdown costs. In the case of the shearing machines, for example, the estimated quarterly cost (depreciation and maintenance) of operating one machine would generally vary between $7000 and $10,000 per quarter in the case of output bundle $k = 1$, and $14,000 and $19,000 in the case of $k = 2$. The increased stress placed on the equipment in the production of flat bar products versus that created in the manufacture of sheet steel products, contributes to the difference in cost between the two product groupings.

Table 11.1 can be used to set AR constraints corresponding to the various pairs of multipliers. See Table 11.2 for the full set. Note that the lower bounds on all constraints $c^k_{iL} \leq \upsilon_i/\upsilon_n \leq c^k_{iU}$, can be expressed as the ratio of the lowest value $\upsilon_i$ can take divided by the highest value $\upsilon_n$ can assume. Similarly, the upper bounds are defined as the ratio of the highest value $\upsilon_i$ can take divided by the lowest value taken by $\upsilon_n$.For example, given that the range for labor cost is $5–$7.5 and the range for shears is $7–$10 in the case of k = 1, the AR constraints corresponding to $\upsilon_2$ and $\upsilon_1$ are expressed as $\frac{7}{7.5} \leq \frac{\upsilon_2}{\upsilon_1} \leq \frac{10}{5}$.

**Table 11.2** AR constraints

| | i = 2 | i = 3 | i = 4 |
|---|---|---|---|
| $K = 1$ | $0.93 \leq \frac{v_2}{v_1} \leq 2$ | $0.8 \leq \frac{v_3}{v_1} \leq 1.92$ | $0.93 \leq \frac{v_4}{v_1} \leq 2.2$ |
| $K = 2$ | $1.87 \leq \frac{v_2}{v_1} \leq 3.8$ | $2.13 \leq \frac{v_3}{v_1} \leq 4.2$ | $0.93 \leq \frac{v_4}{v_1} \leq 2.2$ |
| $K = 3$ | $1.87 \leq \frac{v_2}{v_1} \leq 3.8$ | $1.6 \leq \frac{v_3}{v_1} \leq 3$ | $0.93 \leq \frac{v_4}{v_1} \leq 2.2$ |
| $K = 4$ | $0.93 \leq \frac{v_2}{v_1} \leq 2$ | $0.8 \leq \frac{v_3}{v_1} \leq 1.92$ | $0.93 \leq \frac{v_4}{v_1} \leq 2.2$ |
| $K = 5$ | $1.2 \leq \frac{v_2}{v_1} \leq 2.4$ | $0.67 \leq \frac{v_3}{v_1} \leq 1.8$ | $0.93 \leq \frac{v_4}{v_1} \leq 2.2$ |

All constraints are expressed in terms of labor ($v_1$) which has been chosen as the numeraire. Refer to Appendix 2 for a detailed discussion on generating a single set of AR constraints for each pair of input multipliers.

Following the methodology presented in Sect. 3, model (11.3) is applied to the data of Appendix 3 Tables 11.8 and 11.9. Recall that the purpose of solving this stage 1 problem is to facilitate an apportioning of the inputs to the subunits that make up the DMU. To bound the values of $\alpha$ so that a representative apportioning occurs, survey data from a sample of the plants suggested the following ranges:

N1: (0.15, 0.80)
N2: (0.10, 0.60)
N3: (0.20, 0.90)
N4: (0.20, 0.90)

It is noted that the ranges vary according to the DMU subgroup $N_p$, and is related to the number of subunits $K$ comprising the subgroup. Specifically, the more subunits that $N_p$ contains, the narrower are the ranges. Recall that

$$L_{N_1} = \{R_1, R_2, R_3\},\ L_{N_2} = \{R_2, R_3, R_4, R_5\},\ L_{N_3} = \{R_3, R_5\},\ L_{N_4} = \{R_1, R_3\}.$$

For example, since DMU subgroup $N_2$ contains 4 subunits, a minimum of 10 % and a maximum of 60 % of each input can be assigned to any subunit. In the case of $N_3$, however, which contains only 2 subunits, the alpha range is wider.

Applying model (11.3), the $\hat{\alpha}_{iR_kp}$ for each DMU $j_o$ in $N_p$ have been derived. The results are displayed in Appendix 3, Tables 11.10, 11.11, 11.12, 11.13, 11.14. Recall that $\hat{\alpha}_{iR_kp}$ values are used to adjust the corresponding data in each DMU subgroup $M_{R_k}$, in preparation for the subunit analysis. Specifically, using the appropriately adjusted data, model (11.5) is applied to each DMU in $M_{R_k}$, resulting in the subunit scores displayed in Appendix 3 Table 11.16. To derive an overall efficiency score for each DMU $j_o$ in $N_p$, the relevant subunit scores are combined using the weights $W_{R_kj_o}$, as per Appendix 3 Table 11.15. The resulting overall scores are presented along with their relevant subunit scores in Appendix 3 Table 11.16.

It is noted that none of the DMUs are technically efficient. Recall that a DMU can be efficient only if all subunits for that DMU are efficient as well. However, within each subgroup $R_k$ at least one of the (sub) DMUs in $M_{R_k}$ is inefficient.

To demonstrate the degree of sensitivity of the overall efficiency scores (as per the right-most column in Appendix 3 Table 11.16) to the choice of alpha ranges, the above analysis was repeated, but with two new sets of alpha ranges. A summary of the results is as follows:

| Scenario | Lower and upper limits on $\hat{\alpha}_{iR_kp}$ | Average absolute change in overall efficiency scores |
|---|---|---|
| 1 | (0.10, 0.80) | 0.12219 |
| 2 | (0.05, 0.90) | 0.33868 |

Based on this, it would appear that very wide ranges such as those given as scenario 2, result in substantial swings in the overall efficiency scores when compared with base results as described above. Somewhat tighter ranges such as those in Scenario 1 significantly decrease the variation in efficiency scores vis a vis the base results.

To complete the analysis of this section, it is worth comparing the efficiency results obtained using our model with what would have occurred had conventional DEA analysis been carried out, by simply inserting zeros in the data for any missing outputs. Two levels of analysis were conducted, namely one without any AR constraints, and one with the AR constraints applied. The results are displayed in Appendix 3, Table 11.17. It is noted that having replaced all blank spaces with zeros, a significant number of DMUs are rendered technically efficient. In the non-AR versions of our model and the conventional DEA model, we note that there are 3 efficient DMUs in the former versus 33 in the latter. The existence of the very large number of efficient units (33) with the conventional model is partially due to the large number of outputs and inputs involved, as compared to the total number of DMUs. A somewhat more realistic set of scores arise with the conventional model in the presence of AR constraints, where only 17 of the DMUs are efficient. Specifically, 17 of the 47 DMUs have a score of 100 % and another 8 have scores at the level of 90 % or above. Arguably, part of the problem as well is that the absence of outputs in the various DMUs may be providing the opportunity to DMUs in any given subset $N_p$ to negate the influence of other DMUs that are in different subsets.

## 11.7 Conclusions

This chapter has examined efficiency measurement in a setting where decision making units are non-homogeneous. This environment violates the usual assumption in DEA that DMUs are all in the same 'business', meaning that each DMU produces some amount of each output in a given output bundle, albeit in differing amounts from one DMU to the next. The problem of "missing" outputs has been addressed in the literature, but only in the context that either the missing value exists, but is not available to the analyst, or that the missing item is a quantity that the DMU intended to produce (and resources were expended in an effort to do so), but for whatever

reason none was actually created. In this case, the value assigned to that output is legitimately taken to be *zero*.

Herein we argue that in many situations the output mix can differ substantially from one DMU to another, meaning that the usual assumption of homogeneity does not hold, and therefore the DMUs involved are not directly comparable. Substituting zero or some other computed value when an output is missing, as a means of rendering DMUs 'comparable', appears to be ad hoc, and fails to properly address the efficiency evaluation problem in a direct way. To address this apparent gap in the DEA literature, we present a DEA-like methodology that views the DMU as consisting of a set of business subunits. The overall efficiency of a DMU is then taken to be a weighted average (convex combination) of the efficiency scores for the subgroups that make up that DMU. The methodology is applied to the efficiency evaluation problem for a set of steel fabrication plants.

One criticism of this approach is that it presumes that the DMU can be viewed as being the sum of its parts, meaning that economies or diseconomies of scope are assumed to be non-existent. In cases where this assumption is violated, our approach may fail to accurately capture the performance of the DMU. See, for example, Pulley and Braunstein (1992). Capturing economies of scope is difficult in settings where one does not have the benefit of observing an entity operating by itself as well as in a mode where it is combined with other entities. In noted from the literature, data on mergers and acquisitions can be a way in which one might reasonably examine an entity in both states. There is also the added difficulty of separating economies of scope from economies of scale. Is the increase in output of a given product, when a new product is added to the mix, due to scope or simply a result of increased size of the operation (scale)? Further research is encouraged in this area.

This chapter is based upon (i) W. Cook, R. Imanirad, J. Harrison, P. Rouse and J. Zhu. 2013. Data Envelopment Analysis with Non-Homogeneous Decision Making Units, **Operations Research,** 61(3), 666–676; with permission from the Institute for Operations Research and the Management Sciences, 5521 Research Park Drive, Suite 200, Catonsville, MD 21228 USA, and (ii) W. Cook, J. Harrison, P. Rouse and J. Zhu. 2012. "Relative Efficiency Measurement: The Problem of a Missing Output in a Subset of Decision Making Units", **European Journal of Operational Research**, 220 (1), 79–84; with permission from Elsevier.

## 11.8 Appendix 1: Generating the Maximal Output Groupings

The algorithm for generating the maximal output groupings is as follows:

*Step 1:* Define S to be an empty set.

*Step 2:* For each output $r$, derive $N(r)$, the set of all DMU subgroups $N_p$ that produce $r$. Add $N(r)$ to S.

*Step 3:* For each $N(r)$ in S, compare it with every other $N(r')$ in S, and identify all $N(r')$ that have the same elements as $N(r)$. If no such $r'$ is identified, create set $R^k_{k=1} = \{r\}$. Remove $N(r)$ from S. Otherwise, create set $R^k_{k=1} = \{r\}$ and add all $r'$ to $R_k$. Remove $N(r)$ and all $N(r')$ from S.

**Table 11.3** Multiple sets of AR constraints across all output subgroups

| | i = 2 | i = 3 | i = 4 |
|---|---|---|---|
| $K=1$ | $\frac{7}{7.5} \le \frac{v_2}{v_1} \le \frac{10}{5}$ | $\frac{6}{7.5} \le \frac{v_3}{v_1} \le \frac{9.6}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=2$ | $\frac{14}{7.5} \le \frac{v_2}{v_1} \le \frac{19}{5}$ | $\frac{16}{7.5} \le \frac{v_3}{v_1} \le \frac{21}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=3$ | $\frac{14}{7.5} \le \frac{v_2}{v_1} \le \frac{19}{5}$ | $\frac{12}{7.5} \le \frac{v_3}{v_1} \le \frac{15}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=4$ | $\frac{7}{7.5} \le \frac{v_2}{v_1} \le \frac{10}{5}$ | $\frac{6}{7.5} \le \frac{v_3}{v_1} \le \frac{9.6}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=5$ | $\frac{9}{7.5} \le \frac{v_2}{v_1} \le \frac{12}{5}$ | $\frac{5}{7.5} \le \frac{v_3}{v_1} \le \frac{9}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |

**Table 11.4** Adjustments made to the AR constraints

| | i = 2 | i = 3 | i = 4 |
|---|---|---|---|
| $K=1$ | $\frac{7}{7.5} \le \frac{v_2}{v_1} \le \frac{10}{5}$ | $\frac{5}{6} \times \frac{6}{7.5} \le \frac{5}{6} \times \frac{v_3}{v_1} \le \frac{5}{6} \times \frac{9.6}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=2$ | $\frac{1}{2} \times \frac{14}{7.5} \le \frac{1}{2} \times \frac{v_2}{v_1} \le \frac{1}{2} \times \frac{19}{5}$ | $\frac{5}{16} \times \frac{16}{7.5} \le \frac{5}{16} \times \frac{v_3}{v_1} \le \frac{5}{16} \times \frac{21}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=3$ | $\frac{1}{2} \times \frac{14}{7.5} \le \frac{1}{2} \times \frac{v_2}{v_1} \le \frac{1}{2} \times \frac{19}{5}$ | $\frac{5}{12} \times \frac{12}{7.5} \le \frac{5}{12} \times \frac{v_3}{v_1} \le \frac{5}{12} \times \frac{15}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=4$ | $\frac{7}{7.5} \le \frac{v_2}{v_1} \le \frac{10}{5}$ | $\frac{5}{6} \times \frac{6}{7.5} \le \frac{5}{6} \times \frac{v_3}{v_1} \le \frac{5}{6} \times \frac{9.6}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |
| $K=5$ | $\frac{7}{9} \times \frac{9}{7.5} \le \frac{7}{9} \times \frac{v_2}{v_1} \le \frac{7}{9} \times \frac{12}{5}$ | $\frac{5}{7.5} \le \frac{v_3}{v_1} \le \frac{9}{5}$ | $\frac{7}{7.5} \le \frac{v_4}{v_1} \le \frac{11}{5}$ |

**Table 11.5** Transformation of multipliers

| | i = 2 | i = 3 |
|---|---|---|
| $K=1$ | – | $v_3' = (\frac{5}{6})v_3$ |
| $K=2$ | $v_2' = (\frac{1}{2})v_2$ | $v_3' = (\frac{5}{16})v_3$ |
| $K=3$ | $v_2' = (\frac{1}{2})v_2$ | $v_3' = (\frac{5}{12})v_3$ |
| $K=4$ | – | $v_3' = (\frac{5}{6})v_3$ |
| $K=5$ | $v_2' = (\frac{7}{9})v_2$ | – |

## 11.9 Appendix 2: Generating a Single Set of AR Constraints

Given that $v_1$ is the numeraire against which all other multipliers are compared, Table 11.3 displays all AR constraints across all output subgroups expressed in the form $c_{iL}^k \le v_i/v_1 \le c_{iU}^k$.

In order to arrive at a single set of constraints for each pair of multipliers, we assume that $R_{k=1}$ is the base subgroup for $v_{i=2}$ and $R_{k=5}$ is the base subgroup for $v_{i=3}$ against which all other sets will be compared. Table 11.4 demonstrates the required adjustments made to the AR constraints.

Next, by making the transformations displayed in Table 11.5, the adjusted AR constraints are derived as presented in Table 11.6.

Subsequent to adjusting the corresponding input data within each subgroup $M_{R_k}$ and considering

$$\bar{c}_{iL} = c_{iL}^{k_1}$$

$$\bar{c}_{iU} = \min\{\bar{c}_{iU}^{k_1}, \bar{c}_{iU}^{k_2}, \ldots..\},$$

**Table 11.6** Adjusted AR constraints

| | i = 2 | i = 3 | i = 4 |
|---|---|---|---|
| $K = 1$ | $\frac{7}{7.5} \leq \frac{v_2}{v_1} \leq \frac{10}{5}$ | $\frac{5}{7.5} \leq \frac{v_3'}{v_1} \leq \frac{9.6}{6}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |
| $K = 2$ | $\frac{7}{7.5} \leq \frac{v_2'}{v_1} \leq \frac{19}{10}$ | $\frac{5}{7.5} \leq \frac{v_3'}{v_1} \leq \frac{21}{16}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |
| $K = 3$ | $\frac{7}{7.5} \leq \frac{v_2'}{v_1} \leq \frac{19}{10}$ | $\frac{5}{7.5} \leq \frac{v_3'}{v_1} \leq \frac{15}{12}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |
| $K = 4$ | $\frac{7}{7.5} \leq \frac{v_2}{v_1} \leq \frac{10}{5}$ | $\frac{5}{7.5} \leq \frac{v_3'}{v_1} \leq \frac{9.6}{6}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |
| $K = 5$ | $\frac{7}{7.5} \leq \frac{v_2'}{v_1} \leq \frac{28}{15}$ | $\frac{5}{7.5} \leq \frac{v_3}{v_1} \leq \frac{9}{5}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |

**Table 11.7** Reduced set of AR constraints

| | i = 2 | i = 3 | i = 4 |
|---|---|---|---|
| | $\frac{7}{7.5} \leq \frac{v_2}{v_1} \leq \frac{28}{15}$ | $\frac{5}{7.5} \leq \frac{v_3}{v_1} \leq \frac{5}{4}$ | $\frac{7}{7.5} \leq \frac{v_4}{v_1} \leq \frac{11}{5}$ |

for each subgroup, where $R_{k_1}$ is the base, all AR constraints corresponding to the various pairs of multipliers can now be reduced to a single pair of constraints for each pair of multipliers as displayed in Table 11.7.

## 11.10 Appendix 3: Tables

**Table 11.8** Data on 47 plants-outputs

| Outputs | | | | | | |
|---|---|---|---|---|---|---|
| | Sheet steel | Flat bar | Pipes/ cylinders | Ducts | Structural steel | Storage tanks |
| DMU | *Y1* | *Y2* | *Y3* | *Y4* | *Y5* | *Y6* |
| 1 | 70 | 103 | 100 | – | 60 | – |
| 2 | – | 125 | 90 | 123 | 48 | 133 |
| 3 | 50 | 110 | 105 | – | 170 | – |
| 4 | 80 | 80 | 110 | – | 82 | – |
| 5 | – | – | 60 | – | 100 | 150 |
| 6 | 40 | 95 | 120 | – | 151 | – |
| 7 | 100 | – | 200 | – | 64 | – |
| 8 | – | – | 180 | – | 104 | 66 |
| 9 | 65 | 150 | 125 | – | 93 | – |
| 10 | 40 | 110 | 70 | – | 79 | – |
| 11 | 70 | 117 | 122 | – | 132 | – |
| 12 | – | – | 89 | – | 80 | 189 |
| 13 | 88 | – | 57 | – | 150 | – |

**Table 11.8** (continued)

| Outputs | | | | | | |
|---|---|---|---|---|---|---|
| | Sheet steel | Flat bar | Pipes/ cylinders | Ducts | Structural steel | Storage tanks |
| DMU | *Y1* | *Y2* | *Y3* | *Y4* | *Y5* | *Y6* |
| 14 | 48 | – | 146 | – | 162 | – |
| 15 | – | – | 220 | – | 111 | 73 |
| 16 | 99 | – | 89 | – | 56 | – |
| 17 | – | – | 88 | – | 41 | 161 |
| 18 | – | 55 | 132 | 129 | 112 | 113 |
| 19 | 80 | 97 | 142 | – | 82 | – |
| 20 | 97 | – | 209 | – | 106 | – |
| 21 | – | – | 55 | – | 157 | 130 |
| 22 | – | – | 93 | – | 163 | 55 |
| 23 | 59 | | 218 | – | 79 | – |
| 24 | 61 | – | 58 | – | 75 | – |
| 25 | 68 | – | 110 | – | 48 | – |
| 26 | – | – | 86 | – | 109 | 69 |
| 27 | – | 65 | 166 | 41 | 183 | 137 |
| 28 | – | – | 228 | – | 199 | 71 |
| 29 | – | – | 95 | – | 110 | 54 |
| 30 | 50 | – | 77 | – | 89 | – |
| 31 | – | 138 | 206 | 68 | 102 | 74 |
| 32 | 36 | 106 | 167 | – | 130 | – |
| 33 | – | 84 | 98 | 45 | 176 | 69 |
| 34 | – | 62 | 120 | 57 | 58 | 154 |
| 35 | 24 | 135 | 185 | – | 112 | – |
| 36 | – | – | 144 | – | 196 | 78 |
| 37 | 58 | – | 178 | – | 147 | – |
| 38 | – | 123 | 206 | 63 | 195 | 57 |
| 39 | 41 | 110 | 225 | – | 53 | – |
| 40 | – | – | 188 | – | 60 | 127 |
| 41 | 70 | – | 140 | – | 150 | – |
| 42 | – | – | 55 | – | 70 | 191 |
| 43 | 45 | – | 124 | – | 139 | – |
| 44 | 63 | – | 161 | – | 125 | – |
| 45 | 85 | – | 81 | – | 90 | – |
| 46 | 42 | 78 | 69 | – | 82 | – |
| 47 | 25 | – | 184 | – | 162 | – |

**Table 11.9** Data on 47 plants-inputs

| Inputs | | | | |
|---|---|---|---|---|
| | Labor | Shears | Presses | Torches |
| DMU | *X1* | *X2* | *X3* | *X4* |
| 1 | 30 | 5.0 | 3.0 | 15 |
| 2 | 40 | 4.0 | 6.5 | 18 |
| 3 | 35 | 5.2 | 4.2 | 10 |
| 4 | 38 | 7.0 | 7.6 | 9 |
| 5 | 28 | 9.0 | 5.5 | 13 |
| 6 | 37 | 4.2 | 3.8 | 17 |
| 7 | 31 | 6.0 | 4.1 | 11 |
| 8 | 35 | 5.0 | 7.0 | 15 |
| 9 | 25 | 6.2 | 4.8 | 19 |
| 10 | 30 | 3.0 | 3.2 | 21 |
| 11 | 25 | 4.0 | 6.0 | 12 |
| 12 | 45 | 5.0 | 3.3 | 23 |
| 13 | 35 | 4.1 | 5.0 | 25 |
| 14 | 32 | 5.3 | 3.5 | 11 |
| 15 | 26 | 7.7 | 4.3 | 16 |
| 16 | 19 | 5.3 | 6.2 | 12 |
| 17 | 25 | 8.0 | 3.0 | 9 |
| 18 | 32 | 6.0 | 2.8 | 7 |
| 19 | 33 | 2.8 | 3.9 | 13 |
| 20 | 27 | 3.3 | 4.3 | 22 |
| 21 | 25 | 7.9 | 5.0 | 16 |
| 22 | 34 | 5.0 | 5.4 | 20 |
| 23 | 45 | 4.0 | 4.1 | 12 |
| 24 | 24 | 5.1 | 3.4 | 19 |
| 25 | 33 | 8.6 | 2.7 | 10 |
| 26 | 21 | 9.8 | 5.5 | 5 |
| 27 | 25 | 7.0 | 3.1 | 23 |
| 28 | 38 | 4.5 | 2.4 | 10 |
| 29 | 33 | 3.2 | 4.6 | 24 |
| 30 | 27 | 6.4 | 3.0 | 7 |
| 31 | 20 | 5.8 | 5.1 | 18 |
| 32 | 39 | 8.4 | 3.8 | 16 |

**Table 11.9** (continued)

| Inputs | | | | |
|---|---|---|---|---|
| | Labor | Shears | Presses | Torches |
| DMU | *X1* | *X2* | *X3* | *X4* |
| 33 | 42 | 6.5 | 2.4 | 8 |
| 34 | 44 | 4.3 | 3.0 | 22 |
| 35 | 26 | 3.7 | 6.7 | 20 |
| 36 | 43 | 7.5 | 7.1 | 8 |
| 37 | 35 | 6.8 | 4.7 | 14 |
| 38 | 22 | 3.9 | 3.2 | 25 |
| 39 | 41 | 6.7 | 2.5 | 21 |
| 40 | 21 | 5.2 | 4.9 | 10 |
| 41 | 33 | 3.5 | 5.9 | 7 |
| 42 | 20 | 4.7 | 7.2 | 23 |
| 43 | 39 | 9.5 | 5.4 | 6 |
| 44 | 48 | 6.7 | 6.2 | 18 |
| 45 | 31 | 3.6 | 4.7 | 23 |
| 46 | 28 | 9.2 | 2.6 | 5 |
| 47 | 36 | 7.6 | 5.7 | 10 |

**Table 11.10** $\alpha_{iR_k}$ Values resulting from model (11.11)—$N_1$

| DMU | $X_1$ K=1 | $X_1$ K=2 | $X_1$ K=3 | $X_2$ K=1 | $X_2$ K=2 | $X_2$ K=3 | $X_3$ K=1 | $X_3$ K=2 | $X_3$ K=3 | $X_4$ K=1 | $X_4$ K=2 | $X_4$ K=3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.70 | 0.15 | 0.15 | 0.15 | 0.70 | 0.15 | 0.61 | 0.24 | 0.15 | 0.15 | 0.70 | 0.15 |
| 3 | 0.15 | 0.16 | 0.69 | 0.15 | 0.23 | 0.62 | 0.18 | 0.15 | 0.67 | 0.27 | 0.58 | 0.16 |
| 4 | 0.70 | 0.15 | 0.15 | 0.69 | 0.16 | 0.15 | 0.70 | 0.15 | 0.15 | 0.27 | 0.49 | 0.24 |
| 6 | 0.15 | 0.15 | 0.70 | 0.15 | 0.15 | 0.70 | 0.61 | 0.24 | 0.15 | 0.34 | 0.51 | 0.15 |
| 9 | 0.38 | 0.47 | 0.15 | 0.15 | 0.70 | 0.15 | 0.30 | 0.55 | 0.15 | 0.15 | 0.70 | 0.15 |
| 10 | 0.43 | 0.42 | 0.15 | 0.15 | 0.70 | 0.15 | 0.30 | 0.55 | 0.15 | 0.15 | 0.70 | 0.15 |
| 11 | 0.64 | 0.15 | 0.21 | 0.15 | 0.70 | 0.15 | 0.40 | 0.32 | 0.28 | 0.38 | 0.24 | 0.38 |
| 19 | 0.61 | 0.15 | 0.24 | 0.15 | 0.70 | 0.15 | 0.41 | 0.31 | 0.28 | 0.26 | 0.23 | 0.51 |
| 32 | 0.38 | 0.18 | 0.44 | 0.68 | 0.17 | 0.15 | 0.25 | 0.22 | 0.53 | 0.15 | 0.43 | 0.42 |
| 35 | 0.15 | 0.70 | 0.15 | 0.26 | 0.59 | 0.15 | 0.63 | 0.15 | 0.22 | 0.15 | 0.35 | 0.50 |
| 39 | 0.23 | 0.15 | 0.62 | 0.70 | 0.15 | 0.15 | 0.38 | 0.45 | 0.17 | 0.15 | 0.15 | 0.70 |
| 46 | 0.40 | 0.25 | 0.34 | 0.54 | 0.31 | 0.15 | 0.15 | 0.54 | 0.31 | 0.15 | 0.70 | 0.15 |

**Table 11.11** $\alpha_{iR_k}$ Values resulting from model (11.11)—$N_2$ ($X_1$, $X_2$)

| DMU | $X_1$ K=2 | $X_1$ K=3 | $X_1$ K=4 | $X_1$ K=5 | $X_2$ K=2 | $X_2$ K=3 | $X_2$ K=4 | $X_2$ K=5 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.10 | 0.10 | 0.36 | 0.44 | 0.20 | 0.10 | 0.10 | 0.60 |
| 18 | 0.10 | 0.10 | 0.58 | 0.22 | 0.10 | 0.60 | 0.10 | 0.20 |
| 27 | 0.10 | 0.28 | 0.10 | 0.52 | 0.10 | 0.10 | 0.20 | 0.60 |
| 31 | 0.60 | 0.20 | 0.10 | 0.10 | 0.60 | 0.20 | 0.10 | 0.10 |
| 33 | 0.10 | 0.60 | 0.10 | 0.20 | 0.10 | 0.60 | 0.10 | 0.20 |
| 34 | 0.10 | 0.10 | 0.20 | 0.60 | 0.10 | 0.10 | 0.20 | 0.60 |
| 38 | 0.20 | 0.60 | 0.10 | 0.10 | 0.20 | 0.60 | 0.10 | 0.10 |

**Table 11.12** $\alpha_{iR_k}$ Values resulting from model (11.11)—$N_2$ ($X_3$, $X_4$)

| DMU | $X_3$ K=2 | $X_3$ K=3 | $X_3$ K=4 | $X_3$ K=5 | $X_4$ K=2 | $X_4$ K=3 | $X_4$ K=4 | $X_4$ K=5 |
|---|---|---|---|---|---|---|---|---|
| 2 | 0.60 | 0.10 | 0.10 | 0.20 | 0.60 | 0.10 | 0.10 | 0.20 |
| 18 | 0.10 | 0.60 | 0.20 | 0.10 | 0.44 | 0.36 | 0.10 | 0.10 |
| 27 | 0.10 | 0.60 | 0.10 | 0.20 | 0.10 | 0.48 | 0.10 | 0.32 |
| 31 | 0.60 | 0.20 | 0.10 | 0.10 | 0.39 | 0.41 | 0.10 | 0.10 |
| 33 | 0.60 | 0.20 | 0.10 | 0.10 | 0.27 | 0.53 | 0.10 | 0.10 |
| 34 | 0.38 | 0.17 | 0.10 | 0.34 | 0.18 | 0.26 | 0.10 | 0.46 |
| 38 | 0.60 | 0.20 | 0.10 | 0.10 | 0.35 | 0.45 | 0.10 | 0.10 |

**Table 11.13** $\alpha_{iR_k}$ Values resulting from model (11.11)—$N_3$

| DMU | $X_1$ K=3 | $X_1$ K=5 | $X_2$ K=3 | $X_2$ K=5 | $X_3$ K=3 | $X_3$ K=5 | $X_4$ K=3 | $X_4$ K=5 |
|---|---|---|---|---|---|---|---|---|
| 5 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 |
| 8 | 0.80 | 0.20 | 0.61 | 0.39 | 0.20 | 0.80 | 0.80 | 0.20 |
| 12 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 |
| 15 | 0.80 | 0.20 | 0.34 | 0.66 | 0.46 | 0.54 | 0.80 | 0.20 |
| 17 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 |
| 21 | 0.80 | 0.20 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 |
| 22 | 0.80 | 0.20 | 0.56 | 0.44 | 0.20 | 0.80 | 0.80 | 0.20 |
| 26 | 0.80 | 0.20 | 0.20 | 0.80 | 0.20 | 0.80 | 0.80 | 0.20 |
| 28 | 0.80 | 0.20 | 0.77 | 0.23 | 0.80 | 0.20 | 0.80 | 0.20 |
| 29 | 0.80 | 0.20 | 0.20 | 0.80 | 0.20 | 0.80 | 0.80 | 0.20 |
| 36 | 0.80 | 0.20 | 0.41 | 0.59 | 0.20 | 0.80 | 0.80 | 0.20 |
| 40 | 0.80 | 0.20 | 0.25 | 0.75 | 0.80 | 0.20 | 0.63 | 0.37 |
| 42 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 |

**Table 11.14** $\alpha_{iR_k}$ Values resulting from model (11.11)—$N_4$

| DMU | $X_1$ K=1 | $X_1$ K=3 | $X_2$ K=1 | $X_2$ K=3 | $X_3$ K=1 | $X_3$ K=3 | $X_4$ K=1 | $X_4$ K=3 |
|---|---|---|---|---|---|---|---|---|
| 7 | 0.28 | 0.72 | 0.80 | 0.20 | 0.80 | 0.20 | 0.62 | 0.38 |
| 13 | 0.21 | 0.79 | 0.74 | 0.26 | 0.80 | 0.20 | 0.80 | 0.20 |
| 14 | 0.20 | 0.80 | 0.80 | 0.20 | 0.66 | 0.34 | 0.35 | 0.65 |
| 16 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 |
| 20 | 0.25 | 0.75 | 0.80 | 0.20 | 0.80 | 0.20 | 0.30 | 0.70 |
| 23 | 0.29 | 0.71 | 0.20 | 0.80 | 0.80 | 0.20 | 0.44 | 0.56 |
| 24 | 0.21 | 0.79 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 |
| 25 | 0.32 | 0.68 | 0.80 | 0.20 | 0.80 | 0.20 | 0.68 | 0.32 |
| 30 | 0.20 | 0.80 | 0.80 | 0.20 | 0.77 | 0.23 | 0.43 | 0.57 |
| 37 | 0.20 | 0.80 | 0.80 | 0.20 | 0.48 | 0.52 | 0.60 | 0.40 |
| 41 | 0.20 | 0.80 | 0.36 | 0.64 | 0.43 | 0.57 | 0.80 | 0.20 |
| 43 | 0.20 | 0.80 | 0.80 | 0.20 | 0.55 | 0.45 | 0.44 | 0.56 |
| 44 | 0.22 | 0.78 | 0.69 | 0.31 | 0.68 | 0.32 | 0.51 | 0.49 |
| 45 | 0.21 | 0.79 | 0.80 | 0.20 | 0.80 | 0.20 | 0.80 | 0.20 |
| 47 | 0.20 | 0.80 | 0.80 | 0.20 | 0.80 | 0.20 | 0.40 | 0.60 |

**Table 11.15** $W_{R_k j}$ Values resulting from model (11.11)

| | DMU | K = 1 | K = 2 | K = 3 | K = 4 | K = 5 |
|---|---|---|---|---|---|---|
| $N_1$ | 1 | 0.44408 | 0.39907 | 0.15685 | | |
| | 3 | 0.16682 | 0.29690 | 0.53628 | | |
| | 4 | 0.54295 | 0.26620 | 0.19085 | | |
| | 6 | 0.22609 | 0.29794 | 0.47598 | | |
| | 9 | 0.22350 | 0.62558 | 0.15092 | | |
| | 10 | 0.26127 | 0.58762 | 0.15111 | | |
| | 11 | 0.37901 | 0.34571 | 0.27528 | | |
| | 19 | 0.37866 | 0.28895 | 0.33240 | | |
| | 32 | 0.28143 | 0.29423 | 0.42434 | | |
| | 35 | 0.18438 | 0.54183 | 0.27379 | | |
| | 39 | 0.24116 | 0.19268 | 0.56616 | | |
| | 46 | 0.30087 | 0.42795 | 0.27118 | | |
| $N_2$ | 2 | | 0.38390 | 0.10349 | 0.19334 | 0.31928 |
| | 18 | | 0.18928 | 0.31701 | 0.33098 | 0.16272 |
| | 27 | | 0.11320 | 0.38099 | 0.10099 | 0.40482 |
| | 31 | | 0.59282 | 0.25192 | 0.07614 | 0.07911 |
| | 33 | | 0.20369 | 0.55359 | 0.08866 | 0.15406 |
| | 34 | | 0.17234 | 0.16286 | 0.14964 | 0.51516 |
| | 38 | | 0.33544 | 0.49652 | 0.08301 | 0.08502 |
| $N_3$ | 5 | | | 0.23014 | | 0.76986 |
| | 8 | | | 0.75205 | | 0.24795 |
| | 12 | | | 0.22075 | | 0.77925 |
| | 15 | | | 0.72507 | | 0.27493 |
| | 17 | | | 0.22210 | | 0.77790 |
| | 21 | | | 0.46861 | | 0.53139 |
| | 22 | | | 0.75620 | | 0.24380 |
| | 26 | | | 0.62227 | | 0.37773 |
| | 28 | | | 0.81509 | | 0.18491 |
| | 29 | | | 0.69303 | | 0.30697 |
| | 36 | | | 0.72576 | | 0.27424 |
| | 40 | | | 0.65048 | | 0.34952 |
| | 42 | | | 0.22915 | | 0.77085 |
| $N_4$ | 7 | 0.48752 | | 0.51248 | | |
| | 13 | 0.49139 | | 0.50861 | | |
| | 14 | 0.33427 | | 0.66573 | | |
| | 16 | 0.76776 | | 0.23224 | | |
| | 20 | 0.36724 | | 0.63276 | | |
| | 23 | 0.33658 | | 0.66342 | | |
| | 24 | 0.52081 | | 0.47919 | | |
| | 25 | 0.50691 | | 0.49309 | | |
| | 30 | 0.36657 | | 0.63343 | | |
| | 37 | 0.40675 | | 0.59325 | | |
| | 41 | 0.34990 | | 0.65010 | | |
| | 43 | 0.33357 | | 0.66643 | | |
| | 44 | 0.39367 | | 0.60633 | | |
| | 45 | 0.50200 | | 0.49800 | | |
| | 47 | 0.35032 | | 0.64968 | | |

**Table 11.16** Subunit scores and overall efficiency scores

| DMU | K=1 | K=2 | K=3 | K=4 | K=5 | Overall Score |
|---|---|---|---|---|---|---|
| 1 | 0.64383 | 0.48974 | 0.96625 | | | 0.63291 |
| 2 | | 0.49777 | 1.00000 | 0.72677 | 0.43337 | 0.57346 |
| 3 | 1.00000 | 0.83433 | 0.57926 | | | 0.72518 |
| 4 | 0.48579 | 0.58032 | 0.76521 | | | 0.56428 |
| 5 | | | 0.77758 | | 0.24349 | 0.36640 |
| 6 | 0.55676 | 0.71150 | 0.64337 | | | 0.64409 |
| 7 | 0.82027 | | 0.74042 | | | 0.77935 |
| 8 | | | 0.36178 | | 0.31186 | 0.34940 |
| 9 | 0.91587 | 0.37769 | 1.00000 | | | 0.59189 |
| 10 | 0.50165 | 0.37760 | 1.00000 | | | 0.50406 |
| 11 | 0.63563 | 0.66075 | 1.00000 | | | 0.74462 |
| 12 | | | 0.59878 | | 0.23637 | 0.31637 |
| 13 | 0.49416 | | 0.63887 | | | 0.56776 |
| 14 | 0.57289 | | 0.65031 | | | 0.62443 |
| 15 | | | 0.46475 | | 0.32369 | 0.42597 |
| 16 | 0.55495 | | 0.64928 | | | 0.57686 |
| 17 | | | 0.68029 | | 0.32672 | 0.40525 |
| 18 | | 0.83048 | 0.68179 | 0.70051 | 1.00000 | 0.76791 |
| 19 | 0.71655 | 0.67867 | 0.79608 | | | 0.73204 |
| 20 | 0.95713 | | 0.60450 | | | 0.73400 |
| 21 | | | 0.78876 | | 0.29359 | 0.52563 |
| 22 | | | 0.42537 | | 0.24189 | 0.38064 |
| 23 | 0.53971 | | 0.49664 | | | 0.51114 |
| 24 | 0.43947 | | 0.43417 | | | 0.43693 |
| 25 | 0.50334 | | 0.41704 | | | 0.46079 |
| 26 | | | 0.53215 | | 0.30466 | 0.44622 |
| 27 | | 1.00000 | 0.96923 | 0.48817 | 0.38550 | 0.68783 |
| 28 | | | 0.53566 | | 0.51068 | 0.53104 |
| 29 | | | 0.35708 | | 0.22141 | 0.31544 |
| 30 | 0.66471 | | 0.45856 | | | 0.53413 |
| 31 | | 0.42683 | 1.00000 | 1.00000 | 1.00000 | 0.66021 |
| 32 | 0.35574 | 0.65883 | 0.55675 | | | 0.53021 |
| 33 | | 0.81774 | 0.65852 | 0.71355 | 0.54007 | 0.67758 |
| 34 | | 0.59935 | 0.94789 | 0.43319 | 0.30634 | 0.48030 |

**Table 11.16** (continued)

| DMU | K=1 | K=2 | K=3 | K=4 | K=5 | Overall Score |
|---|---|---|---|---|---|---|
| 35 | 0.39096 | 0.55704 | 0.97542 | | | 0.64097 |
| 36 | | | 0.55920 | | 0.32494 | 0.49495 |
| 37 | 0.48443 | | 0.51519 | | | 0.50268 |
| 38 | | 0.72317 | 0.84442 | 0.92137 | 0.78049 | 0.80470 |
| 39 | 0.48497 | 0.89472 | 0.59068 | | | 0.62377 |
| 40 | | | 0.50195 | | 0.64758 | 0.55285 |
| 41 | 0.83899 | | 0.52029 | | | 0.63181 |
| 42 | | | 0.55009 | | 0.30872 | 0.36403 |
| 43 | 0.48226 | | 0.46341 | | | 0.46970 |
| 44 | 0.42130 | | 0.34355 | | | 0.37416 |
| 45 | 0.52384 | | 0.44516 | | | 0.48466 |
| 46 | 0.62222 | 0.42030 | 0.73324 | | | 0.56591 |
| 47 | 0.24297 | | 0.61788 | | | 0.48654 |

**Table 11.17** Comparison of the proposed model with the conventional DEA model

| DMU | Conventional DEA | Proposed model | AR-conventional DEA | AR-proposed Model |
|---|---|---|---|---|
| 1 | 1.00000 | 0.96259 | 0.85841 | 0.63291 |
| 2 | 1.00000 | 0.87506 | 0.99184 | 0.57346 |
| 3 | 1.00000 | 0.52688 | 1.00000 | 0.72518 |
| 4 | 1.00000 | 0.85181 | 0.83462 | 0.56428 |
| 5 | 0.94100 | 0.32943 | 0.92577 | 0.36640 |
| 6 | 0.93144 | 0.89230 | 0.77672 | 0.64409 |
| 7 | 1.00000 | 0.97524 | 1.00000 | 0.77935 |
| 8 | 0.74458 | 0.45133 | 0.68388 | 0.34940 |
| 9 | 1.00000 | 0.83791 | 1.00000 | 0.59189 |
| 10 | 1.00000 | 0.85270 | 0.74586 | 0.50406 |
| 11 | 1.00000 | 0.94815 | 1.00000 | 0.74462 |
| 12 | 1.00000 | 0.58070 | 0.78612 | 0.31637 |
| 13 | 0.96264 | 0.49230 | 0.82657 | 0.56776 |
| 14 | 1.00000 | 0.72042 | 0.97245 | 0.62443 |
| 15 | 0.99860 | 0.52936 | 0.92594 | 0.42597 |
| 16 | 1.00000 | 0.59822 | 1.00000 | 0.57686 |
| 17 | 1.00000 | 0.46034 | 1.00000 | 0.40525 |

**Table 11.17** (continued)

| DMU | Conventional DEA | Proposed model | AR-conventional DEA | AR-proposed Model |
|---|---|---|---|---|
| 18 | 1.00000 | 1.00000 | 1.00000 | 0.76791 |
| 19 | 1.00000 | 1.00000 | 0.98835 | 0.73204 |
| 20 | 1.00000 | 0.64903 | 1.00000 | 0.73400 |
| 21 | 1.00000 | 0.49218 | 1.00000 | 0.52563 |
| 22 | 0.71282 | 0.41578 | 0.69469 | 0.38064 |
| 23 | 1.00000 | 0.94055 | 0.86776 | 0.51114 |
| 24 | 0.81586 | 0.32452 | 0.66143 | 0.43693 |
| 25 | 1.00000 | 0.49337 | 0.66097 | 0.46079 |
| 26 | 1.00000 | 0.52563 | 0.95684 | 0.44622 |
| 27 | 1.00000 | 0.64045 | 1.00000 | 0.68783 |
| 28 | 1.00000 | 0.60527 | 1.00000 | 0.53104 |
| 29 | 0.80628 | 0.40414 | 0.51204 | 0.31544 |
| 30 | 0.96842 | 0.75725 | 0.75228 | 0.53413 |
| 31 | 1.00000 | 0.77224 | 1.00000 | 0.66021 |
| 32 | 0.91397 | 0.78357 | 0.75751 | 0.53021 |
| 33 | 1.00000 | 0.64628 | 1.00000 | 0.67758 |
| 34 | 1.00000 | 0.60739 | 0.82704 | 0.48030 |
| 35 | 1.00000 | 0.58479 | 0.91254 | 0.64097 |
| 36 | 1.00000 | 0.63989 | 0.91299 | 0.49495 |
| 37 | 0.89794 | 0.57327 | 0.85628 | 0.50268 |
| 38 | 1.00000 | 0.92342 | 1.00000 | 0.80470 |
| 39 | 1.00000 | 1.00000 | 0.81284 | 0.62377 |
| 40 | 1.00000 | 0.85580 | 1.00000 | 0.55285 |
| 41 | 1.00000 | 0.88574 | 1.00000 | 0.63181 |
| 42 | 1.00000 | 0.40143 | 1.00000 | 0.36403 |
| 43 | 1.00000 | 0.47160 | 0.76886 | 0.46970 |
| 44 | 0.62889 | 0.44469 | 0.60419 | 0.37416 |
| 45 | 0.79867 | 0.40053 | 0.74383 | 0.48466 |
| 46 | 1.00000 | 0.74946 | 0.81605 | 0.56591 |
| 47 | 0.90160 | 0.39063 | 0.83234 | 0.48654 |

## References

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Euro J Oper Res 2(6):429–444

Cook WD, Hababou M (2001) Sales performance measurement in bank branches. Omega 29:299–307

Cook WD, Zhu J (2011) Multiple variable proportionality in DEA. Oper Res 59(4):1024–1032

Cook WD, Hababou M, Tuenter F (2000) Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Product Anal 14(3):209–224

Cook WD, Harrison J, Rouse P, Zhu J (2012) Relative efficiency measurement: the problem of a missing output in a subset of decision making units. Euro J Oper Res 220(1):79–84

Färe R, Grosskopf S (1996) Intertemporal production frontiers with dynamic DEA. Kluwer Academic, Boston

Pulley LB, Braunstein YM (1992) A composite cost function for multiproduct firms with an application to economies of scope in banking. Rev Econ Stat 74:221–230

Thompson RG, Langemeir LN, Lee C, Lee L, Thrall RM (1990) The role of multiplier bounds in efficiency analysis with application to Kansas farming. J Econom 46:93–108

Thompson RG, Dharmapala PS, Thrall RM (1993) Importance for DEA of zeros in data, multipliers, and solutions. J Prod Anal 4:379–390

# Chapter 12
# Efficiency Measurement in Data Envelopment Analysis with Fuzzy Data

**Chiang Kao**

**Abstract** Conventional data envelopment analysis (DEA) requires the data to have crisp values, which can be measured precisely. However, there are cases where data is missing and has to be estimated, or the situation has not occurred yet and the data has to be predicted. There are also cases where the factors are qualitative, and thus the data cannot be measured precisely. In these cases, fuzzy numbers can be used to represent the imprecise values, and this paper discusses the corresponding measurement of efficiency. Based on the extension principle, two approaches are proposed; one views the membership function of the fuzzy data vertically, and the results are represented by membership grades. The other views it horizontally, and the results are represented by α-cuts. The former approach is easier to understand, yet is applicable only to very simple problems. The latter, in contrast, can be applied to all problems, and is easier to implement. An example explains the development and implementation of these two approaches.

**Keywords** Data envelopment analysis · Fuzzy data · Two-level programming · Extension principle

## 12.1 Introduction

Data envelopment analysis (DEA), developed by Charnes et al. (1978), is a technique for measuring the relative efficiency of a set of decision making units (DMUs) that use multiple inputs to produce multiple outputs. Due to its solid theoretical foundation and persuasive measurement approach, it has been widely applied to evaluate efficiency for real world cases (see, for example, the survey of Cook and Seiford 2009).

With DEA; the efficiency measures are very sensitive to the data. If there is an outlier, then the efficiency measures of most DMUs will change drastically.

C. Kao (✉)
Department of Industrial and Information Management,
National Cheng Kung University Tainan, Taiwan
e-mail: ckao@mail.ncku.edu.tw

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_12

Therefore, a key to the success of the DEA approach is the accurate measurement of the data. However, in addition to outliers, there are many cases where the data cannot be measured precisely. For example, in measuring the volume of a tree, different persons will get different measures due to the irregular shape of the tree. In other words, there are measurement errors. Even if the data can be collected correctly and precisely, there are still factors which limit precise measurements. For example, the data is missing and has to be estimated (Kao and Liu 2000), or a situation has not occurred yet, and the data has to be predicted (Kao and Liu 2004). In addition, if the input/output factors are qualitative, which are described by linguistic terms, such as strongly satisfactory, satisfactory, and unsatisfactory, then precise measurement is almost impossible (Kao and Lin 2011). This property of impreciseness makes conventional DEA models intractable.

Bellman and Zadeh (1970) introduced the notion of fuzziness to deal quantitatively with imprecision in the decision process, and several DEA models have been developed based on this to handle fuzzy data (Dia 2004; Guo 2009; Jahanshahloo et al. 2004; Kao and Liu 2000a; Leon et al. 2003; Lertworasirkul et al. 2003; Wen and Li 2009). When data is imprecise, it is expected that the measured efficiency will also be imprecise. Unfortunately, most of the above-mentioned studies only provide crisp measures. Although a lot of effort has been devoted to studying DEA under fuzzy environments, it remains less well developed than its deterministic counterpart.

In this paper we will develop two approaches to measure efficiency when the data is fuzzy, based on the extension principle of Zadeh (1978). The measured efficiency is a fuzzy number, and is thus more informative than crisp values for making decisions. In the following, we will first use an example to introduce the concept of efficiency measurement graphically when the data is fuzzy. Then, in Sects 12.3 and 12.4, we will develop two approaches, using an example extended from the graphical one discussed in Sect. 12.2. Finally, some discussions are made and conclusions are drawn in Sect. 12.5.

## 12.2 The Problem

Let $X_{ij}$, $i = 1,\ldots, m$, and $Y_{rj}$, $r = 1,\ldots, s$, denote the $i$th input and $r$th output, respectively, of DMU $j$, $j = 1,\ldots, n$. The output-oriented model for measuring the efficiency of DMU $k$, under the assumption of variable returns to scale, can be formulated as (Banker et al. 1984):

$$\begin{aligned} E_k = \max.\quad & \sum_{r=1}^{s} u_r Y_{rk} \\ \text{s.t.}\quad & v_0 + \sum_{i=1}^{m} v_i X_{ik} = 1 \\ & \sum_{r=1}^{s} u_r Y_{rj} - \left( v_0 + \sum_{i=1}^{m} v_i X_{ij} \right) \le 0, \quad j = 1, \ldots, n \\ & u_r, v_i \ge \varepsilon, \quad r = 1, \ldots, s, \quad i = 1, \ldots, m \\ & v_0 \quad \text{unrestricted in sign,} \end{aligned} \tag{12.1}$$

where ε is a small non-Archimedean number imposed to avoid ignorance of any factor in calculating efficiency (Charnes and Cooper 1984).

Suppose the inputs $X_{ij}$ and outputs $Y_{rj}$ are approximately known, and can be represented by fuzzy numbers $\tilde{X}_{ij}$ and $\tilde{Y}_{rj}$, characterized by membership functions $\mu_{\tilde{X}_{ij}}$ and $\mu_{\tilde{Y}_{rj}}$, respectively. The membership function has a range of [0, 1], where larger values indicate a higher possibility of occurrence. A common fuzzy number is trapezoidal, denoted as $\tilde{X} = (a, b, c, d)$, whose membership function is:

$$\mu_{\tilde{X}}(x) = \begin{cases} (x-a)/(b-a), & a \le x \le b \\ 1, & b \le x \le c \\ (d-x)/(d-c), & c \le x \le d \end{cases}. \tag{12.2}$$

For very simple problems, the membership function of the fuzzy efficiency of a DMU, when there are fuzzy observations, can be derived analytically.

Consider four DMUs, labeled as $A$, $B$, $C$, and $D$ in Fig. 12.1, where 10, 20, 30, and 50 units of input $X$ are applied to produce 5, (6, 7, 8, 9), 9, and 15 units of output $Y$, respectively. Here only one observation, $\tilde{Y}_B = (6, 7, 8, 9)$, is fuzzy, whose membership function is:

$$\mu_{\tilde{Y}_B}(y) = \begin{cases} y-6, & 6 \le y \le 7 \\ 1, & 7 \le y \le 8 \\ 9-y, & 8 \le y \le 9 \end{cases}. \tag{12.3}$$

For $y$ in the range of [6, 7.5], the production frontier constructed from these four DMUs is the line segment connecting $A$ and $D$. As the value of $y$ increases from 7.5 to the upper bound 9, the frontier becomes a kinked line segment $AyD$. No matter what value $y$ is, DMUs $A$ and $D$ always lie on the frontier, and therefore are efficient, with $E_A = E_D = 1$. The efficiencies of $B$ and $C$, on the other hand, change with the value of $y$, and are fuzzy numbers.

The efficiency of DMU $B$ is $y/7.5$ for $y$ less than or equal to 7.5. Since $y$ has a membership grade of $(y-6)$ in the range of [6, 7], $(y-6)$ will also be the membership grade for the corresponding efficiency score of $e = y/7.5$, and it has a range of [6/7.5, 7/7.5]. Expressing the membership grade of $(y-6)$ by $e = y/7.5$, or $y = 7.5e$, one has $(7.5e-6)$. Similarly, $y$ has a membership grade of 1 in the range of [7, 7.5], the efficiency score of $y/7.5$ (with a range of [7/7.5, 1]) thus also has a membership grade of 1. For $y$ in the range of [7.5, 9], it lies on the frontier and becomes efficient. Since the membership grade for $y$ in this range has different values, and the largest is 1, the membership grade for an efficiency score of 1 is 1. Combining these results together, the membership function of the fuzzy efficiency of DMU $B$ is:

$$\mu_{\tilde{E}_B}(e) = \begin{cases} 7.5e-6, & 6/7.5 \le e \le 7/7.5 \\ 1, & 7/7.5 \le e \le 1 \end{cases} \tag{12.4}$$

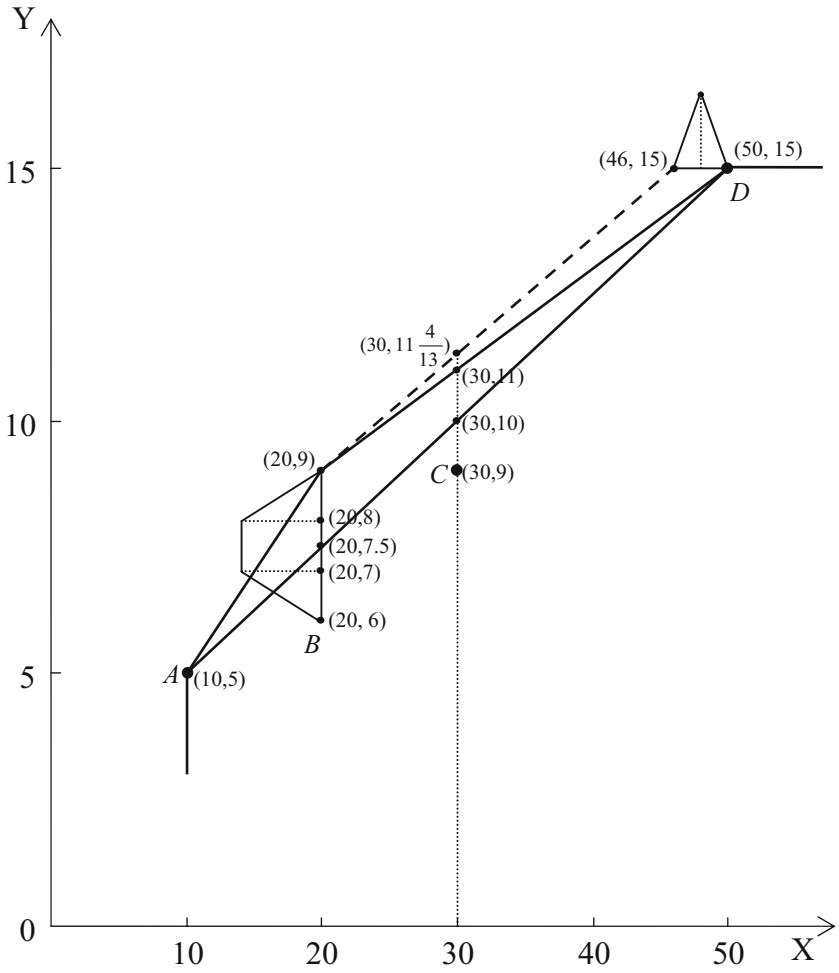



**Fig. 12.1** Production frontier and efficiency measurement of the example

Figure 12.2 shows this membership function, labeled as $\tilde{E}_B$.

For DMU $C$, its efficiency is $e = 9/10$ for $y$ in the range of [6, 7.5], since the frontier in this case is the line segment $AD$. The largest membership grade for $y$ in this range is 1; therefore, $e = 9/10$ has a membership grade of 1. For $y$ in the range of [7.5, 9], the target point of DMU $C$ on the frontier is $y + (30 - 20)(15 - y)/(50 - 20)$, or $(15 + 2y)/3$, which results in an efficiency score of $e = 27/(15 + 2y)$. According to Eq. (12.3), the membership grades for $y$ in the ranges of [7.5, 8] and [8, 9] are 1 and $9 - y$, respectively. Therefore, the corresponding $e = 27/(15 + 2y)$ has the same membership grades in the ranges of [27/31, 9/10] and [9/11, 27/31], respectively. Taking the inverse function of $e = 27/(15 + 2y)$, one obtains $y = (27 - 15e)/2e$. The membership grade of $9 - y$ can thus be expressed as $9 - (27 - 15e)/2e$, or $(33e - 27)/2e$.

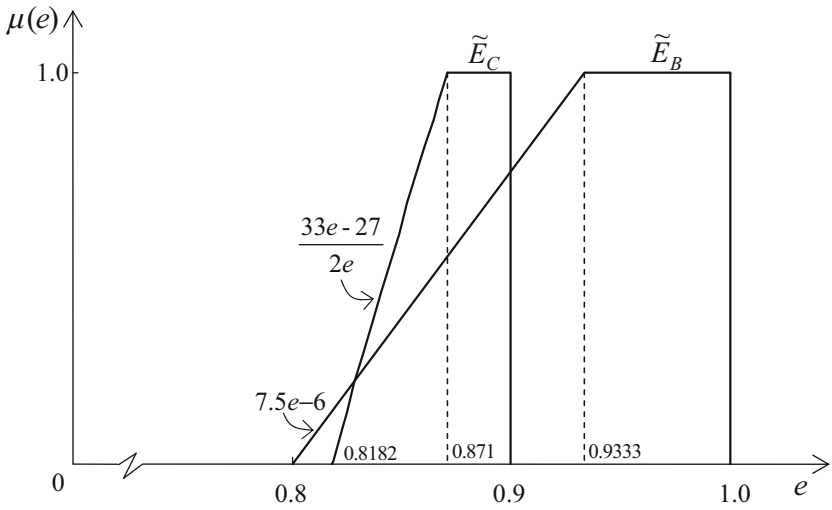



**Fig. 12.2** Membership functions of $\tilde{E}_B$ and $\tilde{E}_C$

Altogether, the membership function of $\tilde{E}_C$ is:

$$\mu_{\tilde{E}_C}(e) = \begin{cases} (33e - 27)/2e, & 9/11 \le e \le 27/31 \\ 1, & 27/31 \le e \le 9/10 \end{cases} \tag{12.5}$$

Figure 12.2 shows this membership function, labeled as $\tilde{E}_C$.

This example shows that when observations are fuzzy numbers, the efficiencies are also fuzzy numbers, and the membership function of the latter can be derived analytically from the former. Notably, although only one DMU has a fuzzy observation (DMU $B$ in this case), other DMUs (DMU $C$ in this case) can also have fuzzy efficiency. Obviously, the analytical derivation is possible only for very simple cases. When more observations are fuzzy numbers, or when more DMUs are involved, one must rely on numerical approaches.

For simplicity of notation, suppose all observations in Model (12.1) are fuzzy numbers. Denote the fuzzy efficiency of DMU $k$ as $\tilde{E}_k$. Based on Zadeh's extension principle (Yager 1986; Zadeh 1978; Zimmermann 1996), the membership function for $\tilde{E}_k$ can be obtained via the following equation:

$$\mu_{\tilde{E}_k}(e) = \sup_{x,y} \min\{\mu_{\tilde{X}_{ij}}(x_{ij}),\ \mu_{\tilde{Y}_{rj}}(y_{rj}), \forall i, j, r | e = E_k(\boldsymbol{x}, \boldsymbol{y})\} \tag{12.6}$$

where $\mu_{\tilde{X}_{ij}}$ and $\mu_{\tilde{Y}_{rj}}$ are the membership functions of $\tilde{X}_{ij}$ and $\tilde{Y}_{rj}$, respectively, and $E_k(\boldsymbol{x}, \boldsymbol{y})$ is defined in Model (12.1). The right-hand side of Eq. (12.6) is a two-level programming problem. At the second level (the inner program), one seeks the minimum membership grade for each set of $x_{ij}$ and $y_{rj}$ values which generates an efficiency score of $e$. At the first level (the outer program), one finds the largest

membership grade from those obtained from the second level program for all sets of $x_{ij}$ and $y_{rj}$ values. To find the membership function $\mu_{\tilde{E}_k}$, it suffices to solve the two-level program. In the following two sections, we will develop two approaches to solve the two-level program.

## 12.3 The Membership Grade Approach

Equation (12.6) shows the relationship between the membership function of $\tilde{E}_k$ and those of $\tilde{X}_{ij}$ and $\tilde{Y}_{rj}$. While its mathematical meaning is clear, it is not solvable in the two-level form, and must be transformed into the conventional one-level program for a solution.

Let $h = \min\{\mu_{\tilde{X}_{ij}}(x_{ij}), \mu_{\tilde{Y}_{rj}}(y_{rj}), \forall i, j, r | e = E_k(\boldsymbol{x}, \boldsymbol{y})\}$, Eq. (12.6) can then be expressed as $\mu_{\tilde{E}_k}(e) = \max_{\mathbf{x},\mathbf{y}} h$. Since $h$ is the minimum of those items inside the brace of Eq. (12.6), it must satisfy the following conditions:

$$\begin{aligned} h &\le \mu_{\tilde{X}_{ij}}(x_{ij}), \quad i = 1, \ldots, m, \ j = 1, \ldots, n \\ h &\le \mu_{\tilde{Y}_{rj}}(y_{rj}), \quad r = 1, \ldots, s, \ j = 1, \ldots, n \end{aligned} \tag{12.7}$$

and the $x_{ij}$ and $y_{rj}$ values used to calculate $\mu_{\tilde{X}_{ij}}(x_{ij})$ and $\mu_{\tilde{Y}_{rj}}(y_{rj})$ in (12.7) must be able to generate an efficiency score of $e$ via Model (12.1). Altogether, Eq. (12.6) can be converted to the following mathematical program:

$$\begin{aligned} &\mu_{\tilde{E}_k}(e) = \max h \\ &\text{s.t.} \quad h \le \mu_{\tilde{X}_{ij}}(x_{ij}), \quad i = 1, \ldots, m, j = 1, \ldots, n \\ &\qquad h \le \mu_{\tilde{Y}_{rj}}(y_{rj}), \quad r = 1, \ldots, s, j = 1, \ldots, n \\ &\qquad e = \{\max. \sum_{r=1}^{s} u_r Y_{rk} \\ &\qquad \text{s.t. } v_0 + \sum_{i=1}^{m} v_i X_{ik} = 1. \\ &\qquad \sum_{r=1}^{s} u_r Y_{rj} - (v_0 + \sum_{i=1}^{m} v_i X_{ij}) \le 0, \quad j = 1, \ldots, n \\ &\qquad u_r, v_i \ge \varepsilon, \quad r = 1, \ldots, s, i = 1, \ldots, m \\ &\qquad v_0 \quad \text{unrestricted in sign}\} \end{aligned} \tag{12.8}$$

This is a special two-level program, where the second-level program is a constraint that must be transformed into a conventional one to be solvable, as shown in the following example.

Consider the example discussed in the preceding section. In addition to $\tilde{Y}_B$, suppose the input of DMU $D$ is also a fuzzy number. Let it be a triangular fuzzy number of $\tilde{X}_D = (46, 48, 50)$, whose membership function is:

$$\mu_{\tilde{X}_D}(x) = \begin{cases} (x - 46)/2, & 46 \le x \le 48 \\ (50 - x)/2, & 48 \le x \le 50 \end{cases}. \tag{12.9}$$

Referring to Fig. 12.1, it is interesting to note that no matter what the value of $x$ is, DMU $D$ always lies on the frontier, and thus has a crisp efficiency of 1. In other words, a DMU with fuzzy observations can have crisp efficiency score. DMU $A$ also has a crisp efficiency of 1, and only $B$ and $C$ have fuzzy efficiencies. We will use DMU $C$ to explain the generation of the membership function of its fuzzy efficiency.

By connecting DMU $A$ with a value $x$ in the domain of $\tilde{X}_D$, the point intersecting with the domain of $\tilde{Y}_B$ is $5+10(15-5)/(x-10)$, or $(5x+50)/(x-10)$. If $y$, in the range of [6, 9], is greater than that value, then the frontier is the kinked line segment $Ayx$; otherwise, it is the line segment $Ax$. In the former case, the target point of DMU $C$ on the frontier is $y + 10(15 - y)/(x-20)$, or $(xy-30y+150)/(x-20)$, which results in an efficiency score of $9/[(xy-30y+150)/(x-20)]$, or $(9x-180)/(xy-30y+150)$ for DMU $C$. Its smallest and largest values are 39/49, occurring at $x = 46$ and $y = 9$, and 0.9, occurring at $x = 50$ and $y \le 7.5$, respectively. The constraint that the calculated efficiency must be equal to $e$ in Model (12.8) can thus be expressed as $e = (9x - 180)/(xy - 30y + 150)$. In the latter case, the target point on the frontier is $5 + 20(15 - 5)/(x - 10)$, or $(5x + 150)/(x - 10)$, which results in an efficiency score of $(9x - 90)/(5x + 150)$ for DMU $C$. In this case, the constraint becomes $e = (9x - 90)/(5x + 150)$.

Regarding the constraints of $h \le \mu_{\tilde{X}_{ij}}(x_{ij})$ and $h \le \mu_{\tilde{Y}_{rj}}(y_{rj})$ in Model (12.8), consider the trapezoidal fuzzy number defined in (12.2). One of the following three situations must hold:

a. if $a \le x \le b$, then $h \le \mu_{\tilde{X}}(x) = (x - a)/(b - a)$
b. if $b \le x \le c$, then $h \le \mu_{\tilde{X}}(x) = 1$
c. if $c \le x \le d$, then $h \le \mu_{\tilde{X}}(x) = (d - x)/(d - c)$.

To describe this set of either-or constraints, three binary variables $\delta_1$, $\delta_2$, and $\delta_3$ are introduced. The formulation is:

a. $a \le x + M\,\delta_1, x \le b + M\,\delta_1, h \le (x - a)/(b - a) + M\,\delta_1$
b. $b \le x + M\,\delta_2, x \le c + M\,\delta_2, h \le 1 + M\,\delta_2$
c. $c \le x + M\,\delta_3, x \le d + M\,\delta_3, h \le (d - x)/(d - c) + M\,\delta_3$,

where $M$ is a very large number. When $\delta_i$ is equal to 1, the associated constraints are redundant, and they are active when $\delta_i$ is equal to 0. Therefore, a constraint of $\delta_1 + \delta_2 + \delta_3 = 2$ ensures that only one of the three situations holds. Triangular membership functions can be handled similarly.

**Table 12.1** Membership grades of $\mu_{\tilde{E}_C}(e)$ at 21 values of $e$

| E | $\mu_{\tilde{E}_C}$(e) | e | $\mu_{\tilde{E}_C}$(e) | e | $\mu_{\tilde{E}_C}$(e) |
|---|---|---|---|---|---|
| 0.800 | 0.0727 | 0.835 | 0.6592 | 0.870 | 1.0000 |
| 0.805 | 0.1604 | 0.840 | 0.7380 | 0.875 | 1.0000 |
| 0.810 | 0.2467 | 0.845 | 0.8158 | 0.880 | 0.8696 |
| 0.815 | 0.3317 | 0.850 | 0.8924 | 0.885 | 0.6557 |
| 0.820 | 0.4155 | 0.855 | 0.9679 | 0.890 | 0.4396 |
| 0.825 | 0.4979 | 0.860 | 1.0000 | 0.895 | 0.2210 |
| 0.830 | 0.5791 | 0.865 | 1.0000 | 0.900 | 0.0000 |

With these representations and transformations, Model (12.8) for calculating the membership grade of the fuzzy efficiency of DMU $C$ is formulated as:

$$
\begin{aligned}
&\mu_{\tilde{E}_C}(e) = \max\ h \\
&\text{s.t.}\quad 6 \le y + 1000\delta_1, \quad y \le 7 + 1000\delta_1, \quad h \le (y-6) + 1000\delta_1 \\
&\qquad 7 \le y + 1000\delta_2, \quad y \le 8 + 1000\delta_2, \quad h \le 1 + 1000\delta_2 \\
&\qquad 8 \le y + 1000\delta_3, \quad y \le 9 + 1000\delta_3, \quad h \le (9-y) + 1000\delta_3 \\
&\qquad 46 \le x + 1000\delta_4, \quad x \le 48 + 1000\delta_4, \quad h \le (x-46)/2 + 1000\delta_4 \\
&\qquad 48 \le x + 1000(1-\delta_4), \quad x \le 50 + 1000\delta_4, \quad h \le (50-x)/2 + 1000(1-\delta_4) \\
&\qquad e = (9x - 180)/(xy - 30y + 150) \\
&\qquad 6 \le y, \quad y \le 9, \quad 46 \le x, \quad x \le 50 \\
&\qquad \delta_1 + \delta_2 + \delta_3 = 2, \quad \delta_i \in \{0,1\}, i = 1,\ldots,4,
\end{aligned}
\tag{12.10}
$$

where the large number $M$ has been replaced with 1000. Note that the above model is for cases of $y \ge (5x+50)/(x-10)$. If $y \le (5x+50)/(x-10)$, then $e = (9x-180)/(xy-30y+150)$ in the above model must be replaced by $e = (9x-90)/(5x+150)$.

By enumerating various values of $e$, the membership function of $\mu_{\tilde{E}_C}$ can be obtained numerically. Table 12.1 shows the membership grade for 21 values of $e$, from 0.8 to 0.9.They are also depicted in Fig. 12.3, with solid circles, to show the shape of $\mu_{\tilde{E}_C}$. Note that for $e$ greater than 81/95, or 0.8526, the constraint for $e$ in Model (12.10) is replaced with $e = (9x-90)/(5x+150)$ to calculate the membership grade of $\mu_{\tilde{E}_C}$. If a finer graph is desired, then one simply enumerates more $e$ values to get more membership grades.

The key point of this approach is in expressing the efficiency score as a function of $x_{ij}$ and $y_{rj}$ in closed form. This is a very difficult task, even for the very simple problem in the example. Moreover, the resulting model is nonlinear, which is relatively difficult to solve. Therefore, it is not a practical method for solving real world problems.

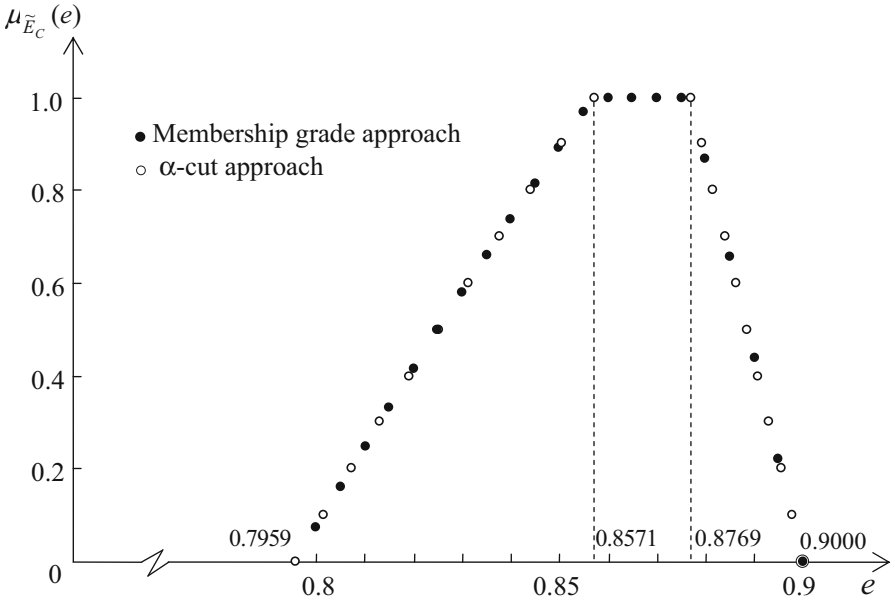



**Fig. 12.3** Membership function of $\tilde{E}_C$ constructed from two approaches

## 12.4 The α-cut Approach

The membership grade views the membership function vertically. The membership function can also be viewed horizontally from the α-cut, which is defined as $(X)_\alpha = \{x \mid \mu_{\tilde{X}}(x) \geq \alpha\}$. Denote $(X)_\alpha^L = \min.\{x \mid \mu_{\tilde{X}}(x) \geq \alpha\}$ and $(X)_\alpha^U = \max.\{x \mid \mu_{\tilde{X}}(x) \geq \alpha\}$, then $(X)_\alpha = \{x \mid (X)_\alpha^L \leq x \leq (X)_\alpha^U\}$. If the α-cuts at all α values are known, then the membership function can be constructed.

According to Eq. (12.6), $\mu_{\tilde{E}_k}(e)$ is the minimum of $\mu_{\tilde{X}_{ij}}(x_{ij})$ and $\mu_{\tilde{Y}_{rj}}(y_{rj})$, $\forall i, j, r$. To satisfy $\mu_{\tilde{E}_k}(e) = \alpha$, one needs $\mu_{\tilde{X}_{ij}}(x_{ij}) \geq \alpha$, $\mu_{\tilde{Y}_{rj}}(y_{rj}) \geq \alpha$, $\forall i, j, r$, and at least one of them is equal to α, and these $x_{ij}$ and $y_{rj}$ must generate an efficiency score of $e = E_k(\boldsymbol{x}, \boldsymbol{y})$. Since $\mu_{\tilde{X}_{ij}}(x_{ij}) \geq \alpha$ and $\mu_{\tilde{X}_{ij}}(x_{ij}) = \alpha$ have the same domain (this also applies to $\mu_{\tilde{Y}_{rj}}(y_{rj})$), one only needs to check the α-cuts of $\mu_{\tilde{X}_{ij}}(x_{ij})$ and $\mu_{\tilde{Y}_{rj}}(y_{rj})$. Denote $(X_{ij})_\alpha = [(X_{ij})_\alpha^L, (X_{ij})_\alpha^U]$, $(Y_{rj})_\alpha = [(Y_{rj})_\alpha^L, (Y_{rj})_\alpha^U]$, and $(E_k)_\alpha = [(E_k)_\alpha^L, (E_k)_\alpha^U]$ as the α-cuts of $\mu_{\tilde{X}_{ij}}(x_{ij})$, $\mu_{\tilde{Y}_{rj}}(y_{rj})$, and $\mu_{\tilde{E}_k}(e)$, respectively. To find the lower bound of the α-cut of $\mu_{\tilde{E}_k}(e)$, $(E_k)_\alpha^L$, it suffices to find the smallest efficiency score for DMU $k$ generated from the $x_{ij}$ and $y_{rj}$ values in their respective α-cuts. Equation (12.6) indicates that this efficiency score has a membership grade of α. By the same token, the upper bound $(E_k)_\alpha^U$ can be found by searching for the maximum efficiency score. In symbols, $(E_k)_\alpha^L$ and $(E_k)_\alpha^U$ can be obtained via

the following models:

$$(E_k)_\alpha^L = \min_{\substack{(X_{ij})_\alpha^L \le x_{ij} \le (X_{ij})_\alpha^U \\ (Y_{rj})_\alpha^L \le y_{rj} \le (Y_{rj})_\alpha^U \\ \forall i,j,r}} E_k(\boldsymbol{x}, \boldsymbol{y}) \tag{12.11a}$$

$$(E_k)_\alpha^U = \max_{\substack{(X_{ij})_\alpha^L \le x_{ij} \le (X_{ij})_\alpha^U \\ (Y_{rj})_\alpha^L \le y_{rj} \le (Y_{rj})_\alpha^U \\ \forall i,j,r}} E_k(\boldsymbol{x}, \boldsymbol{y}) \tag{12.11b}$$

or in full form:

$$(E_k)_\alpha^L = \min_{\substack{(X_{ij})_\alpha^L \le x_{ij} \le (X_{ij})_\alpha^U \\ (Y_{rj})_\alpha^L \le y_{rj} \le (Y_{rj})_\alpha^U \\ \forall i,j,r}} \begin{cases} E_k = \max. \sum_{r=1}^s u_r y_{rk} \\ \text{s.t.} \quad v_0 + \sum_{i=1}^m v_i x_{ik} = 1 \\ \sum_{r=1}^s u_r y_{rj} - (v_0 + \sum_{i=1}^m v_i x_{ij}) \le 0, j = 1, ..., n \\ u_r, v_i \ge 0, r = 1, ..., s, i = 1, ..., m \\ v_0 \text{ unrestricted in sign} \end{cases} \tag{12.12a}$$

$$(E_k)_\alpha^U = \max_{\substack{(X_{ij})_\alpha^L \le x_{ij} \le (X_{ij})_\alpha^U \\ (Y_{rj})_\alpha^L \le y_{rj} \le (Y_{rj})_\alpha^U \\ \forall i,j,r}} \begin{cases} E_k = \max. \sum_{r=1}^s u_r y_{rk} \\ \text{s.t.} \quad v_0 + \sum_{i=1}^m v_i x_{ik} = 1 \\ \sum_{r=1}^s u_r y_{rj} - (v_0 + \sum_{i=1}^m v_i x_{ij}) \le 0, j = 1, ..., n \\ u_r, v_i \ge 0, r = 1, ..., s, i = 1, ..., m \\ v_0 \text{ unrestricted in sign} \end{cases} \tag{12.12b}$$

These two models are two-level programs, and must be transformed into the conventional one-level ones in order to be solved.

The meaning of Model (12.12a) is that for each set of $x_{ij}$ and $y_{rj}$ values given at the first level, the second-level program calculates the corresponding efficiency, and the first-level program determines the set of $x_{ij}$ and $y_{rj}$ values which produce the smallest efficiency score. Based on the concept of relative comparison, the smallest efficiency occurs at the least favorable condition of DMU $k$, which is it uses the largest amount of input $(X_{ik})_\alpha^U$ to produce the smallest amount of output $(Y_{rk})_\alpha^L$, while other DMUs use the smallest amount of input $(X_{ij})_\alpha^L$ to produce the largest amount of output $(Y_{rj})_\alpha^U$. Similarly, DMU $k$ needs the most favorable condition to obtain the largest efficiency score, which is it uses the smallest amount of input $(X_{ik})_\alpha^L$ to produce the largest amount of output $(Y_{rk})_\alpha^U$, while other DMUs use the largest amount of input $(X_{ij})_\alpha^U$ to produce the smallest amount of output $(Y_{rj})_\alpha^L$. Models (12.12a) and (12.12b) can thus be transformed into the following one-level

programs:

$$
\begin{aligned}
&(E_k)^L_\alpha = \max. \sum_{r=1}^{s} u_r (Y_{rk})^L_\alpha \\
&\text{s.t.} \quad v_0 + \sum_{i=1}^{m} v_i (X_{ik})^U_\alpha = 1 \\
&\sum_{r=1}^{s} u_r (Y_{rk})^L_\alpha - (v_0 + \sum_{i=1}^{m} v_i (X_{ik})^U_\alpha) \le 0 \qquad (12.13a) \\
&\sum_{r=1}^{s} u_r (Y_{rj})^U_\alpha - (v_0 + \sum_{i=1}^{m} v_i (X_{ij})^L_\alpha) \le 0, j = 1, \ldots, n, j \ne k \\
&u_r, v_i \ge \varepsilon, r = 1, \ldots, s, i = 1, \ldots, m \\
&v_0 \quad \text{unrestricted in sign,} \\
&(E_k)^U_\alpha = \max. \sum_{r=1}^{s} u_r (Y_{rk})^U_\alpha \\
&\text{s.t.} \quad v_0 + \sum_{i=1}^{m} v_i (X_{ik})^L_\alpha = 1 \\
&\sum_{r=1}^{s} u_r (Y_{rk})^U_\alpha - (v_0 + \sum_{i=1}^{m} v_i (X_{ik})^L_\alpha) \le 0 \qquad (12.13b) \\
&\sum_{r=1}^{s} u_r (Y_{rj})^L_\alpha - (v_0 + \sum_{i=1}^{m} v_i (X_{ij})^U_\alpha) \le 0, j = 1, \ldots, n, j \ne k \\
&u_r, v_i \ge \varepsilon, r = 1, \ldots, s, i = 1, \ldots, m \\
&v_0 \quad \text{unrestricted in sign.}
\end{aligned}
$$

By enumerating various values of α, the membership function $\mu_{\tilde{E}_k}(e)$ is constructed.

For the example discussed in the preceding section, the α-cuts for $\tilde{Y}_B = (6, 7, 8, 9)$ and $\tilde{X}_D = (46, 48, 50)$ are $[6 + \alpha, 9 - \alpha]$ and $[46 + 2\alpha, 50 - 2\alpha]$, respectively. According to Models (12.13a) and (12.13b), the programs for calculating the lower and upper bounds of the α-cut of $\mu_{\tilde{E}_k}$ are:

$$
\begin{aligned}
&(E_c)^L_\alpha = \max. 9u \\
&\text{s.t.} \quad v_0 + 30v_1 = 1 \\
&5u - v_0 - 10v_1 \le 0 \\
&(9 - \alpha)u - v_0 - 20v_1 \le 0 \\
&9u - v_0 - 30v_1 \le 0 \qquad (12.14a) \\
&15u - v_0 - (46 + 2\alpha)v_1 \le 0 \\
&u, v_1 \ge \varepsilon \\
&v_0 \quad \text{unrestricted in sign}
\end{aligned}
$$

**Table 12.2** The α-cuts of $\mu_{\tilde{E}_C}(e)$ at eleven α values

| α | $(E_C)_\alpha^L$ | $(E_C)_\alpha^U$ | α | $(E_C)_\alpha^L$ | $(E_C)_\alpha^U$ |
|---|---|---|---|---|---|
| 0.0 | 0.7959 | 0.9000 | 0.6 | 0.8313 | 0.8863 |
| 0.1 | 0.8016 | 0.8977 | 0.7 | 0.8376 | 0.8840 |
| 0.2 | 0.8073 | 0.8955 | 0.8 | 0.8440 | 0.8816 |
| 0.3 | 0.8131 | 0.8932 | 0.9 | 0.8505 | 0.8793 |
| 0.4 | 0.8191 | 0.8909 | 1.0 | 0.8571 | 0.8769 |
| 0.5 | 0.8251 | 0.8886 | | | |

$$
\begin{aligned}
&(E_c)_\alpha^U = \text{max. } 9u \\
&\text{s.t.} \quad v_0 + 30v_1 = 1 \\
&5u - v_0 - 10v_1 \le 0 \\
&(6+\alpha)u - v_0 - 20v_1 \le 0 \\
&9u - v_0 - 30v_1 \le 0 \\
&15u - v_0 - (50 - 2\alpha)v_1 \le 0 \\
&u, v_1 \ge \varepsilon \\
&v_0 \quad \text{unrestricted in sign.}
\end{aligned}
\tag{12.14b}
$$

Table 12.2 shows the α-cuts at α = 0, 0.1,. . ., 1.0. They are also depicted on Fig. 12.3 by hollow circles. Visually, the two types of circles in Fig. 12.3 show that the membership function $\mu_{\tilde{E}_C}$ constructed from the two approaches, membership grade and α-cut, is the same.

## 12.5 Discussion and Conclusion

In the real world, there are many cases where the data is imprecise, and can be expressed by fuzzy numbers. This paper shows that the measured efficiencies will be fuzzy numbers when the data is fuzzy by using a simple example. Based on the extension principle, this paper develops two approaches, membership grade and α-cut, to find the fuzzy efficiency of a DMU.

The membership grade approach calculates the membership grade for values in the domain of fuzzy efficiency. By aggregating various membership grades, the membership function is constructed. Supposing every fuzzy observation has a trapezoidal membership function, this approach needs ten constraints and three binary variables to express the membership function. For a problem of $t$ fuzzy observations, there will be $10t$ more constraints and $3t$ more binary variables in the associated model. Since the model is a nonlinear integer program, it is relatively difficult to solve. Most disappointingly, this approach is limited to very small problems where the efficiency

of the DMU being measured can be expressed as a function of the observations. A direction for future research is thus to develop a better transformation method, which is applicable to all problems.

In contrast to the membership grade approach, the α-cut approach is able to transform the two-level program into a linear one-level program for all problems. Moreover, the associated models of this approach are the conventional DEA ones. No extra constraints, nor extra variables, are needed in the transformation, which makes the solution process very easy. Therefore, it is a better approach to use.

This approach has another merit; that is, it is easy to rank the fuzzy numbers. When several fuzzy numbers are to be ranked, a very effective method is the one proposed by Chen and Klein (1997), which requires only three or four α-cuts of those fuzzy numbers. Since the results of the α-cut approach are α-cuts of the fuzzy efficiency, no further work is needed for ranking.

To make the model tractable, imprecise data in DEA studies is normally represented by the most likely values. The results, which should be imprecise, thus become precise, and this can make decision makers over-confident. With the fuzzy measures calculated from the models developed in this paper, the decision maker is better informed and can make better decisions.

Finally, production systems in many cases are composed of several processes interrelated with each other. That is, one faces a network system. If the operations of the component processes are not taken into account when measuring efficiency, then this produces misleading results, and many studies have examined this topic (see, for example, the review of Castelli 2010). However, there are few papers that discuss the case of fuzzy data, for example, the two-stage system of Kao and Liu (2011) and the parallel system of Kao and Lin (2012). Within the context of fuzzy data, network systems thus have ample room for exploration in future work.

## References

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 30:1078–1092

Bellman R, Zadeh LA (1970) Decision making in a fuzzy environment. Manag Sci 17B:141–164

Castelli L, Pesenti R, Ukovich W (2010) A classification of DEA models when the internal structure of the decision making units is considered. Ann Oper Res 173:207–235

Charnes A, Cooper WW (1984) The non-Archimedean CCR ratio for efficiency analysis: a rejoinder to Boyd and Färe, Eur J Oper Res 15:333–334

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Chen CB, Klein CM (1997) A simple approach to ranking a group of aggregated fuzzy utilities. IEEE Trans Syst Man Cybern Part B: Cybern 27:26–35

Cook WD, Seiford LM (2009) Data envelopment analysis (DEA)—thirty years on. Eur J Oper Res 192:1–17

Dia M (2004) A model of fuzzy data envelopment analysis. INFOR 42:267–279

Guo PJ (2009) Fuzzy data envelopment analysis and its application to location problems. Inf Sci 179:820–829

Jahanshahloo GR, Soleimani-damaneh M, Nasrabadi E (2004) Measure of efficiency in DEA with fuzzy input-output levels: a methodology for assessing, ranking and imposing of weights restrictions. Appl Maths Comput 156:175–187

Kao C, Lin PH (2011) Qualitative factors in data envelopment analysis: a fuzzy number approach. Eur J Oper Res 211:586–593

Kao C, Lin PL (2012) Efficiency of parallel production systems with fuzzy data. Fuzzy Sets Syst 198:83–98

Kao C, Liu ST (2000a) Fuzzy efficiency measures in data envelopment analysis. Fuzzy Sets Syst 113:427–437

Kao C, Liu ST (2000b) Data envelopment analysis with missing data: an application to university libraries in Taiwan. J Oper Res Soc 51:897–905

Kao C, Liu ST (2004) Predicting bank performance with financial forecasts: a case of Taiwan commercial banks. J Bank Financ 28:2353–2368

Kao C, Liu ST (2011) Efficiencies of two-stage systems with fuzzy data. Fuzzy Sets Syst 176:20–35

Leon T, Liern V, Ruiz JL, Sirvent I (2003) A fuzzy mathematical programming approach to the assessment of efficiency with DEA models. Fuzzy Sets Syst 139:407–419

Lertworasirkul S, Fang SC, Joines JA, Nuttle HLW (2003) Fuzzy data envelopment analysis (DEA): a possibility approach. Fuzzy Sets Syst 139:379–394

Wen ML, Li HS (2009) Fuzzy data envelopment analysis (DEA): model and ranking method. J Comput Appl Maths 223:872–878

Yager RR (1986) A characterization of the extension principle. Fuzzy Sets Syst 18:205–217

Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28

Zimmermann HZ (1996) Fuzzy set theory and its applications (3rd ed). Boston: Kluwer-Nijhoff

# Chapter 13
# Partial Input to Output Impacts in DEA: Production Considerations and Resource Sharing Among Business Sub-Units

**Raha Imanirad, Wade D. Cook and Joe Zhu**

**Abstract** Data envelopment analysis (DEA) is a methodology for evaluating the relative efficiencies of peer decision-making units (DMUs), in a multiple input/output setting. While it is generally assumed that all outputs are impacted by all inputs, there are many situations where this may not be the case. For example, in a food manufacturing setting, certain foods are exempt from nutrition labeling and as a result are not influenced by labeling resources. This chapter extends the conventional DEA methodology to allow for the measurement of technical efficiency in situations where only partial input-to-output impacts exist. The new methodology involves viewing the DMU as a business unit, consisting of a set of mutually exclusive subunits, each of which can be treated in the conventional DEA sense.

**Keywords** DEA · Partial impacts · Business sub units

## 13.1 Introduction

Data Envelopment Analysis (DEA), first introduced by Charnes et al. (1978), has gained widespread appeal as a tool for evaluating the relative efficiency of decision making units (DMUs) in various settings. The conventional DEA model is based on the implicit, if not explicit assumption that in a multiple input, multiple output environment, all inputs impact all outputs. Possibly a more accurate statement is that often the internal workings of the DMU are unknown to the analyst, or are at a level of complexity that prohibits a more precise portrayal of performance than that given by the basic model. However, in many situations the internal processes that define the DMU are more clearly understood, and the assumption of all inputs impacting all outputs should be abandoned. For example, in a hospital setting with multiple

W. D. Cook (✉) · R. Imanirad
Schulich School of Business, York University, 4700 Keele Street,
M3J 1P3 Toronto, ON, Canada
e-mail: wcook@schulich.yorku.ca

J. Zhu
School of Business, Worcester Polytechnic Institute, 100 Institute Road,
01609 Worcester, MA, USA

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations
Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_13

types of staff and activities, not all activities are influenced by all staff. Clinical staff members, for example, do not impact many operational outputs, hence, evaluating those outputs in terms of those particular staff inputs results in distorted efficiency measures. We elaborate on this distortion later in the chapter.

Such partial impacts are often a reflection of the fact that in some environments a DMU is actually a 'business unit' (e.g. a manufacturing facility), consisting of several subunits, where the activity in terms of the products made and resources consumed, differs from one subunit to another. Such a view brings to light a number of important issues. Perhaps the most pertinent issue for management at the outset is how to clearly define what should be considered as separate subunits. What outputs and inputs, and in what amounts define a subunit of the overall business? With such a definition in place, management can then move on to address the problem of resource sharing among the identified subunits; this activity will almost certainly be driven by the efficiencies of the various subunits in relation to the overall efficiency of the DMU/business. This need then gives rise to how one should go about measuring efficiency in such a setting.

In this chapter we examine the problem of evaluating efficiency in the presence of such partial input- to-output interactions within the context of a set of steel fabrication plants. Section 13.2 describes this problem setting. A general methodology is presented in Sect. 13.3 based on the idea that a DMU can be viewed as a business unit comprised of separate subunits, and that efficiency of the DMU can be defined as a weighted average of the efficiencies of the subunits. The methodology revolves around the idea that resources/inputs can be partitioned or separated, and thereby allocated to the defined subunits. We point out that some of the preliminary ideas behind the discussion in Sects. 13.2 and 13.3 were first presented in a working paper by Cook and Imanirad (2010). In Sect. 13.4, we further investigate the phenomenon of partial impacts of inputs on outputs in situations where assurance region (AR) constraints are imposed at the level of the subunit. We specifically address the issue where multiple, often inconsistent AR constraints are imposed on the same pair of variables. Section 13.5 deals with further considerations involving partial impacts of inputs on outputs. In particular, we examine the case wherein AR constraints may cross subunits, and as well we extend the methodology of Sect. 13.3 to accommodate non-separable variables. Section 13.6 demonstrates the application of the methodologies to data on a set of 20 fabrication plants. Conclusions follow in Sect. 13.7.

## 13.2 Efficiency Measurement in Steel Fabrication Plants

To demonstrate the problem of efficiency measurement in DEA settings where partial input to output interactions exist, a set of 20 steel fabrication plants is considered. The following four product groupings are manufactured across all plants:

**Table 13.1** Input-to-output connections

| | **Outputs** | | | |
|---|---|---|---|---|
| | Sheet | Flat | Pipes/ | Cylindrical |
| **Inputs** | Steel | Bar | Cylinders | Bearings |
| Labor | X | X | X | X |
| Shears | X | X | | |
| Presses | X | X | | |
| Lathes | | X | X | X |

1. Sheet steel products (ladders, guards, bumpers and conveyors);
2. Flat bar products used mainly in building construction (brackets, base plates, headers and posts);
3. Pipes and cylinders (storm drains, plumbing products, etc);
4. Cylindrical bearings (automotive and non-automotive).

Similarly plant resources are comprised of: (1) Plant labor, (2) Shearing machines, (3) Presses, and (4) Lathes.

For the purpose of this study, the four main product lines and plant resources constitute the outputs and inputs respectively.

Table 13.1 illustrates the interactions among these inputs and outputs. As demonstrated in the table, not all inputs impact all outputs. For example, the usage of the presses is not required for production of cylindrical bearings, while presses are needed in the production of sheet steel and flat bar products. As discussed above, it is reasonable to argue that the conventional DEA model is based on the assumption that in a multiple input, multiple output setting, all members of an input bundle influence the output bundle. Hence, in settings where partial input to output interactions exist, the application of the conventional model may not be appropriate, and may distort the profile of the efficiencies of the DMUs.

In order to address this problem, we may view each DMU as a business unit consisting of $K$ subunits, where each subunit $k$ is represented by its own input/output bundle $(I_k, R_k)$ so that each output in $R_k$ is impacted by every member of $I_k$. As a result, each subunit k can be treated as a DMU (or sub DMU) to which the conventional DEA models can be applied.

Considering the manufacturing setting described earlier, we number the inputs labor, shears, presses and lathes as 1, 2, 3, and 4 respectively. Similarly, outputs sheet steel, flat bar, pipes and bearings are numbered 1, 2, 3 and 4 respectively. We argue that the DMU may be viewed as consisting of three subunits or bundles $(I_1, R_1), (I_2, R_2), (I_3, R_3)$, where $I_1 = (1, 2, 3), I_2 = (1, 2, 3, 4), I_3 = (1, 4)$ and $R_1 = (1), R_2 = (2), R_3 = (3, 4)$.

It is important to point out that bundles, as described herein do not necessarily correspond to product groupings as might be designated by the company. The organization may well form product groups such as automotive, residential, industrial,

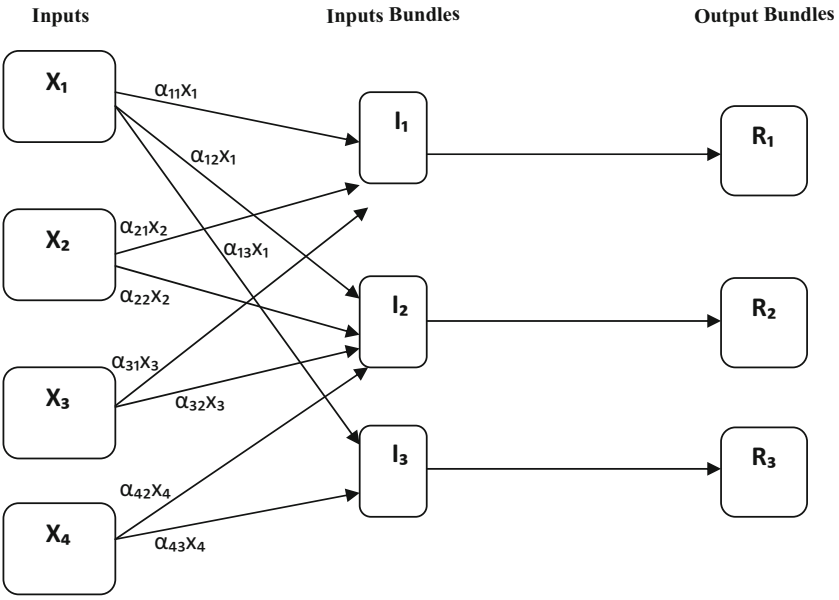



**Fig. 13.1** Resource splitting across business subunits

etc, but the members of these groups may not be impacted by the same inputs. Thus, the bundles, as we have defined them for modeling purposes may not be obvious groupings from the perspective of the organization, and may bear no resemblance to company-defined bundles.

An algorithm for generating these bundles in the general case is discussed in Sect. 13.3.1.

Further, assume that each input $x_i$ is separable, and can be apportioned across those subunits $k$ in which it holds membership, in the amounts $\alpha_{ik}x_i$, where $\Sigma_k\alpha_{ik} = 1$. For example, for each DMU $j$, input number 2 (shears) is a member of subunits $I_1$ and $I_2$, so a portion of the input needs to be allocated across $I_1$ and $I_2$ in some amounts $\alpha_{21}x_{2j}$ and $\alpha_{22}x_{2j}$ respectively. Figure 13.1 demonstrates the splitting of resources across those subunits containing those resources as members.

It is pertinent to note at this point that the conventional DEA model makes no specific provision for resource sharing; it is assumed that the entire input set impacts the entire output set. Thus, how a resource (input) is split across the outputs it impacts is not a consideration in the conventional methodology. In cases where partial impacts exist, however, as is true of the steel fabrication plants example, and where we view the DMU as consisting of a set of subunits, it becomes necessary to examine the sharing of inputs across the bundles that contain them as members. Thus, it becomes natural to decide on how resources will be shared.

In the sections to follow we present a methodology for deriving appropriate proportions $\alpha$.

## 13.3 Modeling Efficiency in the Presence of Partial Input to Output Interactions

In the conventional DEA setting a set of $n$ DMUs is to be evaluated in terms of a set of $I$ inputs $x_{ij}$ and $R$ outputs $y_{rj}$. One of the models used to accomplish this is a radial projection model such as the constant returns to scale (CRS) model (13.1).

$$\begin{aligned} & e = \max \sum_r u_r y_{ro} \Big/ \sum_i v_i x_{io} \\ & \textit{subject to} \\ & \sum_r u_r y_{rj} \Big/ \sum_i v_i x_{ij} \leq 1 \quad j = 1, \ldots\ldots\ldots, n \\ & u_r, v_i \geq 0, \quad \forall r, i \end{aligned} \tag{13.1}$$

In this formulation, as originally proposed by Charnes et al. (1978), $u_r, v_i$ are the decision variables and are intended to denote the prices, weights or multipliers to be assigned to outputs $y_{rj}$ and inputs $x_{ij}$ respectively.

This model was designed for efficiency settings where the entire bundle of $R$ outputs is influenced by the entire bundle of $I$ inputs. In situations such as that outlined in the previous section, however, where only partial input to output impacts exist, model (13.1) may give a distorted view of the efficiencies of the DMUs. To illustrate, consider the simple example of two DMUs with 2 inputs and 2 outputs as illustrated in the table.

| DMU | Y1 | Y2 | X1 | X2 |
|---|---|---|---|---|
| 1 | 100 | 1 | 200 | 2 |
| 2 | 1 | 100 | 1 | 100 |

As an extreme case, let us assume that $Y_1$ is impacted only by $X_1$ and $Y_2$ only by $X_2$. In the language used above, we could consider the DMUs as possessing 2 business units or subunits $(I_1, R_1) = (X_1, Y_1)$ and $(I_2, R_2) = (X_2, Y_2)$. If we apply model (13.1) at the subunit level we would do two analyses, namely an analysis for the first bundle and one for the second. In the first analysis we have only one input $X_1$ and one output $Y_1$. Clearly DMU2 outperforms DMU1 in this subunit in the CRS context as per model (13.1); DMU2 uses one unit of input $X_1$ to produce one unit of output $Y_1$, whereas DMU1 requires 200 units of input $X_1$ to produce 100 units of $Y_1$. Under constant returns to scale in (13.1), the efficiency score for DMU1 would be 50 %, whereas DMU2 would get a score of 100 %. DMU1 would be deemed inefficient and DMU2 efficient. If we go through the same line of argument with the second subunit (second analysis), the same result would happen, namely DMU1 would get a score of 50 % while DMU2 rates at 100 %. Thus, if the overall score for each DMU is to be some weighted average of the two subunit scores, no matter what weights are used, DMU1 will score 50 % and DMU2 100 %.

If one ignores the partial impact information and applies model (13.1) directly, which means assuming both inputs impact both outputs, it would be concluded that DMU #1 has an efficiency score of 100 %. This is arrived at by having that DMU put all output weight on $Y_1$ (zero weight on $Y_2$), and all input weight on $X_2$ (zero weight on $X_1$). In so doing, the conventional model applied at the full DMU level clearly ignores the input to output impacts involved, namely that $Y_1$ is impacted only by $X_1$. At the same time, DMU2 would put its output weight entirely on $Y_2$ and $X_1$, giving it a score of 100 %, but violating the input to output impacts.

This illustration clearly demonstrates that the conventional model generally makes no allowance for adhering to partial impacts that may be present.

The concept of partial input to output impacts has attracted some attention in the literature. In earlier work by Cook and Hababou (2001) and by Cook et al. (2000), a related scenario was presented wherein bank branches were viewed as consisting of two components, sales and service. While the model developed therein provides for an aggregate measure of efficiency of the overall branch, it fails to properly connect that measure to the partial measures for the sales and service components. The current chapter facilitates the link between these two sets of measures. The model given here is, as well, somewhat related to the work on network DEA as proposed by Fare and Grosskopf (1996). Their methodology is aimed primarily at describing the internal sub-processes in the DMU, hence it may be argued that the model here is a type of network DEA approach to efficiency. Arguably, one difference between our methodology and that characterizing network DEA is that our definition of the overall performance of the DMU is that it is a weighted average of the subunit efficiencies. Network DEA provides no clear connection between the efficiency score for the overall DMU and the scores of the sub-processes. We provide that connection in the methodology presented here.

To capture partial interactions, as described above, let us suppose that a DMU is viewed as a business unit comprised of $K$ independent subunits. To derive a measure of efficiency of the individual DMU, we propose proceeding in three steps. Step1 derives the "split" variables $\alpha_{ik}$ representing the portions of inputs $i$ to be assigned to bundles $k$. We point out that the situation involving inputs that are non-separable is examined later in the paper. In step 2 each subunit $k$ can be treated as a DMU, and the conventional DEA model of the type (13.1) is applied, using the outputs for that subunit and inputs given by the $\alpha_{ik}x_{ij}$. Finally, in step 3 the subunit scores as derived in step 2 are combined to give an overall score for the DMU. We discuss these three steps in detail.

**Step 1: Deriving the Split of Inputs Across Subunits** It can be argued that if efficiency scores of the $K$ subunits of a DMU $j_o$ were available, then some weighted average of those scores would reasonably represent the overall efficiency score of that DMU. Below we discuss how appropriate weights might be selected. It should be noted at the outset that splitting the DMU into subunits, and then combining subunit level scores to get an overall score for the DMU, may not be appropriate in cases where economies or diseconomies of scope are present. See, for example Panzar and

Willig (1981), and Pulley and Braunstein (1992). In such an eventuality, it would be necessary to account for such synergies before combining the subunit scores.

Since we wish to maximize the aggregate efficiency of each DMU, and given the fact that this aggregate will be represented (as shown later) as a convex combination of the $K$ subunit efficiencies, using some set of weights $W_{kj}$, we must first determine an appropriate $\alpha$-split of inputs. As a convention, we choose (in keeping with model (13.1)), to apply the CRS input oriented radial projection model for any DMU $j_o$, shown as (13.2). We point out that one might choose alternatively to apply a variable returns to scale (VRS) model along the lines of Banker et al. (1984), in which case the approach taken herein can be easily adapted.

$$e_{agg} = \max \sum_{k=1}^{K} W_{kj_o} \left[ \sum_{r \in R_k} u_r y_{rj_o} \Big/ \sum_{i \in I_k} v_i \alpha_{ik} x_{ij_o} \right] \tag{13.2a}$$

*Subject to* :

$$\sum_{k=1}^{K} W_{kj} \left[ \sum_{r \in R_k} u_r y_{rj} \sum_{i \in I_k} v_i \alpha_{ik} x_{ij} \right] \leq 1, \qquad \forall j \tag{13.2b}$$

$$\sum_{r \in R_k} u_r y_{rj} \Big/ \sum_{i \in I_k} v_i \alpha_{ik} x_{ij} \leq 1, \qquad \forall k, j \tag{13.2c}$$

$$\sum_{k \in L_i} \alpha_{ik} = 1, \qquad \forall i \tag{13.2d}$$

$$a_{ik} \leq \alpha_{ik} \leq b_{ik}, \qquad \forall i, k \tag{13.2e}$$

$$u_r, v_i, \alpha_{ik} \geq \varepsilon, \qquad \forall i, k \tag{13.2f}$$

In this formulation, we use the notation $e_{agg}$ to denote the 'aggregate' efficiency score for the DMU. The weights $W_{kj}$ are intended to represent the importance to be attached to each subunit for the DMU $j$ under consideration. It is appropriate in many situations to represent the importance of a subunit (to the overall business) by the proportion of inputs assigned to or consumed by that subunit. For example, a weight of 30 % is assigned to the efficiency ratio of a subunit if 30 % of the inputs are consumed by that subunit. Adopting this line of argument, we therefore define

$$W_{kj} = \sum_{i \in I_k} v_i \alpha_{ik} x_{ij} \Big/ \sum_{k=1}^{K} \left[ \sum_{i \in I_k} v_i \alpha_{ik} x_{ij} \right] \tag{13.3}$$

In addition, as specified by constraints (13.2c), the variables $\alpha_{ik}$ should be selected in such a way that the efficiency score pertaining to each subunit $k$ of DMU $j$ not exceed unity for some values of the multipliers $u_r, v_i$. We point out that in the presence of constraints (13.2c), constraints (13.2b) are rendered redundant, and may therefore be

dropped from the model. Constraints (13.2d) impose the usual convexity restriction on the $\alpha_{ik}$ values within each subunit $k$, and for each input $i$ that applies to that subunit. The set $L_i$ in constraints (13.2d) is defined as the set of all subunits $k$ that have $i$ as a member. Finally, to place limits on the size of the $\alpha$ variables, constraints (13.2e) are imposed.

By virtue of the definition of $W_{kj}$ as proposed in (13.3), the objective function (13.2a) may be written as

$$e_o = \max \sum_{k=1}^{K} \sum_{r \in R_k} u_r y_{rj_o} \Big/ \sum_i v_i x_{ij_o} \tag{13.1a'}$$

To transform the current nonlinear structure of (13.2) to a more tractable form, we make the change of variables $z_{ik} = v_i \alpha_{ik}$, and note that

$$\sum_{k \in L_i} \alpha_{ik} = 1 \Rightarrow \sum_{k \in L_i} v_i \alpha_{ik} = v_i \Rightarrow \sum_{k \in L_i} z_{ik} = v_i$$

Employing the standard Charnes and Cooper (1962) transformation $t = 1/\Sigma_i v_i x_{ij_o}$, and defining $\mu_r = tu_r, \upsilon_i = tv_i, \gamma_{ik} = tz_{ik}$. Problem (13.2) becomes:

$$e_{agg} = \sum_{k=1}^{K} \sum_{r \in R_k} \mu_r y_{rj_o} \tag{13.4a}$$

*Subject to:*

$$\sum_i \upsilon_i x_{ij_o} = 1 \tag{13.4b}$$

$$\sum_{r \in R_k} \mu_r y_{rj} - \sum_{i \in I_k} \gamma_{ik} x_{ij} \le 0, \forall j, k \tag{13.4c}$$

$$\sum_{k \in L_i} \gamma_{ik} = \upsilon_i, \quad \forall i \tag{13.4d}$$

$$\upsilon_i a_{ik} \le \gamma_{ik} \le \upsilon_i b_{ik}, \tag{13.4e}$$

$$\mu_r, \upsilon_i, \gamma_{ik} \ge \varepsilon, \quad \forall r, i, k \tag{13.4f}$$

**Step 2: Deriving Subunit Efficiencies** From the solution of (13.4) we can derive the resource splitting variables $\alpha_{ik}$, specifically $\alpha_{ik} = \gamma_{ik}/\upsilon_i$. This provides an appropriate apportioning of inputs ($\alpha_{ik} x_{ij}$) to their respective subunits. We now wish to evaluate the efficiencies of those subunits within the DMU. To this end, model (13.1) is applied, but with the understanding that the "DMU" being evaluated is the $k$th subunit whose outputs are the members of $R_k$ in the amounts $y_{rj}$, and whose inputs are the members of $I_k$ in the amounts $\alpha_{ik} x_{ij}$.

**Step 3: Deriving the Overall Efficiency Scores for the DMUs** The overall efficiency score $e_{ove}$ of the DMU is derived in this stage by taking a weighted average of the $k$ subunit scores obtained in Stage 2, using the $W_{kj}$ defined in (13.3). It should be pointed out that in computing $W_{kj}$ an appropriate set of input multipliers $v_i$ needs to be chosen. Furthermore, the multipliers need to be computed in an environment where all subunits are being compared simultaneously. The aggregate model (13.4) provides such an environment. That is, in (13.4) when DMU $j_o$ is being evaluated, the input portion of expression (13.4c), namely $\Sigma_{i\in I_k}\gamma_{ik}x_{ij_o}$ (for $j = j_o$), represents the *value* of that DMU's resources that are assigned to subunit $k$. The *total value* of all resources consumed by DMU $j_o$ is given by $\Sigma_{i\in I}v_i x_{ij_o}$, which of course is scaled to unity as per constraint (13.4b). Hence, the weights $W_{kj_o}$ reduce to $W_{kj_o} = \Sigma_{i\in I_k}\gamma_{ik}x_{ij_o}$. Note again that this set of weights is dependent on the particular DMU $j_o$ under investigation, to reflect the fact that the proportion of inputs allocated to the $k$th subunit is at the discretion of the DMU under consideration. The following theorem (see proof in Appendix 1) establishes that the overall efficiency score $e_{ove}$ for a DMU arising from Step 3 is greater than or equal to the aggregate score $e_{agg}$ derived from Model (13.4)

**Theorem 3.1**

$$e_{ove} \geq e_{agg}$$

The above discussion is centered on the idea of partitioning the input-output set into $K$ bundles $(I_k, R_k)$.We now discuss the formation of these bundles.

**Generating the Input/Output Bundles** In a multiple input/output setting, for each $k = 1,\ldots K$, let $I_k$ and $R_k$ represent a set of inputs and outputs respectively. We need to first generate input/output bundles $(I_1, R_1), (I_2, R_2), \ldots (I_k, R_k)$ in a way that the $R_{k=1}^{K}$ form a mutually exclusive set, and for each $k$, $(I_k, R_k)$ is maximal.

**Definition 3.1** An input/output bundle $(I_k, R_k)$ is said to be *maximal* if it possesses the following two properties:

1) Every output $r$ in $R_k$ is influenced by every input $i$ in $I_k$, and no other input outside of $I_k$ influences any output $r$ in $R_k$; and
2) There exists no output outside of $R_k$ whose input bundle is identical to that of $R_k$.

There can, however, be an input $i_0$ in a given bundle $I_k$ that influences an output $r_0$ outside of $R_k$, but at least one $i$ in $I_k$ does not influence $r_0$.

An algorithm for generating maximal bundles in a given multiple input/output setting appears in Appendix 2.

**Theorem 3.2** The generated set of maximal input/output bundles is unique.

See Appendix 1 for proof.

## 13.4 AR Restrictions on Pairs of Input Variables

Since the introduction of the original DEA model, a number of extensions such as the assurance region (AR) model of Thompson et al. (1990) have been proposed as a means of restricting the relative sizes of multipliers. It is generally argued that in the absence of such restrictions, the conventional DEA model may fail to deliver acceptable results. In the present setting involving the measurement of the technical efficiencies of a set of steel fabrication plants, the per unit costs of inputs are known, at least within certain bounds, and therefore AR restrictions are needed to insure that these cost bounds are adhered to.

In the conventional setting where we wish to restrict the magnitude of input multipliers relative to one another, AR constraints might take the form: $c_L \leq \nu_2/\nu_1 \leq c_U$ or alternatively $c_L \nu_1 \leq \nu_2 \leq c_U \nu_1$. Thus, for example, if $c_L = 2$ and $c_U = 3$, this constraint stipulates that the magnitude of the multiplier for input 2 must be at least twice that of the multiplier for input 1, but not more than three times the size of that multiplier. Thompson et al. (1990) go on to suggest that an appropriate format in which to specifying AR constraint, when we wish to impose several restrictions on pairs of multipliers, is to choose one of the multipliers as the *numeraire* or base against which the other multipliers would be compared. Hence, in this case $\nu_1$ is taken as the base against which to express the relative importance of the various multipliers.

In this section, we re-examine the model presented in the previous section and present a modified version of the model that allows for the imposition of multiple multiplier restrictions in the form of AR constraints. Although our focus in this paper is on input multipliers, the methodology given here is applicable to both input and output multipliers.

Consider a DEA setting in which partial input-to-output interactions are present. Based on the methodology presented in Sect. 13.3, each DMU can be viewed as a business unit comprised of separate subunits. In this setting the imposition of AR constraints presents challenges that do not arise in the conventional DEA situation, with the major challenge being that of having more than one AR restriction involving the same pair of variables. This may happen when say two inputs are part of two or more input bundles $I_k$, and the relative importance of these two variables is different in one input bundle than in another. For example, in the setting described in Sect. 13.2, inputs 2 and 3 both appear as members of $I_1$ and of $I_2$. Thus, an AR restriction on the pair in $I_1$might take the form $2 \leq \nu_3/\nu_2 \leq 4$, while in $I_2$ the restriction on the same pair might appear as $3 \leq \nu_3/\nu_2 \leq 9$. In a bank branch setting for instance, the relative importance attached to counter staff versus financial services staff can be different when they are performing routine service activities versus when they perform tasks relating to loans and investments. Thus, the need to evaluate performance at both unit and subunit levels, calls for a methodology that can handle multiple and often conflicting AR constraints.

Let us now assume there are multiple AR constraints imposed on each subunit $k$. We will assume for the moment that no AR restrictions involve pairs of inputs

wherein one member of the pair is in one input bundle while the other member of the pair is in a different bundle. We examine this phenomenon later.

For notational purposes, let us denote an AR pair by $(\upsilon_{i_1}, \upsilon_{i_2})$ meaning that a pair of assurance region constraint of the form $c_L \leq \upsilon_{i_1}/\upsilon_{i_2} \leq c_U$ is to be imposed. In addition, let $AR_k$ represent the set of all pairs $(\upsilon_{i_1}, \upsilon_{i_2})$ corresponding to AR constraints $c_L \leq \upsilon_{i_1}/\upsilon_{i_2} \leq c_U$ on pairs of multipliers $\upsilon_i$, $i \in I_k$, $k = 1,..,K$.

In settings where partial input to output interactions are present, we shall assume for the time being that AR constraints arise at the level of the subunit. That is, as stated above, we assume for now that there are no AR constraints that cross subunits. Further, we assume that within any subunit, all AR restrictions are stated in terms of a single numeraire, that is one of the members of that subunit will have been chosen as the numeraire against which other members will be compared. A difficulty that arises is the fact that a numeraire may not exist that is common across all subunits, and as a result, it may not be possible to express all constraints in terms of a single numeraire. One approach to this problem is to partition all AR constraints into L *mutually exclusive* sets $\Theta_l$ so that in any $\Theta_l$ all AR constraints can be expressed in terms of a single numeraire. Specifically, all AR constraints in any set $\Theta_l$ take the form

$$c_{iL}^k \leq \upsilon_i/\upsilon_{i\theta_l} \leq c_{iU}^k \tag{13.5}$$

where $\upsilon_{\iota\Theta_l}$ denotes the numeraire in set $\Theta_l$. We index the upper and lower bounds with the superscript $k$ to signify the subunit in which the particular AR constraints originate.

The algorithm for generating these mutually exclusive sets is discussed in appendix 1.

It should be noted that each generated set $\Theta_l$ is comprised of AR pairs that are either connected to each other explicitly, through a common multiplier, or implicitly through an AR set involving a common multiplier. For example, consider the following AR sets: $AR_1 = \{(\upsilon_2, \upsilon_3)\}, AR_2 = \{(\upsilon_3, \upsilon_4)\}, AR_3 = \{(\upsilon_4, \upsilon_5)\}, AR_4 = \{(\upsilon_6, \upsilon_7)\}$.

It can be observed that $AR_1$ is connected explicitly to $AR_2$ through their common multiplier $\upsilon_3$, while it is only *implicitly* connected to $AR_3$ through its connection to $AR_2$.$AR_4$, however, is not connected to any of the other sets, and as a result, the four mentioned AR sets are divided into the following two mutually exclusive sets

$$\Theta_1 = \{(\upsilon_2, \upsilon_3), (\upsilon_3, \upsilon_4), (\upsilon_4, \upsilon_5)\}, \Theta_2 = \{(\upsilon_6, \upsilon_7)\}.$$

Specifically, there are no common multipliers connecting any two AR pairs in $\Theta_1$ and $\Theta_2$.

As indicated earlier, it can turn out that for any given pair of inputs, there can be multiple AR constraints involving that pair. Reiterating statements made earlier, we assume that AR constraints originate from within the subunits $k$, prompting the use of the superscript $k$ in (13.5). In order to convert multiple sets of AR restrictions (involving a given pair of input multipliers) to a single AR constraint, we focus on one of the bounds in each case, say the lower bound $c_{iL}^k$. Consider the example

given earlier, namely $2 \leq \upsilon_3/\upsilon_2 \leq 4$, reflecting the relative importance of inputs 3 and 2 in subunit $k=1$, while $3 \leq \upsilon_3/\upsilon_2 \leq 9$ reflects the relative importance of the two inputs from the standpoint of subunit $k=2$. It is observed that the expression $3 \leq \upsilon_3/\upsilon_2 \leq 9$ can be replaced by $\frac{2}{3}3 \leq \frac{2}{3}\upsilon_3/\upsilon_2 \leq \frac{2}{3}9$, hence $2 \leq \frac{2}{3}\upsilon_3/\upsilon_2 \leq 6$. Let us define $\upsilon_3' = \frac{2}{3}\upsilon_3$. Now in $I_2$ the weighted version of input #3, namely $\upsilon_3 x_3$, can be replaced by $\frac{2}{3}\upsilon_3(\frac{3}{2}x_3)$ or $\upsilon_3'(\frac{3}{2}x_3)$. Specifically, if we agree to *scale up* the data for $x_3$ in $I_2$by a factor 3/2, we can then *scale down* the multiplier $\upsilon_3$ by the reciprocal of that factor (thereby creating a new factor $\upsilon_3'$). In this way, multiple lower bounds arising from the presence of multiple AR constraints can be reduced to a single lower bound by scaling the data in those subunits where the lower bounds are undergoing adjustment.

To formalize these ideas, suppose that there are two AR restrictions arising from two units $k_1$ and $k_2$, namely $c_{iL}^{k_1} \leq \upsilon_i/\upsilon_{i\Theta_l} \leq c_{iU}^{k_1}$ and $c_{iL}^{k_2} \leq \upsilon_i/\upsilon_{i\Theta_l} \leq c_{iU}^{k_2}$, involving multiplier $\upsilon_i$ and the associated numeraire $\upsilon_{i\Theta_l}$. Further, suppose that we wish to use the lower limit $c_{iL}^{k_1}$ as the base value to which we aim to reduce all other lower limits for AR restrictions involving these two variables. Following the logic of the above example we note that

$$c_{iL}^{k_2} \leq \upsilon_i/\upsilon_{i\Theta_l} \leq c_{iU}^{k_2} \Leftrightarrow \frac{c_{iL}^{k_1}}{c_{iL}^{k_2}} c_{iL}^{k_2} \leq \frac{c_{iL}^{k_1}}{c_{iL}^{k_2}} \upsilon_i/\upsilon_{i\Theta_l} \leq \frac{c_{iL}^{k_1}}{c_{iL}^{k_2}} c_{iU}^{k_2}$$

$$\Leftrightarrow c_{iL}^{k_1} \leq \frac{c_{iL}^{k_1}}{c_{iL}^{k_2}} \upsilon_i/\upsilon_{i\Theta_l} \leq \frac{c_{iL}^{k_1}}{c_{iL}^{k_2}} c_{iU}^{k_2}$$

If we define the transformed variable $\upsilon_{ik_2}' = c_{iL}^{k_1}\upsilon_i/c_{iL}^{k_2}$, then in $I_{k_2}$ the weighted input $\upsilon_i x_i = \upsilon_{ik_2}'(\frac{c_{iL}^{k_2}}{c_{iL}^{k_1}})x_i$, where the lower limit $(c_{iL}^{k_1})$ on the AR restriction involving $\upsilon_{ik_2}'/\upsilon_{i\Theta_l}$ is the same as the lower limit involving $\upsilon_i/\upsilon_{i\Theta_l}$. Hence, in $I_{k_2}$ we may proceed by replacing the data for $x_i$ by the scaled data $(\frac{c_{iL}^{k_2}}{c_{iL}^{k_1}})x_i$, and replace the notation $\upsilon_{ik_2}'$ by $\upsilon_i$, where the lower limit in the AR restriction on $\upsilon_i$ relative to $\upsilon_{i\Theta_l}$ (in $I_{k_2}$) is now given by $c_{iL}^{k_1}$ rather than $c_{iL}^{k_2}$. Hence, we may scale 'down' the lower limit for $\upsilon_i$ in $I_{k_2}$ provided we scale 'up' the data $x_i$ in that subunit. This exercise can now be repeated for all other subunits $k_3, k_4, \ldots$. that have AR restrictions involving these two variables. In each case the lower limit on the ratio of the two multipliers is

$$\bar{c}_{iL} = c_{iL}^{k_1}. \tag{13.6}$$

Along the same lines, the upper bound $c_{iU}^{k_2}$ on $\upsilon_i/\upsilon_{i\Theta_l}$ is replaced by $\bar{c}_{iU}^{k_2} = (\frac{c_{iL}^{k_1}}{c_{iL}^{k_2}})c_{iU}^{k_2}$. This is again repeated for all other subunits $k_3, k_4, \ldots$. Now let

$$\bar{c}_{iU} = \min\{\bar{c}_{iU}^{k_1}.\bar{c}_{iU}^{k_2}, \ldots..\} \tag{13.7}$$

Expression (13.7) can now be replaced by

$$\bar{c}_{iL} \leq \frac{\upsilon_i}{\upsilon_{i\theta_l}} \leq \bar{c}_{iU} \tag{13.8}$$

After the required data adjustments are made, model (13.4) with any applicable AR restrictions can be applied to derive the split of inputs across the subunits. This analysis is then followed by running model (13.1) for each subunit as the DMU, and again all applicable AR constraints are imposed.

## 13.5 Other Considerations

**The Case of Non-Separable Inputs** In many instances, there can be inputs that do not lend themselves to subdivision in the manner described above. If, for example, in the analysis of the steel fabrication plants, one wished to include as an input a quality measure pertaining to supplier reliability, it would appear to be unreasonable to suggest subdividing this factor, and assigning portions of it across the various subunits. This factor, in its entirety, is assumed to affect the outputs in each subunit $k$. Generalizing, let us assume that there are $I_{ns}$ inputs of this non-separable type. The efficiency ratio for a given $k$ within DMU $j_o$ can now be expressed in the form $\Sigma_{r\in R_k} u_r y_{rj_o}/(\Sigma_{i\in I_k} v_i \alpha_{ik} x_{io} + \Sigma_{i\in I_{ns}} v_i^k x_{io})$ where $v_i^k$ is the worth or weight assigned to the non-separable input $x_{io}, i \in I_{ns}$, and represents the impact of that input on the outputs in subunit $k$. Note that we are permitting this weight to be different from one subgroup to another.

Following the logic of (13.3) we define the weight attached to the $k^{th}$ efficiency ratio by

$$W_k = \Big(\sum_{i\in I_k} v_i \alpha_{ik} x_{io} + \sum_{i\in I_{ns}} v_i^k x_{io}\Big) \Big/ \Big(\sum_k \Big[\sum_{i\in I_k} v_i \alpha_{ik} x_{io} + \sum_{i\in I_{ns}} v_i^k x_{io}\Big] \tag{13.9}$$

The optimization model for this more general case would be identical in form to (13.4) with the exception that constraint (13.4b) and (13.4c) are replaced by

$$\sum_{i\in I} \upsilon_i x_{io} + \sum_{i\in I_{ns}} \Big(\sum_k \upsilon_i^k\Big) x_{io} = 1 \tag{13.4b'}$$

and

$$\sum_{r\in R_k} \mu_r y_{rj} - \sum_{i\in I_k} \gamma_{ki} x_{ij} - \sum_{i\in I_{ns}} \upsilon_i^k x_{ij} \le 0, \forall k, j, \tag{13.4c'}$$

respectively, to account for the two types of inputs. Furthermore, constraints (13.4d) and (13.4e) apply only to separable inputs $i$. Here, $\upsilon_i^k = t v_i^k$ under the usual transformation as discussed above.

**Inter-Subunit Assurance Regions** In the above development, it is assumed that all AR constraints originate from within subunits. This means that when solving the stage 2 problem for any given subunit $k$, only AR constraints pertaining to that subunit, hence only multipliers for inputs contained in that subunit, are considered. There would be no reason to consider AR restrictions external to that subunit, since

they would relate to inputs and multipliers not relevant to $k$, and therefore would not affect the solution.

In the event that an AR restriction connects an input from a subunit $k$ to an input from a different subunit, then an argument can be made for including the restrictions relating to that other subunit during the evaluation of subunit $k$. In that regard if such 'connecting' AR constraints are present, it is recommend that a more complete version of (13.1), namely (13.1') be applied. Note that we explicitly index the multipliers (e.g. $u_{rk}, v_{ik}$) to denote that they are permitted to take different values from one subunit to another. As well, we use here the notation $x_{ikj}$ to denote the adjusted (by $\alpha_{ik}$) inputs.

$$
\begin{aligned}
& e = \max \sum_{r \in R_{k_o}} u_{rk_0} y_{r j_0} \Big/ \sum_{i \in I_{k_o}} v_{ik_o} x_{ik_0 j_0} \\
& \textit{subject to} \\
& \sum_{r \in R_k} u_{rk} y_{rj} \Big/ \sum_{i \in i_k} v_{ik} x_{ikj} \leq 1, \quad k = 1, \ldots K, j = 1, \ldots n \\
& All \quad AR_k \\
& u_{rk}, v_{ik} \geq 0, \quad \forall r, i, k
\end{aligned}
\tag{13.1'}
$$

## 13.6 Application

In this section, we demonstrate the application of the models presented in the previous sections to data on 20 steel fabrication plants as displayed in Appendix Table 13.3. We point out this data arose from a large survey of companies in the steel fabrication industry. The desire was to arrive at a relatively homogeneous set of companies, in the sense that they would all be producing similar product lines. Out of 230 companies contacted, 47 completed the questionnaire, and of those only 20 of the respondents supplied sufficient detail to enable a full scale analysis. The primary information collected included (1) a list of major products produced, from which Table 13.1 was constructed, (2) the input to output impacts as displayed in Fig. 13.1, and (3) cost estimates per quarter, from which Table 13.2 was constructed. The latter cost data point to the need to control the input multipliers in any DEA analysis undertaken. This requirement has been addressed through the construction of AR constraints.

The creation of AR limits in many applications is a challenge in that input multipliers, for example, may not have economic meaning. In the current setting one can directly interpret the input multipliers as per unit costs incurred during the analysis period. In this particular case the analysis period was the last quarter of 2010, meaning that $\upsilon_1, \upsilon_2, \upsilon_3, \upsilon_4$ represent that quarter's per unit cost to the plant relating to labor, shears, presses and lathes. In the case of labor ($x_1$) a range of estimates from \$5000 to \$7500 per plant employee (wages and benefits) for the quarter was provided. The reason for a range of wage rates has to do with the fact that the rate can

**Table 13.2** Input cost rates per machine per quarter

| Quarterly costs Thousands of dollars | | | |
|---|---|---|---|
| Input | k= 1 | k= 2 | k= 3 |
| Labor | \$5–\$7.5 | \$5–\$7.5 | \$5–\$7.5 |
| Shears | \$7–\$10 | \$14–\$19 | |
| Presses | \$6–\$9.6 | \$16–\$21 | |
| Lathes | | \$2–\$3.5 | \$4–\$7.4 |

vary from plant to plant and over time due to the mix of full time and part time labor used, and the amount of overtime during peak demand times. There was no implied variation in labor costs across the three bundles, $k = 1,2,3$. In the case of the other three inputs, machine 'rates' were taken to be the estimated quarterly costs of depreciation, routine maintenance and unforeseen breakdown costs. In the case of the shearing machines, for example, it was estimated that the quarterly cost (depreciation and maintenance) of operating one machine would generally vary between \$7000 and \$10,000 per quarter in the case of output bundle $k = 1$, and \$14,000 and \$19,000 in the case of $k = 2$. The difference in cost between the two product groupings is explained by the increased stress placed on the equipment in the production of flat bar products versus that created in the manufacture of sheet steel products. Table 13.2 displays the ranges of quarterly costs.

This table allows one to set AR constraints corresponding to the various pairs of multipliers. Since labor is common to all subgroups, we have chosen it as the numeraire for expressing all ratio constraints. As an example, since the range for labor cost is \$5–\$7.5, and for shears the range is \$7–\$10 in the case of k = 1, we maintain that the AR constraints linking $\upsilon_2$ and $\upsilon_1$is given by $\frac{7}{7.5} \leq \frac{\upsilon_2}{\upsilon_1} \leq \frac{10}{5}$; that is, the lower limit on the ratio of the two multipliers is defined as the ratio of the lowest value $\upsilon_2$ can take, divided by the highest value $\upsilon_1$ can assume, etc. Summarizing, the following are the requisite AR constraints corresponding to the supplied cost ranges on the four inputs.

*Subunit k = 1:* $.93 \leq \frac{\upsilon_2}{\upsilon_1} \leq 2, .8 \leq \frac{\upsilon_3}{\upsilon_1} \leq 1.92$

*Subunit k = 2:* $1.87 \leq \frac{\upsilon_2}{\upsilon_1} \leq 3/8, 2.13 \leq \frac{\upsilon_3}{\upsilon_1} \leq 4.2, .27 \leq \frac{\upsilon_4}{\upsilon_1} \leq .7$

*Subunit k = 3:* $.53 \leq \frac{\upsilon_4}{\upsilon_1} \leq 1.48$

Prior to running model (13.4) to derive the $\alpha_{ik}$, it is useful to compute a set of overall efficiency scores, using (13.1). The resulting scores appear in Table 13.4. In this analysis we ignore the partial relationships displayed in Fig. 13.1, and note that half of the DMUs are efficient. This provides a useful starting point for evaluating the performance of the various plants.

Now, applying model (13.4), in the presence of the above AR restrictions, we determine an aggregate score for each DMU (Table 13.5), and the corresponding $\alpha_{ik}$ as displayed in Table 13.6. We point out that the choice of an appropriate set of

upper and lower limits on the alpha variables, $a_{ik} \leq \alpha_{ik} \leq b_{ik}$, as per constraints (13.2e) presented some difficulty. While a secondary survey was conducted to get management's input to the issue of such limits, after establishing the input to output bundles, there tended to be reluctance to provide accurate estimates, meaning that no range was supplied in most cases. In the analysis herein we chose the range $0.1 \leq \alpha_{ik} \leq 0.6$ for input $i = 1$ (which has 3 subunits), and $0.4 \leq \alpha_{ik} \leq 0.6$ for $i = 2,3$ and 4 (which each have 2 subunits).

It is worth pointing out that in the majority of the cases, the alpha values arising from model (13.4) tended to gravitate toward their upper or lower limits. This is a common occurrence when optimizing a linear combination of two or more quantities. Specifically, the optimal thing to do is to force one of the quantities as small as possible thus freeing up resources for the other. The loss in one is generally dominated by the gain in the other, hence the highest overall efficiency for the DMU is often obtained by making one of the quantities involved (alpha in this case) as high as possible, with the other being as low as possible. Of course, one can impose limits on how much resource (the alpha split) one wants to take away from one subunit and give to another.

The $\alpha_{ik}$ are then used to scale the inputs to the levels $\alpha_{ik} x_{ij}$, and model (13.1) is now applied at the subunit level along with the necessary AR constraints for that subunit. The resulting subunit scores appear in Table 13.8. To combine these subunit scores to get an overall efficiency score for each DMU (Step 3), we use the subunit weights $W_{kj_o} = \Sigma_{i \in I_k} \gamma_{ik} x_{ij_o}$ extracted from (13.4c) when the aggregate scores (Table 2) were being derived. These weights are presented in Table 13.7. The corresponding "Overall Scores" for the 20 plants appear as the last column in Table 13.8.

It is observed that the $W_{kj}$ for any subunit $k$ show a wide variation across the 20 DMUs. In the case of subunit $k = 1$, for example, this variation is from 12.7 % for DMU #8 to 47.7 % for DMU #17. These weights are derived as part of the process of allocating resources among the subunits of a DMU such as to maximize its aggregate performance. No attempt is made to restrict weight variation across DMUs; the size of the weight on any given subunit $k$ for a DMU $j$, reflects the proportion of resources that DMU is dedicating to the subunit in question.

## 13.7 Discussion and Conclusions

This chapter has presented a methodology for efficiency measurement of DMUs in situations where not all inputs impact all outputs. The model is based on viewing a DMU as a business unit comprised of a set of subunits in each of which the conventional DEA model properly applies. The overall efficiency score of the DMU is then derived by combining the efficiency scores of the subunits. This model is then modified to allow for the imposition of restrictions on multipliers in form of AR constraints.

Our approach conveys important information about the inner-workings of the DMU, providing insights as to which parts of the 'business unit' are operating at what

levels of efficiency. A somewhat related type of analysis, designed to examine the internal structure of the DMU, is network DEA, as discussed in Fare and Grosskopf (1996), Cook et al. (2010a, 2010b). In the text herein we compare our work to this previous literature. From a production standpoint, it is important to gain an understanding about how the different parts of an organization are performing, and to characterize, to the greatest extent possible, the actual input-output interactions. Resource sharing among subunits can then be addressed.

Arguably, the most relevant information for management in regard to the operating efficiency of the business unit (the DMU) under their control is the partial efficiencies for the set of subunits making up that business unit (Table 13.8), and the corresponding weights displayed in Table 13.7. The subunit scores point to where the strengths and weaknesses of the business unit lie, while the weights indicate the proportion of resources (weighted by value) dedicated to each of those subunits. The input oriented view of performance revolves around the idea that inefficient DMUs, if projected to the frontier can attain a 100 % efficient status by reducing resource consumption. In that regard some managers view the subunit efficiencies as a vehicle for undertaking resource sharing. In simplistic terms, management may undertake to extract resources from the less efficient subunits and move those resources to the more efficient subunits. The role of the weights in this resource shifting exercise is to signal whether the amount of resource involved is significant or not. Consider, for example, the case of DMU #8. Here, the subunit scores for $k = 1,2,3$ are 22.7, 42 and 65.3 % respectively. One might argue that subunit #3 is approximately 3 times as successful as subunit #1 at transforming inputs to outputs, and it would appear that a resource shift from #1 to #3 might be advisable. However, the percentage of DMU #8's resources consumed by the first subunit, as per Table 13.7, is only 12.7 %, meaning that the amount of transferrable resource is likely rather minimal, and in fact may already be at a critically low level. If one compares the proportions of DMU resources among all 20 of the $k = 1$ subunits, the proportion for the 8$^{th}$ DMU is the lowest among the peers (see column $k = 1$ in Table 13.7). Subunit #2, however, currently consumes 35.8 % of the DMUs resources, and a significant portion of that may be transferrable to subunit #3.

There is no obvious formula for guiding resource transfers within a DMU. The primary role of DEA and its offshoots has been to signal where inefficiencies lie, and it is left to management to use efficiency scores and the related resource usage to guide the efficiency improvement exercise. Subunit performance measures (Table 13.8) and the related resource consumption figures (Table 13.7) provide a more in-depth view of the inner workings of the DMU, and will hopefully facilitate more informed decision making at the operations level.

Future work will investigate how features such as nondiscretionary variables and qualitative data might be treated in this partial input-to-output environment. Furthermore, as indicated earlier, the approach taken herein is only applicable when one assumes economies/diseconomies of scope are not present. Future research will examine how to model partial impacts when such a phenomenon is to be taken into account.

This chapter is based upon I. Imanirad, W. Cook and J. Zhu. 2013. Partial input to output impacts in DEA: Production considerations and resource sharing among business sub-units, **Naval Research Logistics,** 60(3), 190–207.

## 13.8 Appendix 1: Proofs of Theorems

Theorem 3.1 proof: The aggregate score (objective function value (13.4a)) is the weighted average of the $K$ subunit ratios of outputs to inputs $\Sigma_{r\in R_k}\mu_r y_{rj}/\Sigma_{i\in I_k}\gamma_{ik}x_{ij}$ arising from constraint (13.4c). The weights $W_{kj_o} = \Sigma_{i\in I_k}\gamma_{ik}x_{ij_o}$ are the same values as those used in step 3 where they are applied to the *maximal* values of these ratios (obtained in step 2) to arrive at $e_{ove}$. Since these latter subunit scores must be at least as large as those arising out of constraints (13.4c), the result follows.

Theorem 3.2 proof: Assuming that the set of maximal bundles is not unique then there must be at least two different sets of maximal bundles, S1 and S2. This implies that there must be at least one input/output bundle $B_k$ in S1 that is different from every bundle $B'_{k'}$ in S2. $B_k$ and $B'_{k'}$ may differ in terms of their respective input or/and output sets. If the input sets $I_k$ and $I'_{k'}$ are different, there must exist at least one input $i_k$ such that $i_k \in I_k$ and $i_k \notin I'_{k'}$. If $i_k$ influences any r $\in R_k$ then bundle $B'_{k'}$ violates the first requirement of a maximal bundle since there exists input $i_k$ outside of $I'_{k'}$ that influences r $\in R_k$. Otherwise, bundle $B_k$ violates the first requirement of a maximal bundle since $i_k \in I_k$ does not influence any r $\in R_k$. In either case there is only one maximal bundle.

In case of a difference between output sets $R_k$ and $R'_{k'}$, there must be at least one output $r_k$ such that $r_k \in R_k$ and $r_k \notin R'_{k'}$. If the input bundle of $r_k$ is not $I_k$ then bundle $B_k$ violates the first requirement of a maximal bundle since there exists an input $i \in I_k$ that does not influence $r_k$ or there exists an input outside of $I_k$ that influences $r_k$. Otherwise, if the input bundle of $r_k$ is equal to $I_k$ then bundle $R'_{k'}$ violates the second requirement of a maximal bundle since there exists output r outside of $R'_{k'}$ with an input bundle identical to that of $R'_{k'}$ In either case there can only be one maximal bundle. This completes the proof.

## 13.9 Appendix 2: Algorithms

**Generating Maximal Bundles** *Step 1:* Define S to be an empty set.

*Step 2:* For each output $r$, derive $I(r)$, the set of all inputs $i$ that influence $r$. Add $I(r)$ to S. Set the bundle counter as $k = 1$.

*Step 3:* For each $I(r)$ in S, compare it with every other $I(r')$ in S, and identify *all* $I(r')$ that have the same input elements as in $I(r)$. If no such $r'$ is identified, create bundle $(I_k, R_k)$ using $I(r)$ and $r$ so that $(I_k, R_k) = (I(r), r)$. Remove $I(r)$ from S. Go to Step 4. Otherwise, group outputs $r$ and all identified $r'$ (having the same input sets) together to derive $R_k$, and create bundle $(I_k, R_k)$ using $I(r)$ and $R_k$ so that $(I_k, R_k) = (I(r), R_k)$. Remove $I(r)$ and all identified $I(r')$ from S. Go to step 4.

*Step 4:* If $S$ is non-empty, set $k = k + 1$ and go to Step 3. Otherwise, terminate having formed the set of all bundles**.**

Example:

Step 1: Define the empty set *S.*

Step 2: Here, $I(1) = (1, 2, 3), I(2) = (1, 2, 3, 4), I(3) = (1, 4), I(4) = (1, 4)$. $S$ $=\{I(1), I(2), I(3), I(4)\}$. Set $k = 1$.

Step 3/4: For $r = 1$, $I(1)$ is different from all other $I(r')$**.** Thus, $(I_1, R_1)$ $= ((1, 2, 3), (1))$. Remove $I(1)$ from *S. S* now becomes the reduces set $\{I(2), I(3),$ $I(4)\}$. Here the process takes us to Step 4 where we discover S is non-empty, the counter $k$ is set to 2, and we return to Step 3. This time through, $I(2)$ is discovered not to have an input set identical with that of any of the remaining members of *S*, meaning that $(I_2, R_2) = ((1, 2, 3, 4), (1))$, and $S$ is reduced further to $\{I(3),$ $I(4)\}$. Step 4 now sets the counter to $k = 3$, and on reentering Step 3, $I(3)$ is checked against $I(4)$, revealing that they are the same, namely $I(3) = I(4) = (1, 4)$. The bundle $(I_3, R_3) = ((1, 4), (3, 4))$ is created, and these two members of *S* are now removed and in Step 4 the algorithm terminates with the three identified bundles above.

**Generating the Mutually Exclusive Assurance Region (AR) Sets** For a given multiple input/output setting in which partial interactions among inputs and outputs are present, the algorithm for generating the mutually exclusive AR sets works as follows:

*Step 1*: Let $\Theta_{l=1}$ be an empty set and S $= AR_1 \cup AR_2 \cup \ldots . AR_k$ represent the set of all AR pairs $(\upsilon_{i_1k}, \upsilon_{i_2k})$, $i \in I_k$, $k = 1,..,K$ in a given setting.

*Step 2*: Let $(\upsilon_{i_1k}, \upsilon_{i_2k})$ be any AR pair in *S*. Remove $(\upsilon_{i_1k}, \upsilon_{i_2k})$ from S and add it to $\Theta_l$.

*Step 3:* Compare each $(\upsilon_{i_1k}, \upsilon_{i_2k}) \in \Theta_l$ with every $(\upsilon'_{i_1k}, \upsilon'_{i_2k}) \in S$. If there exists a multiplier $\upsilon_i$ so that $\upsilon_i \in (\upsilon_{i_1k}, \upsilon_{i_2k})$ and $\upsilon_i \in (\upsilon'_{i_1k}, \upsilon'_{i_2k})$, remove $(\upsilon'_{i_1k}, \upsilon'_{i_2k})$ from S and add it to $\Theta_l$.

*Step 4*: If S is not empty, create an empty set $\Theta_{l=l+1}$ and go to *Step 2*.

## 13.10 Appendix 3: Tables

**Table 13.3** Data on 20 plants

| Outputs | | | | | Inputs | | | |
|---|---|---|---|---|---|---|---|---|
| | Sheet steel | Flat bar | Pipes/ Cylinders | Bearings | Labor | Shears | Presses | Lathes |
| DMU | *Y1* | *Y2* | *Y3* | *Y4* | *X1* | *X2* | *X3* | *X4* |
| 1 | 70 | 103 | 100 | 80 | 30 | 5 | 5 | 15 |
| 2 | 60 | 125 | 90 | 90 | 40 | 4 | 4 | 18 |
| 3 | 50 | 110 | 105 | 85 | 35 | 5.2 | 4.2 | 10 |
| 4 | 80 | 80 | 110 | 90 | 38 | 7 | 4.6 | 8.5 |
| 5 | 56 | 40 | 60 | 55 | 28 | 9 | 5.5 | 12.5 |
| 6 | 40 | 95 | 120 | 110 | 37 | 4.2 | 3.8 | 14 |
| 7 | 100 | 180 | 200 | 210 | 31 | 6 | 4.1 | 11 |
| 8 | 25 | 55 | 180 | 160 | 35 | 5 | 5 | 15 |
| 9 | 65 | 150 | 125 | 145 | 25 | 6.2 | 4.8 | 19 |
| 10 | 40 | 110 | 70 | 115 | 30 | 3 | 3.2 | 21 |
| 11 | 70 | 117 | 122 | 115 | 25 | 4 | 4 | 12 |
| 12 | 92 | 135 | 89 | 64 | 45 | 5 | 3.3 | 23 |
| 13 | 88 | 47 | 57 | 109 | 35 | 4.1 | 6 | 20.5 |
| 14 | 48 | 68 | 146 | 99 | 32 | 5.3 | 3.4 | 11.2 |
| 15 | 79 | 123 | 220 | 122 | 26 | 7.7 | 4.3 | 15.6 |
| 16 | 99 | 114 | 89 | 49 | 19 | 5.3 | 4.2 | 12.4 |
| 17 | 97 | 101 | 88 | 55 | 25 | 8 | 3 | 8.8 |
| 18 | 55 | 55 | 132 | 116 | 32 | 6 | 2.8 | 6.8 |
| 19 | 80 | 97 | 142 | 168 | 33 | 2.8 | 3.9 | 13.4 |
| 20 | 97 | 68 | 209 | 122 | 27 | 3.3 | 4.3 | 21.6 |

**Table 13.4** Efficiency scores—conventional DEA model

| DMU | Efficiency score |
|---|---|
| 1 | 0.71104 |
| 2 | 0.91729 |
| 3 | 0.69367 |
| 4 | 0.92629 |
| 5 | 0.48891 |
| 6 | 0.72387 |
| 7 | 1.00000 |
| 8 | 0.87044 |
| 9 | 1.00000 |
| 10 | 1.00000 |
| 11 | 0.94440 |
| 12 | 1.00000 |
| 13 | 0.78425 |
| 14 | 0.85418 |
| 15 | 1.00000 |
| 16 | 1.00000 |
| 17 | 1.00000 |
| 18 | 1.00000 |
| 19 | 1.00000 |
| 20 | 1.00000 |

**Table 13.5** Aggregate efficiency scores

| DMU | Aggregate score [Problem 3.4] |
|---|---|
| 1 | 0.56593 |
| 2 | 0.57443 |
| 3 | 0.55358 |
| 4 | 0.47876 |
| 5 | 0.30023 |
| 6 | 0.50491 |
| 7 | 0.97958 |
| 8 | 0.49988 |
| 9 | 0.78207 |
| 10 | 0.64049 |
| 11 | 0.78011 |
| 12 | 0.59318 |
| 13 | 0.44862 |
| 14 | 0.54204 |
| 15 | 0.81013 |
| 16 | 0.85452 |
| 17 | 0.70308 |
| 18 | 0.55393 |
| 19 | 0.74080 |
| 20 | 0.78130 |

**Table 13.6** $\alpha_{ik}$ values resulting from Model (13.4)

| DMU | $X_1$ K = 1 | $X_1$ K = 2 | $X_1$ K = 3 | $X_2$ K = 1 | $X_2$ K = 2 | $X_3$ K = 1 | $X_3$ K = 2 | $X_4$ K = 2 | $X_4$ K = 3 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.16101 | 0.6 | 0.23899 | 0.4 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 |
| 2 | 0.15770 | 0.6 | 0.24230 | 0.4 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 |
| 3 | 0.1 | 0.6 | 0.3 | 0.4 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 |
| 4 | 0.43178 | 0.1 | 0.46822 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |
| 5 | 0.6 | 0.1 | 0.3 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |
| 6 | 0.1 | 0.3 | 0.6 | 0.4 | 0.6 | 0.4 | 0.6 | 0.4 | 0.6 |
| 7 | 0.1 | 0.45220 | 0.44780 | 0.4 | 0.6 | 0.6 | 0.4 | 0.4 | 0.6 |
| 8 | 0.1 | 0.3 | 0.6 | 0.52476 | 0.47524 | 0.6 | 0.4 | 0.4 | 0.6 |
| 9 | 0.1 | 0.6 | 0.3 | 0.4 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 |
| 10 | 0.1 | 0.6 | 0.3 | 0.4 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 |
| 11 | 0.16101 | 0.6 | 0.23899 | 0.4 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 |
| 12 | 0.3 | 0.6 | 0.1 | 0.4 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 |
| 13 | 0.6 | 0.1 | 0.3 | 0.6 | 0.4 | 0.6 | 0.4 | 0.6 | 0.4 |
| 14 | 0.11586 | 0.28414 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |
| 15 | 0.11586 | 0.28414 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |
| 16 | 0.3 | 0.6 | 0.1 | 0.6 | 0.4 | 0.6 | 0.4 | 0.6 | 0.4 |
| 17 | 0.6 | 0.3 | 0.1 | 0.6 | 0.4 | 0.6 | 0.4 | 0.6 | 0.4 |
| 18 | 0.3 | 0.1 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |
| 19 | 0.16101 | 0.23899 | 0.6 | 0.4 | 0.6 | 0.6 | 0.4 | 0.4 | 0.6 |
| 20 | 0.3 | 0.1 | 0.6 | 0.6 | 0.4 | 0.6 | 0.4 | 0.4 | 0.6 |

**Table 13.7** $W_{kj}$ values arising from model (13.4)

| DMU | K = 1 | K = 2 | K = 3 |
|---|---|---|---|
| 1 | 0.19194 | 0.62761 | 0.18045 |
| 2 | 0.16604 | 0.64466 | 0.18931 |
| 3 | 0.13676 | 0.64520 | 0.21804 |
| 4 | 0.36738 | 0.31385 | 0.31876 |
| 5 | 0.46694 | 0.29654 | 0.23652 |
| 6 | 0.13714 | 0.46883 | 0.39403 |
| 7 | 0.14616 | 0.48707 | 0.36676 |
| 8 | 0.12775 | 0.35810 | 0.51415 |
| 9 | 0.12952 | 0.63485 | 0.23563 |
| 10 | 0.12791 | 0.63576 | 0.23633 |
| 11 | 0.19129 | 0.62702 | 0.18169 |
| 12 | 0.24474 | 0.63646 | 0.11880 |
| 13 | 0.46385 | 0.29455 | 0.24161 |
| 14 | 0.14834 | 0.34083 | 0.51083 |
| 15 | 0.15841 | 0.35727 | 0.48432 |
| 16 | 0.29551 | 0.57877 | 0.12572 |
| 17 | 0.47872 | 0.42737 | 0.09391 |
| 18 | 0.28475 | 0.26070 | 0.45455 |
| 19 | 0.17764 | 0.36293 | 0.45944 |
| 20 | 0.27453 | 0.26986 | 0.45561 |

**Table 13.8** Subunit scores from model (13.1) and overall efficiency scores

| DMU | Aggregate score [Problem 3.4] | Score K1 | Score K2 | Score K3 | Overall score |
|---|---|---|---|---|---|
| 1 | 0.56593 | 0.68155 | 0.63854 | 0.62323 | 0.64403 |
| 2 | 0.57443 | 0.79875 | 0.88356 | 0.50189 | 0.79722 |
| 3 | 0.55358 | 0.68705 | 0.66773 | 0.85252 | 0.71066 |
| 4 | 0.47876 | 0.51918 | 1.00000 | 0.68551 | 0.72311 |
| 5 | 0.30023 | 0.31978 | 0.63056 | 0.33280 | 0.41502 |
| 6 | 0.50491 | 0.59198 | 0.69895 | 0.46376 | 0.59161 |
| 7 | 0.97958 | 1.00000 | 1.00000 | 1.00000 | 1.00000 |
| 8 | 0.49988 | 0.22669 | 0.42008 | 0.65347 | 0.51537 |
| 9 | 0.78207 | 0.82999 | 0.75074 | 0.95109 | 0.80821 |
| 10 | 0.64049 | 0.73777 | 1.00000 | 0.63739 | 0.88076 |
| 11 | 0.78011 | 0.83780 | 0.90663 | 1.00000 | 0.91042 |
| 12 | 0.59318 | 1.00000 | 0.83659 | 0.59033 | 0.84733 |
| 13 | 0.44862 | 0.59288 | 0.55396 | 0.54155 | 0.56901 |
| 14 | 0.54204 | 0.55415 | 0.60451 | 0.69851 | 0.64506 |
| 15 | 0.81013 | 0.81281 | 1.00000 | 0.81231 | 0.87945 |
| 16 | 0.85452 | 0.89494 | 0.96515 | 1.00000 | 0.94878 |
| 17 | 0.70308 | 0.84762 | 0.94450 | 1.00000 | 0.90334 |
| 18 | 0.55393 | 0.59430 | 0.92209 | 1.00000 | 0.86417 |
| 19 | 0.74080 | 1.00000 | 1.00000 | 0.65473 | 0.84137 |
| 20 | 0.78130 | 0.91408 | 1.00000 | 0.59135 | 0.79023 |

## References

Banker RA, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Man Sci 30:1078–1092

Charnes A, Cooper WW (1962) Programming with linear fractional functionals. Naval Res Logist Q 9:181–185

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2(6):429–444

Cook WD, Hababou M (2001) Sales performance measurement in bank branches. OMEGA 29:299–307

Cook WD, Imanirad R (2010) DEA in the presence of partial input to output impacts, Working Paper, Schulich School of Business, York University, Toronto, Canada

Cook WD, Hababou M, Tuenter H (2000) Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Prod Anal 14(3):209–224

Cook WD, Liang L, Zhu J (2010a) Measuring performance of two-stage network structures by DEA: a review and future perspective. OMEGA 38:423–430

Cook WD, Zhu J, Bi G, Yang F (2010b) Network DEA: additive efficiency decomposition. Eur J Oper Res 207(2):1122–1129

Färe R, Grosskopf S (1996) Productivity and intermediate products: a frontier approach. Econ Lett 50:65–70

Panzar JC, Willig RD (1981) Economies of scope. Am Econ Rev 71:286–272

Pulley LB, Braunstein YM (1992) A composite cost function for multiproduct firms with an application to economies of scope in banking. Rev Econ Stat 74:221–230

Thompson RG, Langemeir LN, Lee C, Lee E, Thrall RM (1990) The role of multiplier bounds in efficiency analysis with application to Kansas farming. J Econom 46:93–108

# Chapter 14
# Super-Efficiency in Data Envelopment Analysis

**Yao Chen and Juan Du**

**Abstract** In an effort to discriminate the performance of efficient decision making units (DMUs), the concept of super-efficiency is proposed, whose basic idea is to eliminate the DMU under evaluation from the reference set. When applied to the variable returns to scale (VRS) situation, the resulting super-efficiency model may become infeasible for certain DMUs due to the convexity constraint. Infeasibility restricts a wider use of super-efficiency DEA. Therefore, taking different viewpoints, a significant amount of studies tackle this problem by developing various new VRS super-efficiency models.

**Keywords** Data envelopment analysis (DEA) · Infeasibility · Super-efficiency · Variable returns to scale (VRS)

## 14.1 Introduction

In an effort to differentiate and rank the performance of efficient decision-making units (DMUs), Andersen and Petersen (1993) propose the concept of super-efficiency and develop a super-efficiency model based on constant returns to scale (CRS). The basic idea is to eliminate the DMU under evaluation from the reference set of the

J. Du (✉)
School of Economics and Management, Tongji University,
1239 Siping Road, 200092 Shanghai, P.R. China
e-mail: dujuan@tongji.edu.cn

Y. Chen
International Center for Auditing and Evaluation, Nanjing Audit University,
211815 Nanjing, P.R. China
e-mail: yao_chen@uml.edu

Manning School of Business, University of Massachusetts at Lowell,
01845 Lowell, MA, USA

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations
Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_14

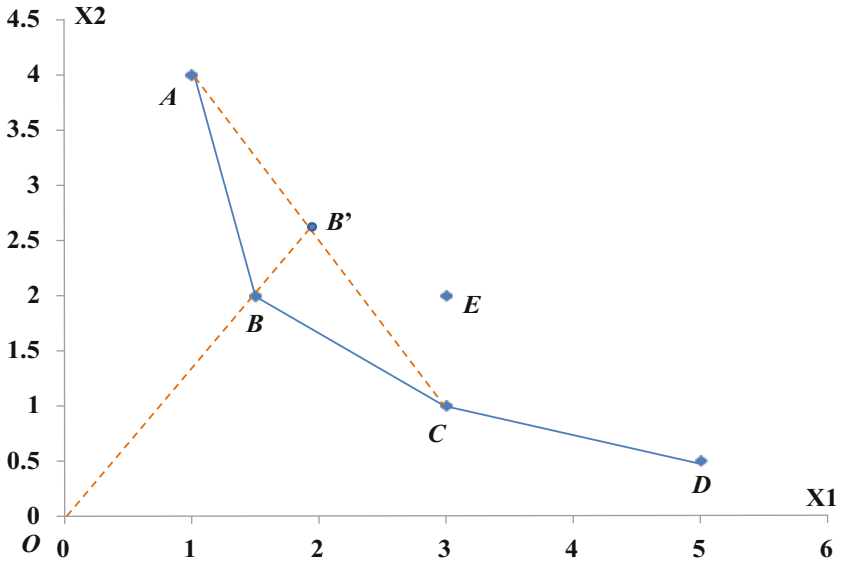



**Fig. 14.1** An illustration of radial CRS super-efficiency

envelopment models. The CRS super-efficiency model can be expressed as

$$\begin{aligned}
\min \quad & \theta - \varepsilon \left( \sum_{i=1}^{m} s_{io}^{-} + \sum_{r=1}^{s} s_{ro}^{+} \right) \\
s.\,t. \quad & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} + s_{io}^{-} = \theta x_{io}, i = 1,2,...,m \\
& \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j y_{rj} - s_{ro}^{+} = y_{ro}, r = 1,2,...,s \\
& \lambda_j, s_{io}^{-}, s_{ro}^{+} \geq 0, j = 1,2,...,n, j \neq o, i = 1,2,...,m, r = 1,2,...,s
\end{aligned} \tag{14.1}$$

where $\varepsilon > 0$ is the non-Archimedean infinitesimal.

Model (14.1) is commonly referred to as a "radial CRS super-efficiency", and allows for an efficiency score greater than one. An efficient DMU is projected onto the frontier constructed by the remaining DMUs and obtains an efficiency score no less than one. Solutions to model (14.1) always exist as long as all input and output elements are positive, i.e., $x_{io}, y_{ro} > 0$ for all $i$ and $r$.

We illustrate the above super-efficiency concept using Fig. 14.1, where there are five DMUs (DMU $A$, $B$, $C$, $D$, $E$) with two inputs and one equal output.

In Fig. 14.1, the original efficient frontier is composed of line segments connecting $A$, $B$, $C$, and $D$. Thus those four points are efficient DMUs, while $E$ is an

inefficient one. If Andersen and Petersen's (1993) model (14.1) is used to make further differentiation, on DMU $B$ for example, $B$ is omitted from the DEA reference set (the left-hand-side of model (14.1) and the new efficient frontier is made up by the line segments connecting $A$, $C$, and $D$. $B$' on line segment $AC$ is the projected point of DMU $B$ on the new frontier, and its super-efficiency defined by model (14.1) is calculated as $OB'/OB$, which is greater than 1. On the other hand, excluding the inefficient DMU $E$ from the reference set has no impact on its efficiency assessment. Thus for DMU $E$, its super-efficiency measure is exactly the same with its standard efficiency score.

Note that Andersen and Petersen's (1993) model development is under the CRS assumption. When the concept of super-efficiency is applied to the variable returns to scale (VRS) case, the resulting model may become infeasible for certain DMUs due to the convexity constraint (Seiford and Zhu 1999). By adding the convexity constraint $\sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n$ to CRS model (14.1), we obtain its VRS version as

$$\begin{aligned} \min \quad & \theta \\ s.t. \quad & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} \leq \theta x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j y_{rj} \geq y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.2}$$

Model (14.2) becomes infeasible if at least for one output $r_o$, all possible convex combinations of this output of the remaining DMUs are less than that output of the evaluated DMU, i.e., $y_{r_o o}$ (see Seiford and Zhu (1999)). As we will discuss later in detail, most of the super-efficiency studies focus on addressing this infeasibility issue in a VRS situation by developing various new super-efficiency models.

## 14.2 Infeasibility

Under the assumption of constant returns to scale (CRS), Zhu (1996) shows that the super-efficiency model becomes infeasible if and only if certain zero patterns appear in the data set. However, when the concept of super-efficiency is applied to the variable returns to scale (VRS) situation, the resulting DEA model must be infeasible for certain DMUs. Seiford and Zhu (1999) investigate necessary and sufficient conditions for infeasibility of super-efficiency DEA models. They can further identify

the position of the DMU under evaluation when infeasibility occurs, based upon the returns to scale (RTS) classifications.

Seiford and Zhu (1999) begin the analysis with presenting the super-efficiency (SE) DEA models under various returns to scale.

*Output-based*

$$\begin{aligned} \max \quad & \varphi \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} \leq x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} \geq \varphi y_{ro}, r = 1,2,...,s \\ & \varphi, \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned}$$

*Input-based*

$$\begin{aligned} \min \quad & \rho \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} \leq \rho x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} \geq y_{ro}, r = 1,2,...,s \\ & \rho, \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned}$$

(14.3)

*For SE-CRS append nothing.*

*For SE-VRS append* $\sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1.$

*For SE-NIRS append* $\sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j \leq 1.$

*For SE-NDRS append* $\sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j \geq 1.$

As pointed out by Charnes et al. (1991), the DMUs can be into four types $E$, $E'$, $F$ and $N$ as follows. $E$ is the set of extreme efficient DMUs, and $E'$ is the set of efficient DMUs that are not extreme points. The DMUs in set $E'$ can be expressed as linear combinations of the DMUs in set $E$. The DMUs in set $F$ is frontier points with non-zero slacks, and are usually referred to as weakly efficient. Finally, $N$ is the set of inefficient DMUs. Based on the above classifications, if a specific $DMU_o$ belongs to any of $E'$, $F$ or $N$ and is eliminated from the reference set, the efficient frontiers (constructed by the DMUs in set $E$) remain unchanged. Therefore, the super-efficiency models are feasible and equivalent to the original DEA models when $DMU_o \in E', For N$. Thus the infeasibility of super-efficiency only occurs for DMUs in set $E$.

Thrall (1996) shows that if the SE-CRS model (or super-efficiency CRS model) is infeasible, then $DMU_o \in E$. But he fails to recognize that the output-oriented

SE-CRS model is always feasible for the trivial solution which sets all variables equal to zero. Zhu (1996) showed that the input-oriented SE-CRS model is infeasible if and only if a certain pattern of zero data occurs in inputs and output metrics. One example is $DMU_o$ has a zero value in some inputs while all the other DMUs have positive data in those inputs, or similarly, $DMU_o$ has positive values in some outputs while all the other DMUs are set zero to those outputs. Later Seiford and Zhu (1999) further discuss the infeasibility of other super-efficiency DEA models with positive data, by first proposing two propositions.

**Proposition 2.1** $DMU_o \in E$ under the VRS model if and only if $DMU_o \in E$ under the NIRS model or NDRS model.

**Proposition 2.2** Let $\varphi^*$ and $\rho^*$ represent, respectively, the optimal objectives of the output-based and input-based super-efficiency DEA models when evaluating an extreme efficient DMU, then

a. Either $\varphi^* < 1$ or the specific output-based super-efficiency DEA model is infeasible;
b. Either $\rho^* > 1$ or the specific input-based super-efficiency DEA model is infeasible.

Based on the above propositions, Seiford and Zhu (1999) investigate the necessary and sufficient conditions for the infeasibility of various super-efficiency models.

For output-based SE-VRS model, they propose the following Theorems 2.1–2.5.

**Theorem 2.1** For a specific extreme efficient $DMU_o = (x_o, y_o)$, the output-based SE-VRS model is infeasible if and only if $(x_o, \delta y_o)$ is efficient under the original VRS model for any $0 < \delta \leq 1$.

**Theorem 2.2** The output-based SE-VRS model is infeasible if and only if $h^* > 1$, where $h^*$ is the optimal value to (14.4).

$$\begin{aligned} h^* = \ & \min \ h \\ s.t. \ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} \leq h x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.4}$$

**Theorem 2.3** If the output-based SE-VRS model is infeasible, then the DMU under evaluation exhibits IRS or CRS.

**Theorem 2.4** The output-based SE-NIRS model is always feasible.

**Theorem 2.5** For a specific extreme efficient $DMU_o = (x_o, y_o)$,

a. The output-based SE-NDRS model is infeasible if and only if $(x_o, \delta y_o)$ is efficient under the original VRS model for any $0 < \delta \leq 1$;

b. The output-based SE-NDRS model is infeasible if and only if $h^* > 1$, where $h^*$ is the optimal value to (14.4).

In a similar way, Seiford and Zhu (1999) explore the infeasibility issue of input-based super-efficiency DEA models by presenting Theorems 2.6–2.10.

**Theorem 2.6** For a specific extreme efficient $DMU_o = (x_o, y_o)$, the input-based SE-VRS model is infeasible if and only if $(\chi x_o, y_o)$ is efficient under the original VRS model for any $1 \leq \chi < +\infty$.

**Theorem 2.7** The input-based SE-VRS model is infeasible if and only if $g^* < 1$, where $g^*$ is the optimal value to (14.5).

$$\begin{aligned} g^* = \max\ & g \\ s.t.\quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} \geq g y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.5}$$

**Theorem 2.8** If the input-based SE-VRS model is infeasible, then the DMU under evaluation exhibits DRS or CRS.

**Theorem 2.9** The input-based SE-NDRS model is always feasible.

**Theorem 2.10** For a specific extreme efficient $DMU_o = (x_o, y_o)$,

a. The input-based SE-NIRS model is infeasible if and only if $(\chi x_o, y_o)$ is efficient under the original VRS model for any $1 \leq \chi < +\infty$;
b. The input-based SE-NIRS model is infeasible if and only if $g^* < 1$, where $g^*$ is the optimal value to (14.5).

Since IRS and DRS are not allowed in the NIRS and NDRS models respectively, the following corollary is proposed in Seiford and Zhu (1999).

**Corollary 2.1**

a. If $DMU_o \in E$ exhibits DRS, then all output-based super-efficiency DEA models are feasible;
b. If $DMU_o \in E$ exhibits IRS, then all input-based super-efficiency DEA models are feasible.

The above necessary and sufficient conditions for infeasibility provided by Seiford and Zhu (1999) indicate that the use of the super-efficiency DEA models should be restricted in some situations and the super-efficiency VRS models could be used to estimate RTS.

## 14.3 Alternative VRS Super-Efficiency Models

Infeasibility restricts a wider use of super-efficiency DEA. Recent years have seen a significant amount studies tackling this problem by developing various new VRS super-efficiency models.

### *14.3.1 Equivalent Standard Super-Efficiency Models (Lovell and Rouse 2003)*

Lovell and Rouse (2003) modify the standard radial super-efficiency models by scaling up the concerned input vector for input-orientation, or by scaling down the concerned output vector for output-orientation. Their idea can be presented in the following model (14.6), where inputs for each efficient DMU are multiplied by a scalar $\alpha > 1$ sufficiently large to make the DMU inefficient via model (14.6), with an optimal objective $\theta_2^* < 1$.

$$\begin{aligned} \min \quad & \theta_2 \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} + \alpha x_{io}\lambda_o \leq \alpha x_{io}\theta_2, i = 1,2,...,m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{ro}, r = 1,2,...,s \\ & \sum_{j=1}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n \end{aligned} \tag{14.6}$$

As mentioned before, extreme efficient DMUs may have no feasible solution when evaluated via the standard VRS super-efficiency program (14.2). However, as proved by Lovell and Rouse (2003), their modified model (14.6) is guaranteed to generate feasible solutions for all DMUs. Furthermore, they demonstrate that their model (14.6) and the standard VRS super-efficiency model are equivalent in providing the same optimal solutions for those DMUs that are feasible under the latter.

Lovell and Rouse (2003) also develop the output-oriented version of model (14.6), where outputs of each efficient DMU are multiplied by a scalar $0 < \beta < 1$ sufficiently small to make the DMU inefficient via model (14.7), with an optimal objective $\varphi_2^{*-1} < 1$.

$$\begin{aligned} \max \quad & \varphi_2 \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \le x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j \ne o}}^{n} \lambda_j y_{rj} + \beta y_{ro} \lambda_o \ge \beta y_{ro} \varphi_2, r = 1,2,...,s \\ & \sum_{j=1}^{n} \lambda_j = 1 \\ & \lambda_j \ge 0, j = 1,2,...,n \end{aligned} \tag{14.7}$$

The scaling parameters $\alpha$ and $\beta$ are specified by Lovell and Rouse (2003) as $\alpha = \max(\alpha_1, \ldots, \alpha_m) + 1$, where $\alpha_i = \max_j x_{ij} / \min_j x_{ij}$ and $\min_j x_{ij}$ are selected to be positive, while $\beta = \{\max(\beta_1, \ldots, \beta_s)\}^{-1}$, where $\beta_r = \max_j y_{rj} / \min_j y_{rj} + 1$ and $\min_j y_{rj}$ are selected to be positive.

### 14.3.2 *Super-Efficiency Based on Efficient Projections (Chen 2004, 2005)*

In order to overcome the infeasibility problem with respect to the conventional VRS super-efficiency, Chen (2004, 2005) suggests characterizing the super-efficiency through both input- and output-oriented super-efficiency models with input savings and output surplus. The basic idea of her methods is to replace the original inefficient observations with the efficient projections, and then the super-efficiency analysis is performed on this revised data set.

According to Chen (2004, 2005), the super-efficiency in terms of input savings is characterized via model (14.8).

$$\begin{aligned} \min \quad & \tilde{\theta}_o^{VRS-\text{super}} \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} \leq \tilde{\theta}_o^{VRS-\text{super}} x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j \hat{y}_{rj} \geq \hat{y}_{ro} = y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.8}$$

where $\hat{y}_{rj} = \phi_j^* y_{rj}$ and $\phi_j^*$ is the optimal objective to the following output-oriented VRS DEA model (14.9).

$$\begin{aligned} \phi_o^* = \max \quad & \phi_o \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{io}, i = 1,2,\ldots,m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} \geq \phi_o y_{ro}, r = 1,2,\ldots,s \\ & \sum_{j=1}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,\ldots,n \end{aligned} \tag{14.9}$$

Applying model (14.8) is equivalent to applying the standard VRS super-efficiency model (14.2) after all inefficient DMUs are projected onto the VRS frontier through proportional output augmentation by model (14.9). Model (14.8) measures the possible input saving achieved by the evaluated DMU against all other DMUs' input levels.

Chen (2004, 2005) pointed out that model (14.8) is still possible to be infeasible, which indicates that the evaluated DMU has the greatest input levels given the current output levels, thus cannot be moved onto the frontier formed by the remaining DMUs simply through input increases. In this case, let $\theta_o^{VRS-super^*} = \tilde{\theta}_o^{VRS-super^*} = 1$, implying for a zero input super-efficiency. Based on the above analysis, Chen (2004, 2005) uses $\gamma_o$ to represent the super-efficiency with respect to input savings.

$$\gamma_o = \begin{cases} \theta_o^{VRS-super^*}, \textit{if standard VRS super} - \textit{efficiency model (1.42) is feasible} \\ \tilde{\theta}_o^{VRS-super^*}, \textit{if model (14.2) is infeasible and model (14.8) is feasible} \\ 1, \qquad \textit{if model (14.8) is infeasible} \end{cases} \tag{14.10}$$

Chen (2004, 2005) then set up model (14.11) to describe the super-efficiency in terms of output surplus.

$$\begin{aligned} \max \quad & \tilde{\phi}_o^{VRS-super} \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j \hat{x}_{ij} \leq \hat{x}_{io} = x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} \geq \tilde{\phi}_o^{VRS-super} y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.11}$$

where $\hat{x}_{ij} = \theta_j^* x_{ij}$ and $\theta_j^*$ is the standard input-oriented VRS efficiency score for $DMU_j$.

Then $\tau_o$ is used to represent the super-efficiency with respect to output surplus.

$$\tau_o = \begin{cases} \phi_o^{VRS-super^*}, & \textit{if standard output} - \textit{oriented VRS super} \\ & - \textit{efficiency modelis feasible} \\ \tilde{\phi}_o^{VRS-super^*}, & \textit{if standard output} - \textit{oriented VRS super} \\ & - \textit{efficiency is infeasible} \\ & \textit{and model (14.11) is feasible} \\ 1, & \textit{if model (14.11) is infeasible} \end{cases} \tag{14.12}$$

As indicated by Chen (2004, 2005), infeasibility occurs when only input saving or output surplus is used to characterize super-efficiency. Therefore, input super-efficiency $\gamma_o$ and output super-efficiency $\tau_o$ should be integrated into one super-efficiency measure. One example suggested by Chen (2004, 2005) is to select $w_\gamma$ and $w_\tau$ such that $w_\gamma + w_\tau = 1$ and to define $S_o = w_\gamma \gamma_o + w_\tau \frac{1}{\tau_o}$ or $\hat{S}_o = w_\gamma \frac{1}{\gamma_o} + w_\tau \tau_o$. It is obvious that $S_o \geq 1$ and $\hat{S}_o \leq 1$, and a greater $S_o$ or a smaller $\hat{S}_o$ implies a better super-efficiency performance.

### 14.3.3 A Modified Super-Efficiency Measure (Cook et al. 2009)

Cook et al. (2009) suggest an alternative approach to address infeasibility in the VRS super-efficiency, which provides equivalent super-efficiency scores to those obtained from the standard VRS super-efficiency model (14.2) for feasible DMUs. When the infeasibility occurs, their approach determines a (virtual) reference DMU formed by the remaining DMUs and produces a score characterizing the super-efficient status. For an efficient $DMU_o$, consider the following model (14.13), where $M$ is a user-specified large positive number.

$$\begin{aligned} \min \quad & \tau + M \times \beta \\ s.t. \quad & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} \leq (1+\tau)\, x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j y_{rj} \geq (1-\beta)\, y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j = 1 \\ & \beta, \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.13}$$

Unlike the standard super-efficiency models which require a specific orientation, the approach proposed by Cook et al. (2009) moves the efficient DMU under evaluation onto the frontier by way of projection in both input and output directions. Thus their model illustrates the minimum movement of the concerned efficient DMU in both directions needed to reach the frontier constructed by the remaining DMUs.

Cook et al. (2009) relate the standard VRS super-efficiency model (14.2) with their modified model (14.13) via Theorem 3.1.

**Theorem 3.1** Model (1.2) is infeasible if and only if $\beta^* > 0$, where $\beta^*$ is the optimal solution in model (14.13).

Theorem 3.1 implies that model (14.2) is feasible if and only if $\beta^* = 0$, and further $1 + \tau^* = \theta^*$, where (*) denotes the optimal values in both models. This indicates that when standard VRS super-efficiency model (14.2) is feasible, Cook et al.'s (2009) model (14.13) is equivalent to (14.2) in that the objective values of both models are identical.

Then a theorem concerning the optimal solution to model (14.13) is presented as

**Theorem 3.2** $1 > \beta^* \geq 0$ and $\tau^* > -1$ in model (14.13).

Theorem 3.2 shows that when infeasibility occurs to standard model (14.2), $1/(1-\beta^*) > 1$ and $1 + \tau^* > 0$, indicating that to have a feasible solution, $DMU_o$ must decrease its outputs. Cook et al. (2009) further define the super-efficiency score

as $1 + \tau^* + 1/(1 - \beta^*)$, which consists of a component for input super-efficiency $1 + \tau^* > 0$, and a component for output super-efficiency $1/(1 - \beta^*) > 1$.

### 14.3.4 Two-Stage Procedure (Lee et al. 2011) and Its One Model Approach (Chen and Liang 2011)

Lee et al. (2011) extend the basic ideas in Chen (2005) and Cook et al. (2009), and present a two-stage process for calculating super-efficiency scores regardless of the feasibility of the standard VRS super-efficiency. When the standard VRS model (14.2) is feasible, their approach yields identical super-efficiency scores to those obtained from model (14.2). While for efficient DMUs which have no solutions via conventional method, their approach produces a super-efficiency measure characterizing input savings and output surplus.

Lee et al. (2011) have pointed out that, in an input-oriented case, the infeasibility of super-efficiency occurs when outputs of the DMU under evaluation is outside the production possibility set generated by the outputs of the remaining DMUs. In an output-oriented case, this infeasibility occurs when inputs of the concerned DMU is outside the production possibility set formed by the inputs of the remaining DMUs.

For the input-oriented VRS super-efficiency model (14.2), if an efficient $DMU_o$ is infeasible, it may be caused by the fact that $DMU_o$ does not exhibit input savings but only output surplus, which characterizes the super-efficiency (Seiford and Zhu 1999; Chen 2005). Thus Lee et al. (2011) develop a linear programming problem (14.14), which seeks to determine potential surplus in each individual output.

$$\begin{aligned} \min \quad & \sum_{r=1}^{s} s_r \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} + s_r y_{ro} \geq y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & s_r, \lambda_j \geq 0, j = 1,2,...,n, j \neq o, r = 1,2,...,s \end{aligned} \tag{14.14}$$

Let $(s_1^*, \ldots, s_s^*)$ denote an optimal solution to model (14.14). Lee et al. (2011) present Theorem 3.3.

**Theorem 3.3** Standard VRS super-efficiency model (14.2) is feasible if and only if $s_r^* = 0$ for all $r$.

Theorem 3.3 indicates that the input-oriented VRS super-efficiency model is infeasible if and only if there are some $s_r^* > 0$.

Lee et al. (2011) then propose their modified unit-invariant VRS super-efficiency model (14.15).

$$\begin{aligned} \min \quad & \hat{\theta} \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} \leq \hat{\theta} x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} + s_r^* y_{ro} \geq y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.15}$$

where $(s_1^*, \ldots, s_s^*)$ is an optimal solution to model (14.14).

Let $\hat{\theta}^*$ and $\theta^*$ be the optimal objective of model (14.15) and standard model (14.2), respectively. Then for feasible DMUs, there is $\hat{\theta}^* = \theta^*$, implying that both models yield the identical super-efficiency score. Then by showing that projection or benchmark for the evaluated DMU is constantly on the frontier formed by the remaining DMUs, Lee et al. (2011) prove that model (14.15) is always feasible. They further modify the super-efficiency score obtained from model (14.15) and define the composite super-efficiency measure as

$$\breve{\theta} = \begin{cases} \dfrac{\sum_{r\in R} \frac{y_{ro}}{y_{ro} - s_r^* y_{ro}}}{|R|} + \hat{\theta}^*, if\ R \neq \phi \\ \hat{\theta}^*, \qquad\qquad\qquad if\ R = \phi \end{cases} \tag{14.16}$$

where $R = \left\{ r \,|\, s_r^* > 0 \right\}$ based upon model (14.13) and $|R|$ is the cardinality of the set $R$.

As for this composite super-efficiency, Lee et al. (2011) demonstrate that $\breve{\theta} = \hat{\theta}^*$ if standard VRS super-efficiency is feasible and $\breve{\theta} > 1$ if standard VRS super-efficiency is infeasible.

For the output-oriented case, Lee et al. (2011) propose a similar method by first solving the linear programming problem (14.17), which seeks to determine potential input savings of the concerned efficient $DMU_o$ compared against the frontier

generated by all other DMUs.

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} t_i \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} - t_i x_{io} \leq x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & t_i, \lambda_j \geq 0, j = 1,2,...,n, j \neq o, i = 1,2,...,m \end{aligned} \tag{14.17}$$

Let $(t_1^*, \ldots, t_m^*)$ denote an optimal solution to model (14.17). Lee et al. (2011) further establish their modified output-oriented VRS super-efficiency model (14.18), which is proved to be always feasible.

$$\begin{aligned} \max \quad & \hat{\beta} \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j x_{ij} - t_i^* x_{io} \leq x_{io}, i = 1,2,...,m \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j y_{rj} \geq \hat{\beta} y_{ro}, r = 1,2,...,s \\ & \sum_{\substack{j=1 \\ j\neq o}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1,2,...,n, j \neq o \end{aligned} \tag{14.18}$$

Based on the optimal values of model (14.18), Lee et al. (2011) define the output-oriented composite super-efficiency measure as

$$\frac{1}{\bar{\beta}} = \begin{cases} \dfrac{\sum_{i\in I} \frac{x_{io} + t_i^* x_{io}}{x_{io}}}{|I|} + \dfrac{1}{\hat{\beta}^*}, if\, I \neq \phi \\ \dfrac{1}{\hat{\beta}^*}, \qquad\qquad if\, I = \phi \end{cases} \tag{14.19}$$

where $I = \left\{ i \,|t_i^* > 0 \right\}$ based upon model (14.17) and $|I|$ is the cardinality of the set $I$.

Based on the study in Cook et al. (2009), Chen and Liang (2011) demonstrate that the two-stage approach developed by Lee et al. (2011) can actually be solved

through the following equivalent linear programming (LP) model (14.20), where $M$ is a user-defined large positive number

**Input-orientation:**

$$\begin{aligned}
\min \quad & \tau + M \times \sum_{r=1}^{s} \beta_r \\
s.t. \quad & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} \leq (1+\tau)\, x_{io}, i = 1,2,...,m \\
& \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j y_{rj} \geq (1-\beta_r)\, y_{ro}, r = 1,2,...,s \\
& \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j = 1 \\
& \beta_r, \lambda_j \geq 0, j = 1,2,...,n, j \neq o, r = 1,2,...,s
\end{aligned} \tag{14.20}$$

Chen and Liang (2011) have pointed out that the conventional VRS super-efficiency model (14.2) is infeasible if and only if some $\beta_r^* > 0$, and $\beta_r^* = s_r^*$ where $\beta_r^*$ and $s_r^*$ are optimal values to the above model (14.20) and Lee et al.'s (2011) model (14.14), respectively. They further indicate that the two-stage procedure proposed in Lee et al. (2011) is equivalent to their model (14.20), thus Lee et al.'s (2011) approach can be equivalently solved through on single model (14.20).

Based on the optimal solution to model (14.20), the super-efficiency score can be defined as $1 + \tau^* + \frac{1}{s}\sum_{r=1}^{s} \frac{1}{1-\beta_r^*}$ according to Cook et al. (2009), or $1 + \tau^* + \frac{1}{|R|}\sum_{r \in R} \frac{1}{1-\beta_r^*}$ where $R$ is the set of $\beta_r^* > 0$, according to Lee et al. (2011).

Finally, the output-oriented version of model (14.20) is developed as

$$\begin{aligned}
\min \quad & \gamma + M \times \sum_{i=1}^{m} \delta_i \\
s.t. \quad & \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j x_{ij} \leq (1+\delta_i)\, x_{io}, i = 1,2,...,m \\
& \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j y_{rj} \geq (1-\gamma)\, y_{ro}, r = 1,2,...,s \\
& \sum_{\substack{j=1 \\ j \neq o}}^{n} \lambda_j = 1 \\
& \delta_i, \lambda_j \geq 0, j = 1,2,...,n, j \neq o, i = 1,2,...,m
\end{aligned} \tag{14.21}$$

**Table 14.1** Data for a numerical example

| DMU | Input 1 | Input 2 | Output | VRS efficiency |
|---|---|---|---|---|
| 1 | 2 | 4 | 2 | 1 |
| 2 | 1.5 | 2 | 1 | 1 |
| 3 | 4 | 1 | 3 | 1 |
| 4 | 5 | 2 | 4 | 1 |
| 5 | 3 | 2 | 1 | 0.8125 |

### *14.3.5 DDF-Based Super-Efficiency and SBM Super-Efficiency*

Based on the directional distance function (DDF) (Chambers et al. 1996), Ray (2008) introduces the VRS Nerlove-Luenberger (N-L) measure of super-efficiency that adjusts both input and output levels at the same proportion. Although this N-L super-efficiency model does not pose a similar infeasibility problem with the conventional VRS method, it fails in two special situations (Ray 2008). To tackle the two exceptions, Chen et al. (2013a) select a different input-output bundle to construct a new DDF from the one used in Ray (2008), based on which a modified VRS super-efficiency is proposed. These DDF-based super-efficiency measures will be discussed in detail later in Sect. 14.5.

Another typical group of super-efficiency measures are developed by dealing directly with input and output slacks. These non-radial measures include SBM super-efficiency (Tone 2002) and additive super-efficiency (Du et al. 2010), and will be further introduced in Sect. 14.4.

### *14.3.6 A Numerical Example for Comparison*

Consider a numerical example presented in Table 14.1, which consists of five DMUs with two inputs and one output. Among the five DMUs, only DMU 5 is VRS inefficient with an efficiency score of 0.8125, while the other four (DMUs 1, 2, 3, and 4) are all efficient. If the standard input-oriented VRS super-efficiency model (14.2) is applied, DMU 4 has no feasible solutions. If its output-oriented version is used, DMUs 2 and 3 become infeasible. The super-efficiency results from the standard models are listed in columns 2 and 3 of Table 14.2, respectively.

Then those alternative approaches introduced from Sects. 14.3.1–14.3.4 are used to address this infeasibility problem. Their results are demonstrated from columns 4 to 7, respectively. Note that the super-efficiency scores obtained from various measures can be very different. For example, DMU 4 is infeasible according to the standard input-oriented model (14.2), but is given the highest super-efficiency score via both the equivalent standard model (Lovell and Rouse 2003) and the two-stage procedure (Lee et al. 2011) or its one-model equivalence (Chen and Liang 2011).

**Table 14.2** Super-efficiency results

| DMU | Input-based standard | Output-based standard | Equivalent standard (Lovell and Rouse 2003) | Efficient projections (Chen 2004, 2005) | Modified (Cook et al. 2009) | Two-stage procedure (Lee et al. 2011) and its equivalence (Chen and Liang 2011) |
|---|---|---|---|---|---|---|
| 1 | 1.3333 | 0.7143 | 1.3333 | 1.3667 | 2.3333 | 1.3333 |
| 2 | 1.6 | Infeasible | 1.6 | 1.3 | 2.6 | 1.6 |
| 3 | 2 | Infeasible | 2 | 1.5 | 3 | 2 |
| 4 | Infeasible | 0.75 | 5 | 1.1667 | 2.1333 | 2.1333 |

However, DMU 4 is evaluated with the lowest super-efficiency based on Chen's (2004, 2005) projection models and Cook et al.'s (2009) modified model.

## 14.4 Slacks-Based Super-Efficiency

### *14.4.1 SBM Super-Efficiency*

Tone (2001) introduces a slacks-based measure (SBM) of non-radial efficiency by directly dealing with input and output slacks. As in the radial DEA model, it provides with an efficiency score between zero and one, and returns unity if and only if the DMU under evaluation is on the frontier of the production possibility set with no input or output slacks.

Based on the SBM-efficiency definition (Tone 2001, 2002) further presents a SBM super-efficiency measure to differentiate those SBM-efficient DMUs. For radial DEA models, super-efficiency models are obtained simply by removing the DMU concerned from the reference set, as in Andersen and Petersen (1993). However, this practice cannot be directly applied to non-radial models such as the additive DEA model (Charnes et al. 1982) or the SBM model (Tone 2001). As indicated in Tone (2002), for non-radial or slacks based DEA models, efficient DMUs need to be identified first to modify the relevant models.

Suppose that $DMU_o$ is SBM-efficient. Its SBM super-efficiency is calculated as the optimal objective function of the following problem (Tone 2002):

$$\begin{aligned}
\delta_o^* = \min \quad & \delta_o = \frac{\frac{1}{m}\sum_{i=1}^{m} \bar{x}_{io}/x_{io}}{\frac{1}{s}\sum_{r=1}^{s} \bar{y}_{ro}/y_{ro}} \\
s.t. \quad & \bar{x}_{io} \geq \sum_{j=1, j\neq o}^{n} \lambda_j x_{ij}, i = 1,2,\ldots,m \\
& \bar{y}_{ro} \leq \sum_{j=1, j\neq o}^{n} \lambda_j y_{rj}, r = 1,2,\ldots,s \\
& \bar{x}_{io} \geq x_{io}, i = 1,2,\ldots,m \\
& \bar{y}_{ro} \leq y_{ro}, r = 1,2,\ldots,s \\
& \lambda_j, \bar{y}_{ro} \geq 0, j = 1,2,\ldots,n, j \neq o, r = 1,2,\ldots,s
\end{aligned} \tag{14.22}$$

The form of the objective function requires positive input and output values for SBM-efficient DMUs, i.e., $x_{ij} > 0$ and $y_{rj} > 0$. Using the Charnes-Cooper transformation (Charnes and Cooper 1962), fractional model (14.22) can be equivalently converted into a linear program.

Two propositions are presented with respect to SBM super-efficiency model (14.22) (Tone 2002).

**Proposition 4.1** The SBM super-efficiency score is unit-invariant, i.e., it is independent of the units in which inputs and outputs are measured as longas these units are the same for every DMU.

**Proposition 4.2** Let $(\alpha x_{io}, i = 1,2,\ldots,m; \beta y_{ro}, r = 1,2,\ldots,s)$ with $0 < \alpha \leq 1$ and $\beta \geq 1$ be a DMU with reduced inputs and enlarged outputs than $(x_{io}, i = 1,2,\ldots,m; y_{ro}, r = 1,2,\ldots,s)$. Then the SBM super-efficiency score of $(\alpha x_{io}, \beta y_{ro})$ is not less than that of $(x_{io}, y_{ro})$.

### *14.4.2 Additive Super-Efficiency*

Du et al. (2010) extends the above SBM super-efficiency to the additive DEA model (Charnes et al. 1982). Alternative slacks-based objective functions are used. Unlike the traditional radial super-efficiency DEA, additive super-efficiency models proposed by Du et al. (2010) are consistently feasible under either constant or variable returns to scale.

Suppose $DMU_o$ is efficient via the additive model (14.23) (Charnes et al. 1982), which is equivalent to all zero slacks.

$$\begin{aligned} \max \quad & \sum_{i=1}^{m} s_{io}^{-} + \sum_{r=1}^{s} s_{ro}^{+} \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} + s_{io}^{-} = x_{io}, i = 1,2,\ldots,m \\ & \sum_{j=1}^{n} \lambda_j y_{rj} - s_{ro}^{+} = y_{ro}, r = 1,2,\ldots,s \\ & \lambda_j, s_{io}^{-}, s_{ro}^{+} \geq 0, j = 1,2,\ldots,n, i = 1,2,\ldots,m, r = 1,2,\ldots,s \end{aligned} \tag{14.23}$$

To obtain the super-efficiency of $DMU_o$, Du et al. (2010) propose two additive super-efficiency models (14.24) and (14.25) with different objectives.

$$\begin{aligned} \alpha_o^* = \min \quad & \alpha_o = \sum_{i=1}^{m} t_{io}^{-} + \sum_{r=1}^{s} t_{ro}^{+} \\ s.t. \quad & \sum_{j=1, j\neq o}^{n} \lambda_j x_{ij} \leq x_{io} + t_{io}^{-}, i = 1,2,\ldots,m \\ & \sum_{j=1, j\neq o}^{n} \lambda_j y_{rj} \geq y_{ro} - t_{ro}^{+}, r = 1,2,\ldots,s \\ & \lambda_j, t_{io}^{-}, t_{ro}^{+} \geq 0, j = 1,2,\ldots,n, j \neq o, i = 1,2,\ldots,m, r = 1,2,\ldots,s \end{aligned} \tag{14.24}$$

$$\begin{aligned} \beta_o^* = \min \quad & \beta_o = \frac{1}{m+s}\left(\sum_{i=1}^{m} \frac{t_{io}^{-}}{x_{io}} + \sum_{r=1}^{s} \frac{t_{ro}^{+}}{y_{ro}}\right) \\ s.t. \quad & x_{io} + t_{io}^{-} \geq \sum_{j=1, j\neq o}^{n} \lambda_j x_{ij}, i = 1,2,\ldots,m \\ & y_{ro} - t_{ro}^{+} \leq \sum_{j=1, j\neq o}^{n} \lambda_j y_{rj}, r = 1,2,\ldots,s \\ & \lambda_j, t_{io}^{-}, t_{ro}^{+} \geq 0, j = 1,2,\ldots,n, j \neq o, i = 1,2,\ldots,m, r = 1,2,\ldots,s \end{aligned} \tag{14.25}$$

After the evaluated $DMU_o$ is removed from the reference set, the constraints and objective function of model (14.23) are modified to get the super-efficiency models. The constraints should be modified because inputs or outputs are supposed to be increased or decreased to reach the frontier constructed by the remaining DMUs

(besides $DMU_o$). The objective is changed from maximization to minimization so that the resulting model is bounded.

Note that the constraints of model (14.24) or (14.25) can be equally converted into the constraints of model (14.22). In that sense, Tone's (2002) SBM model (14.22) is also a super-efficiency version of the additive DEA model (14.23).

Let $\left\{\alpha_o^*; \lambda_j^*(\alpha),\ j = 1, 2, \ldots, n,\ j \neq o; t_{io}^{-*}(\alpha),\ i = 1, 2, \ldots, m; t_{ro}^{+*}(\alpha),\ r = 1, 2, \ldots, s\right\}$ and $\left\{\beta_o^*; \lambda_j^*(\beta),\ j = 1, 2, \ldots, n,\ j \neq o; t_{io}^{-*}(\beta),\ i = 1, 2, \ldots, m; t_{ro}^{+*}(\beta),\ r = 1, 2, \ldots, s\right\}$ be an optimal solution to models (14.24) and (14.25), respectively. Du et al. (2010) use the same format taken by the objective function of Tone's (2002) SBM super-efficiency model (14.22), and define $\hat{\delta}_o^*(\alpha) = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_{io}+t_{io}^{-*}(\alpha))/x_{io}}{\frac{1}{s}\sum_{r=1}^{s}(y_{ro}-t_{ro}^{+*}(\alpha))/y_{ro}} \geq 1$ and $\hat{\delta}_o^*(\beta) = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_{io}+t_{io}^{-*}(\beta))/x_{io}}{\frac{1}{s}\sum_{r=1}^{s}(y_{ro}-t_{ro}^{+*}(\beta))/y_{ro}} \geq 1$ as the additive super-efficiency measures.

If the convexity constraint $\sum_{j=1,j\neq o}^{n}\lambda_j = 1$ is added into the above super-efficiency models, the VRS versions of SBM and additive super-efficiency measures are obtained. Unlike the radial VRS super-efficiency DEA models, all the slacks-based super-efficiency models (Tone 2002; Du et al. 2010) will not encounter the no-solution problem under either constant or variable returns to scale.

Similar to the Proposition 2 in Tone (2002), Du et al. (2010) present and prove their version of this proposition as

**Proposition 4.3** Let $(ax_{io}, i = 1,2,\ldots,m; by_{ro}, r = 1,2,\ldots,s)$ with $0 < a \leq 1$ and $b \geq 1$ be a DMU with reduced inputs and enlarged outputs than $(x_{io}, i = 1,2,\ldots,m; y_{ro}, r = 1,2,\ldots,s)$. Then the optimal objective from additive super-efficiency model (14.24) or (14.25) of $(ax_{io}, i = 1,2,\ldots,m; by_{ro}, r = 1,2,\ldots,s)$ is not less than that of $(x_{io}, i = 1,2,\ldots,m; y_{ro}, r = 1,2,\ldots,s)$.

### *14.4.3 A Numerical Example*

Table 14.3 presents data for seven DMUs with two inputs and one output (Tone 2002). DMUs $C$, $D$ and $E$ are efficient. The SBM super-efficiency scores of the three efficient DMUs via model (14.22) (Tone 2002) are displayed in column $\delta^*$. Table 14.4 reports the scores of $\hat{\delta}_o^*(\alpha)$ and $\hat{\delta}_o^*(\beta)$ defined by Du et al. (2010), along with their rankings for DMUs $C$, $D$ and $E$. It is noted that all DMUs have exactly the same rank according to three different super-efficiency models.

## 14.5 DDF-Based Super-Efficiency

Chambers et al. (1996) defined the directional distance function (DDF) as

$$D(x_{ik}, y_{rk}; g^x, g^y) = \max\beta : (x_{ik} + \beta g^x, y_{rk} + \beta g^y) \in T \tag{14.26}$$

**Table 14.3** Data and results from SBM super-efficiency model (Tone 2002)

| DMU | Data | | | SBM | Super-efficiency | | | |
|---|---|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $y_1$ | $\delta^*$ | Rank by $\delta^*$ | $\bar{x}_1^*$ | $\bar{x}_2^*$ | $\bar{y}_1^*$ |
| A | 4 | 3 | 1 | | 5 | | | |
| B | 7 | 3 | 1 | | 7 | | | |
| C | 8 | 1 | 1 | 1.125 | 3 | 10 | 1 | 1 |
| D | 4 | 2 | 1 | 1.25 | 2 | 4 | 3 | 1 |
| E | 2 | 4 | 1 | 1.5 | 1 | 4 | 4 | 1 |
| F | 10 | 1 | 1 | | 4 | | | |
| G | 12 | 1 | 1 | | 5 | | | |

**Table 14.4** Results from additive super-efficiency models (14.24) and (14.25)

| DMU | Model (14.24) | | | | | | Model (14.25) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha^*$ | $\hat{\delta}^*(\alpha)$ | Rank by $\hat{\delta}^*(\alpha)$ | $t_1^{-*}$ | $t_2^{-*}$ | $t_1^{+*}$ | $\beta^*$ | $\hat{\delta}^*(\beta)$ | Rank by $\hat{\delta}^*(\beta)$ | $t_1^{-*}$ | $t_2^{-*}$ | $t_1^{+*}$ |
| A | | | 5 | | | | | | 5 | | | |
| B | | | 7 | | | | | | 7 | | | |
| C | 0.125 | 1.1429 | 3 | 0 | 0 | 0.125 | 0.0417 | 1.1429 | 3 | 0 | 0 | 0.125 |
| D | 0.2 | 1.25 | 2 | 0 | 0 | 0.2 | 0.0667 | 1.25 | 2 | 0 | 0 | 0.2 |
| E | 0.5 | 2 | 1 | 0 | 0 | 0.5 | 0.1667 | 2 | 1 | 0 | 0 | 0.5 |
| F | | | 4 | | | | | | 4 | | | |
| G | | | 5 | | | | | | 5 | | | |

where $(g^x, g^y)$ is a reference input-output bundle, and $T$ represents the production possibility set (PPS) under the standard assumptions of convexity and free disposability, i.e.,

$$T = \left\{ (x_i, y_r) | \, x_i \geq \sum_{j=1}^{n} \lambda_j x_{ij}, i = 1, \ldots, m; y_r \right.$$
$$\left. \leq \sum_{j=1}^{n} \lambda_j y_{rj}, r = 1, \ldots, s; \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n \right\}.$$

### 14.5.1 N-L Super-Efficiency

For any $DMU_k(x_{ik}, y_{rk})$, Ray (2008) selects $(-x_{ik}, y_{rk})$ for $(g^x, g^y)$, and modifies the above PPS for super-efficiency as

$$T_k = \left\{ (x_i, y_r) \middle| \, x_i \geq \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij}, i = 1, ..., m; y_r \right. \\ \left. \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj}, r = 1, ..., s; \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, ..., n, j \neq k \right\} \tag{14.27}$$

Then the VRS Nerlove-Luenberger (N-L) measure of super-efficiency concerning PPS $T_k$ is developed by Ray (2008) as

$$\begin{aligned} \beta_k^* = \max \quad & \beta_k \\ s.t. \quad & \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij} \leq (1 - \beta_k) \, x_{ik}, i = 1, ..., m \\ & \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj} \geq (1 + \beta_k) \, y_{rk}, r = 1, ..., s \\ & \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1, ..., n, j \neq k \end{aligned} \tag{14.28}$$

A negative optimal value of $\beta_k$, or $\beta_k^*$, indicates that the output bundle of $DMU_k$ should be scaled down and its input bundle should be scaled up by the same proportion to get an attainable input-output mix in PPS $T_k$. The VRS N-L super-efficiency for $DMU_k$ under evaluation is determined as $(1 - \beta_k^*)$. A smaller $\beta_k^*$ implies for a more N-L super-efficient DMU.

By proportionally adjusting input and output levels in a single model, this N-L super-efficiency model eliminates a similar infeasibility problem as in the standard VRS super-efficiency model. Although in most cases feasible, as pointed out by Ray (2008), this VRS N-L super-efficiency model fails in two exceptions. First, the model becomes infeasible if at least one zero input is present in the DMU under evaluation while all other DMUs in the reference set have positive values in that input. In such a case, the first set of constraints in model (14.28) cannot be satisfied. Second, for some input $i_o$, there is $2x_{i_o k} < \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{i_o j}$ for all $\lambda_j$ combinations

satisfying $\sum_{\substack{j=1 \\ j\neq k}}^{n} \lambda_j = 1$ and $\lambda_j \geq 0, j = 1, \ldots, n, j \neq k$.Then $\beta_k$ is restricted to a value lower than $-1$, and the N-L super-efficiency score is greater than 2. The model will yield a reference point with negative output values.

As a matter of fact, zero data are problematic in any super-efficiency models, besides the DDF-based N-L measure. For example, Zhu (1996) shows that the super-efficiency CRS model is infeasible if and only if certain patters appear in the data set. Lee and Zhu (2012) claim that either the conventional or the modified VRS super-efficiency models, such as the two-stage procedure provided by Cook et al. (2009) with an attempt to address infeasibility, will become infeasible when zero data are present. In order to deal with the two exceptions of the N-L super-efficiency, especially the zero data problem, Chen et al. (2013) develop a modified VRS super-efficiency model based on a new DDF.

### *14.5.2 Modified DDF-Based Super-Efficiency*

To tackle the above two exceptions, Chen et al. (2013a) choose a different input-output bundle $(-ax_{io} - 1, by_{ro} + 1)$ for the DDF from the one used in Ray (2008), and construct a new DDF and a modified VRS super-efficiency based on this new DDF as (14.29) and (14.30), respectively.

$$D(x_{io}, y_{ro}) = \max \beta : ((1 - \beta a)x_{io} - \beta, (1 + \beta b)y_{ro} + \beta) \in T \tag{14.29}$$

$$\begin{aligned} \max \quad & \beta_k \\ s.t. \quad & \sum_{\substack{j=1 \\ j\neq k}}^{n} \lambda_j x_{ij} \leq (1 - \beta_k a)\, x_{ik} - \beta_k, i = 1, ..., m \\ & \sum_{\substack{j=1 \\ j\neq k}}^{n} \lambda_j y_{rj} \geq (1 + \beta_k b)\, y_{rk} + \beta_k, r = 1, ..., s \\ & \sum_{\substack{j=1 \\ j\neq k}}^{n} \lambda_j = 1 \\ & \lambda_j \geq 0, j = 1, ..., n, j \neq k \end{aligned} \tag{14.30}$$

where $a$ and $b$ are pre-determined positive parameters. Model (14.30) is feasible even if zero data exist in inputs. To eliminate the second infeasibility issue, Chen et al. (2013a) develop a procedure to select proper values for parameters $a$ and $b$ to prevent directional output targets $(1 + \beta_k b)y_{rk} + \beta_k$ from taking negative values. Specifically, in the completely positive input case, $a$ and $b$ should satisfy

$$a > \frac{\left(\max_{r=1,..,s} \max_{j=1,..,n} \frac{1}{y_{rj}}\right) \left[\max_{i=1,..,m} \left(\max_{j=1,..,n} x_{ij} - \min_{j=1,..,n} x_{ij}\right) + 1\right] - 1}{\min_{i=1,..,m} \min_{j=1,..,n} x_{ij}} \tag{14.31}$$

**Table 14.5** A numerical example (Seiford and Zhu 1999)

| DMU | Input 1 $x_1$ | Input 2 $x_2$ | Input 3 $x_3$ | Output 1 $y_1$ | Output 2 $y_2$ |
|---|---|---|---|---|---|
| 1 | 182 | 237 | 468 | 5008 | 5303 |
| 2 | 74 | 82 | 148 | 1857 | 2336 |
| 3 | 160 | 195 | 400 | 4041 | 5001 |
| 4 | 183 | 150 | 339 | 2779 | 2418 |
| 5 | 133 | 155 | 329 | 3506 | 3602 |
| 6 | 106 | 120 | 138 | 1306 | 956 |
| 7 | 109 | 110 | 188 | 1515 | 2282 |
| 8 | 240 | 243 | 806 | 7763 | 9601 |
| 9 | 276 | 188 | 574 | 4577 | 6493 |
| 10 | 191 | 117 | 466 | 3322 | 4233 |

$$0 < b \leq \frac{a \min_{i=1,..,m} \min_{j=1,..,n} x_{ij} + 1}{\max_{i=1,..,m} (\max_{j=1,..,n} x_{ij} - \min_{j=1,..,n} x_{ij}) + 1} - \max_{r=1,..,s} \max_{j=1,..,n} \frac{1}{y_{rj}} \quad (14.32)$$

While in the zero input case, $a$ can take any positive value, and $b$ is subjected to

$$0 < b \leq \frac{1}{\max_{i=1,..,m} (\max_{j=1,..,n} x_{ij} - \min_{j=1,..,n} x_{ij})} - \max_{r=1,..,s} \max_{j=1,..,n} \frac{1}{y_{rj}} \quad (14.33)$$

In either case, any value combination taken from the corresponding ranges is a reasonable candidate for parameters of the new DDF (14.29). Moreover, Chen et al. (2013a) provide a referable way to determine $a$ and $b$ if the integer-valued parameters are expected. Specifically, for the completely positive input case, the smallest integer satisfying (14.31) can be chosen as $a$ and the greatest integer satisfying (14.32) can be chosen as $b$; while for the zero input case, $a$ is set to 1 and $b$ is selected as the greatest positive integer less than $\frac{1}{\max_{i=1,..,m} (\max_{j=1,..,n} \bar{x}_{ij} - \min_{j=1,..,n} \bar{x}_{ij})} - \max_{r=1,..,s} \max_{j=1,..,n} \frac{1}{y_{rj}}$, where $\bar{x}_{ij}$ are proportionally scaled-down values of $x_{ij}$ to make $\frac{1}{\max_{i=1,..,m} (\max_{j=1,..,n} \bar{x}_{ij} - \min_{j=1,..,n} \bar{x}_{ij})} - \max_{r=1,..,s} \max_{j=1,..,n} \frac{1}{y_{rj}} > 1$.

Next a data set presented in Table 14.5, which was previously studied in Seiford and Zhu (1999) and Ray (2008), are used to demonstrate different results obtained from various VRS super-efficiency models introduced in this section.

Columns 2–4 in Table 14.6 report the results calculated from the standard VRS super-efficiency model, the N-L super-efficiency measure (14.28) (Ray 2008), and the modified DDF-based super-efficiency model (14.30) (Chen et al. 2013a), respectively.

The column identified as "Standard radial" presents the efficiency and super-efficiency scores obtained from the conventional input-oriented VRS super-efficiency model. DMUs 3, 4, 5, 7, 9 are inefficient units, while DMUs 1, 2, 6, 8, 10 are efficient. Among these five efficient units, four (DMUs 1, 2, 6, 10) have super-efficiency greater than one. DMU 8 does not have a feasible solution to the conventional VRS super-efficiency problem.

**Table 14.6** Alternative measures of VRS super-efficiency

| DMU | Standard radial | N-L | Modified DDF |
|---|---|---|---|
| 1 | 1.0626 | 1.0285 | 1.00559 |
| 2 | 1.5277 | 1.4430 | 1.05113 |
| 3 | 0.9765 | 0.9889 | 0.99789 |
| 4 | 0.7354 | 0.8566 | 0.97561 |
| 5 | 0.9752 | 0.9881 | 0.99776 |
| 6 | 1.0725 | 1.0725 | 1.00724 |
| 7 | 0.7852 | 0.8863 | 0.97871 |
| 8 | Infeasible | 1.3836 | 1.38354 |
| 9 | 0.9246 | 0.9581 | 0.99302 |
| 10 | 1.0602 | 1.0334 | 1.00557 |

Column "N-L" displays Ray's (2008) N-L measure of super-efficiency $(1 - \beta_k)$ for all ten DMUs, and column "Modified DDF" shows the results obtained from Chen et al.'s (2013a) modified DDF-based super-efficiency model, with parameters $a$ and $b$ pre-determined as 10 and 1. Similar to the N-L super-efficiency, for super-efficient DMUs 1, 2, 6, 8, 10, their optimal values to model (14.30) are negative, making their super-efficiency measures all exceed one. Comparing the ranking results from various super-efficiency measures, two approaches, namely the standard radial model and the modified DDF-based model, lead to exactly the same rank, which is DMU 8, 2, 6, 1, 10, 3, 5, 9, 7, 4 from high to low. The N-L super-efficiency measure, however, provides a quite different rank for super-efficient units, which is DMU 2, 8, 6, 10, 1, 3, 5, 9, 7, 4 from high to low.

Lin and Chen (2014) indicate that Chen et al.'s (2013a) modified DDF-based super-efficiency model cannot fully address the infeasibility problem in Ray (2008). In some very special situations, Chen et al.'s (2013a) method fails to provide with reasonable results. Lin and Chen (2014) choose a new reference input-output bundle for the DDF and propose an alternative modified DDF-based VRS super-efficiency model.

## 14.6 Integer Super-Efficiency

Conventional DEA methods assume continuous values for input and output metrics. However, in many real managerial cases, some inputs and/or outputs can only take integer values. Take the scientific research in a university for example. Inputs such as the number of research staff and outputs such as the number of patents approved are restricted to non-negative integers. As pointed out by Kuosmanen and Kazemi Matin (2009), simply rounding the optimal solution to the nearest whole numbers can result in misleading efficiency assessment and reference targets. Researchers including Lozano and Villa (2006) and Kuosmanen and Kazemi Matin (2009) provides with their respective versions of mixed integer linear programming (MILP) formulations.

### 14.6.1 Integer-Valued Additive Super-Efficiency

In order to achieve a thorough differentiation of the performance among DMUs, Du et al. (2012) extend the efficiency analysis for integer-valued data into super-efficiency measurement by dealing directly with input and output slacks. They propose additive integer-valued super-efficiency models, which demonstrate a stronger discriminating power among DMUs compared with radial super-efficiency measures.

Du et al. (2012) first add the integer requirement in an additive DEA model (Charnes et al. 1982), and obtain the additive integer-valued model (14.34).

$$
\begin{aligned}
\hat{\rho}_o^* = \max \quad & \hat{\rho}_o = \sum_{i_{NI} \in I^{NI}} s_{i_{NI}o}^- + \sum_{r_{NI} \in O^{NI}} s_{r_{NI}o}^+ + \sum_{i_I \in I^I} \tilde{s}_{i_Io}^- + \sum_{r_I \in O^I} \tilde{s}_{r_Io}^+ \\
s.t. \quad & x_{i_{NI}o} - s_{i_{NI}o}^- = \sum_{j=1}^{n} \lambda_j x_{i_{NI}j}, i_{NI} \in I^{NI} \\
& y_{r_{NI}o} + s_{r_{NI}o}^+ = \sum_{j=1}^{n} \lambda_j y_{r_{NI}j}, r_{NI} \in O^{NI} \\
& \tilde{x}_{i_Io} \geq \sum_{j=1}^{n} \lambda_j x_{i_Ij}, i_I \in I^I \\
& x_{i_Io} - \tilde{s}_{i_Io}^- = \tilde{x}_{i_Io}, i_I \in I^I \\
& \tilde{y}_{r_Io} \leq \sum_{j=1}^{n} \lambda_j y_{r_Ij}, r_I \in O^I \\
& y_{r_Io} + \tilde{s}_{r_Io}^+ = \tilde{y}_{r_Io}, r_I \in O^I \\
& \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n \\
& \tilde{x}_{i_Io}, \tilde{y}_{r_Io} \in Z^+, i_I \in I^I, r_I \in O^I \\
& s_{i_{NI}o}^-, s_{r_{NI}o}^+, \tilde{s}_{i_Io}^-, \tilde{s}_{r_Io}^+ \geq 0, i_{NI} \in I^{NI}, r_{NI} \in O^{NI}, i_I \in I^I, r_I \in O^I
\end{aligned}
\tag{14.34}
$$

where $I^I$, $I^{NI}$, $O^I$ and $O^{NI}$ denote the subsets of integer-valued and real-valued inputs and outputs, respectively. In model (14.34), $\tilde{x}_{i_Io} \in Z^+$ and $\tilde{y}_{r_Io} \in Z^+$ are the integer targets for input $i_I$ and output $r_I$ of $DMU_o$. Non-radial slacks $s_{i_{NI}o}^-, s_{r_{NI}o}^+, \tilde{s}_{i_Io}^-, \tilde{s}_{r_Io}^+$ represent the actual inputs that can be reduced and actual outputs that can be increased in order to realize the best feasible target.

Based on an optimal solution to model (6.1), which is represented by $\left\{\lambda_j^*, j = 1, \ldots, n; s_{i_{NI}o}^{-*}, i_{NI} \in I^{NI}; s_{r_{NI}o}^{+*}, r_{NI} \in O^{NI}; \tilde{s}_{i_Io}^{-*}, i_I \in I^I; \tilde{s}_{r_Io}^{+*}, r_I \in O^I\right\}$, Du et al. (2012) define $\hat{\sigma}_o^* = \frac{1 - \frac{1}{m}\left[\sum_{i_{NI} \in I^{NI}} s_{i_{NI}o}^{-*}/x_{i_{NI}o} + \sum_{i_I \in I^I} \tilde{s}_{i_Io}^{-*}/x_{i_Io}\right]}{1 + \frac{1}{s}\left[\sum_{r_{NI} \in O^{NI}} s_{r_{NI}o}^{+*}/y_{r_{NI}o} + \sum_{r_I \in O^I} \tilde{s}_{r_Io}^{+*}/y_{r_Io}\right]}$ as the additive efficiency measure. This additive efficiency falls between zero and one, and a larger

value represents a better performance in reaching the efficient frontier. $DMU_o$ is called additive efficient if and only if $\hat{\sigma}_o^* = 1$, or equivalently, all of its optimal slacks are zero.

Suppose $DMU_o$ is additive efficient according to model (14.34). Du et al. (2012) propose an additive super-efficiency model designed for integer data.

$$\begin{aligned}
\hat{\alpha}_o^* = \min \quad & \hat{\alpha}_o = \sum_{i=1}^{m} t_{io}^- + \sum_{r=1}^{s} t_{ro}^+ \\
s.t. \quad & \sum_{j=1, j\neq o}^{n} \lambda_j x_{ij} \leq x_{io} + t_{io}^-, i = 1, \ldots, m \\
& \sum_{j=1, j\neq o}^{n} \lambda_j y_{rj} \geq y_{ro} - t_{ro}^+, r = 1, \ldots, s \\
& t_{i_I o}^-, t_{r_I o}^+ \in Z^+, i_I \in I^I, r_I \in O^I \\
& \sum_{j=1, j\neq o}^{n} \lambda_j = 1 \\
& \lambda_j, t_{i_{NI} o}^-, t_{r_{NI} o}^+ \geq 0, j = 1, \ldots, n, j \neq o, i_{NI} \in I^{NI}, r_{NI} \in O^{NI}
\end{aligned} \tag{14.35}$$

Alternative objective functions can be used for model (14.35) so that the resulting super-efficiency model is unit-invariant, for example,

$$\hat{\beta}_o^* = \min \hat{\beta}_o = \frac{1}{m+s} \left( \sum_{i=1}^{m} \frac{t_{io}^-}{x_{io}} + \sum_{r=1}^{s} \frac{t_{ro}^+}{y_{ro}} \right) \tag{14.36}$$

It is further proved by Du et al. (2012) that their additive integer-valued VRS super-efficiency model are consistently feasible.

Let $\left\{ \hat{\alpha}_o^*; \lambda_j^*, j = 1, \ldots, n, j \neq o; t_{io}^{-*}, i = 1, \ldots, m; t_{ro}^{+*}, r = 1, \ldots, s \right\}$ be an optimal solution to model (14.35). Du et al. (2012) use $\hat{\delta}_o^* = \frac{\frac{1}{m}\sum_{i=1}^{m}(x_{io} + t_{io}^{-*})/x_{io}}{\frac{1}{s}\sum_{r=1}^{s}(y_{ro} - t_{ro}^{+*})/y_{ro}} \geq 1$ to define for the additive super-efficiency score. Note that a greater $\hat{\delta}_o^*$ implies a superior performance compared with other efficient DMUs.

### 14.6.2 Additive Super-Efficiency for Undesirable Integer-Restricted Data

Extending the work of Du et al. (2012) to integer-restricted undesirable data, Chen et al. (2012) formulate an additive super-efficiency model.

$m_{GR}, s_{GR}, m_{BR}, s_{BR}, m_{GI}, s_{GI}, m_{BI}, s_{BI}$ are used to respectively represent the number of variables in the eight variable sets, which are characterized by inputs or outputs, continuous or integer, desirable or undesirable. Specifically, the subscripts "G" and "B" stand for "good" and "bad" inputs/outputs, respectively; the subscripts "R" and "I" stand for "real-valued" and "integer-valued" variables, respectively. All inputs

and outputs are assumed to be non-negative. They use the same VRS production possibility set (PPS) with Liu et al. (2010) as:

$$P = \left\{ \begin{array}{l} (X^{GR}, X^{GI}, X^{BR}, X^{BI}, Y^{GR}, Y^{GI}, Y^{BR}, Y^{BI}) \Big| \begin{pmatrix} X^{GI}, Y^{GI} \\ X^{BI}, Y^{BI} \end{pmatrix} \\ \in Z^{m_{GI}+m_{BI}+m_{GI}+m_{BI}}; \begin{pmatrix} X^{GR} \\ X^{GI} \end{pmatrix} \geq \sum_{j=1}^{n} \lambda_j \begin{pmatrix} X_j^{GR} \\ X_j^{GI} \end{pmatrix}, \\ \begin{pmatrix} X^{BR} \\ X^{BI} \end{pmatrix} \leq \sum_{j=1}^{n} \lambda_j \begin{pmatrix} X_j^{BR} \\ X_j^{BI} \end{pmatrix}; \begin{pmatrix} Y^{GR} \\ Y^{GI} \end{pmatrix} \leq \sum_{j=1}^{n} \lambda_j \begin{pmatrix} Y_j^{GR} \\ Y_j^{GI} \end{pmatrix}, \begin{pmatrix} Y^{BR} \\ Y^{BI} \end{pmatrix} \\ \geq \sum_{j=1}^{n} \lambda_j \begin{pmatrix} Y_j^{BR} \\ Y_j^{BI} \end{pmatrix}; \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0 \end{array} \right\} \tag{14.37}$$

In order to simultaneously tackle the integrality and undesirable factors in one model, Chen et al. (2012) modify the standard additive DEA model (Charnes et al. 1982) as:

$$\begin{aligned} Max \quad & \frac{1}{m_{GR} + m_{BR} + m_{GI} + m_{BI} + s_{GR} + s_{BR} + s_{GI} + s_{BI}} \\ & \left( \sum \frac{s_{io}^{GR-}}{x_{io}^{GR}} + \sum \frac{s_{io}^{BR-}}{x_{io}^{BR}} + \sum \frac{s_{ro}^{GR+}}{y_{ro}^{GR}} + \sum \frac{s_{ro}^{BR+}}{y_{ro}^{BR}} \right. \\ & \left. + \sum \frac{s_{io}^{GI-}}{x_{io}^{GI}} + \sum \frac{s_{io}^{BI-}}{x_{io}^{BI}} + \sum \frac{s_{ro}^{GI+}}{y_{ro}^{GI}} + \sum \frac{s_{ro}^{BI+}}{y_{ro}^{BI}} \right) \\ s.t. \quad & X_o^{GR} - S_o^{GR-} = \sum_{j=1}^{n} \lambda_j X_j^{GR}, X_o^{BR} + S_o^{BR-} = \sum_{j=1}^{n} \lambda_j X_j^{BR} \\ & Y_o^{GR} + S_o^{GR+} = \sum_{j=1}^{n} \lambda_j Y_j^{GR}, Y_o^{BR} - S_o^{BR+} = \sum_{j=1}^{n} \lambda_j Y_j^{BR} \\ & X_o^{GI} - S_o^{GI-} \geq \sum_{j=1}^{n} \lambda_j X_j^{GI}, X_o^{BI} + S_o^{BI-} \leq \sum_{j=1}^{n} \lambda_j X_j^{BI} \\ & Y_o^{GI} + S_o^{GI+} \leq \sum_{j=1}^{n} \lambda_j Y_j^{GI}, Y_o^{BI} - S_o^{BI+} \geq \sum_{j=1}^{n} \lambda_j Y_j^{BI} \\ & \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n \\ & S_o^{GI-} \in Z_+^{m_{GI}}, S_o^{BI-} \in Z_+^{m_{BI}}, S_o^{GI+} \in Z_+^{s_{GI}}, S_o^{BI+} \in Z_+^{s_{BI}} \\ & S_o^{GR-}, S_o^{BR-}, S_o^{GR+}, S_o^{BR+} \geq 0 \end{aligned} \tag{14.38}$$

where slack variables $S_o^{GR-}, S_o^{BR-}, S_o^{GR+}, S_o^{BR+}, S_o^{GI-}, S_o^{BI-}, S_o^{GI+}, S_o^{BI+}$ represent the absolute differences between the original input/output values and their respective

reference points. $DMU_o$ is regarded as additive efficient if and only if all optimal slacks in model (14.38) are zero.

To further discriminate efficient DMUs, Chen et al. (2012) remove the evaluated DMU from the reference set and modify the constraints and objective of model (14.37) to get the unit-invariant super-efficiency model:

$$
\begin{aligned}
Min \quad & \frac{1}{m_{GR}+m_{BR}+m_{GI}+m_{BI}+s_{GR}+s_{BR}+s_{GI}+s_{BI}} \\
& \left( \sum \frac{t_{io}^{GR-}}{x_{io}^{GR}} + \sum \frac{t_{io}^{BR-}}{x_{io}^{BR}} + \sum \frac{t_{ro}^{GR+}}{y_{ro}^{GR}} + \sum \frac{t_{ro}^{BR+}}{y_{ro}^{BR}} \right. \\
& \left. + \sum \frac{t_{io}^{GI-}}{x_{io}^{GI}} + \sum \frac{t_{io}^{BI-}}{x_{io}^{BI}} + \sum \frac{t_{ro}^{GI+}}{y_{ro}^{GI}} + \sum \frac{t_{ro}^{BI+}}{y_{ro}^{BI}} \right) \\
s.t. \quad & X_o^{GR} + T_o^{GR-} \geq \sum_{j=1,j\neq o}^{n} \lambda_j X_j^{GR}, X_o^{BR} - T_o^{BR-} \leq \sum_{j=1,j\neq o}^{n} \lambda_j X_j^{BR} \\
& Y_o^{GR} - T_o^{GR+} \leq \sum_{j=1,j\neq o}^{n} \lambda_j Y_j^{GR}, Y_o^{BR} + T_o^{BR+} \geq \sum_{j=1,j\neq o}^{n} \lambda_j Y_j^{BR} \\
& X_o^{GI} + T_o^{GI-} \geq \sum_{j=1,j\neq o}^{n} \lambda_j X_j^{GI}, X_o^{BI} - T_o^{BI-} \leq \sum_{j=1,j\neq o}^{n} \lambda_j X_j^{BI} \\
& Y_o^{GI} - T_o^{GI+} \leq \sum_{j=1,j\neq o}^{n} \lambda_j Y_j^{GI}, Y_o^{BI} + T_o^{BI+} \geq \sum_{j=1,j\neq o}^{n} \lambda_j Y_j^{BI} \\
& \sum_{j=1,j\neq o}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1,\ldots,n, j \neq o \\
& T_o^{BR-} \leq X_o^{BR}, T_o^{GR+} \leq Y_o^{GR}, T_o^{BI-} \leq X_o^{BI}, T_o^{GI+} \leq Y_o^{GI} \\
& T_o^{GI-} \in Z_+^{m_{GI}}, T_o^{BI-} \in Z_+^{m_{BI}}, T_o^{GI+} \in Z_+^{s_{GI}}, T_o^{BI+} \in Z_+^{s_{BI}} \\
& T_o^{GR-}, T_o^{BR-}, T_o^{GR+}, T_o^{BR+} \geq 0
\end{aligned}
\tag{14.39}
$$

VRS super-efficiency model (14.39) overcomes the infeasibility problem of the standard VRS super-efficiency measure, by using slacks to scale up the inputs (or undesirable outputs) and scale down the outputs (or undesirable inputs) of the assessed DMU. Moreover, additive super-efficiency model (14.39) helps in determining the maximum allowable increase in each desirable input or undesirable output, as well as the maximum allowable decrease in each desirable output or undesirable input, given that the efficiency status of an efficient DMU stay unchanged.

Chen et al. (2012) define

$$\delta_o^* = \frac{\frac{1}{m_{GR}+m_{GI}+s_{BR}+s_{BI}}\left(\sum \frac{x_{io}^{GR}+t_{io}^{GR-*}}{x_{io}^{GR}} + \sum \frac{x_{io}^{GI}+t_{io}^{GI-*}}{x_{io}^{GI}} + \sum \frac{y_{ro}^{BR}+t_{ro}^{BR+*}}{y_{ro}^{BR}} + \sum \frac{y_{ro}^{BI}+t_{ro}^{BI+*}}{y_{ro}^{BI}}\right)}{\frac{1}{s_{GR}+s_{GI}+m_{BR}+m_{BI}}\left(\sum \frac{y_{ro}^{GR}-t_{ro}^{GR+*}}{y_{ro}^{GR}} + \sum \frac{y_{ro}^{GI}-t_{ro}^{GI+*}}{y_{ro}^{GI}} + \sum \frac{x_{io}^{BR}-t_{io}^{BR-*}}{x_{io}^{BR}} + \sum \frac{x_{io}^{BI}-t_{io}^{BI-*}}{x_{io}^{BI}}\right)} \tag{14.40}$$

as the additive super-efficiency for $DMU_o$, where $(\lambda_j^*, j \neq o; T_{io}^{GI-*}, T_{io}^{BI-*}, T_{ro}^{GI+*}, T_{ro}^{BI+*}, T_{io}^{GR-*}, T_{io}^{BR-*}, T_{ro}^{GR+*}, T_{ro}^{BR+*})$ is an optimal solution to model (14.39). Super-efficiency measure $\delta_o^*$ values no less than one ( $\delta_o^* \geq 1$), and increases monotonically in both input and output slacks. Thus a greater score represents a superior performance compared with other efficient DMUs.

### *14.6.3 DDF-Based Integer Super-Efficiency*

In order to accommodate integer data, Chen et al. (2013b) seek to modify Ray's (2008) Nerlove-Luenberger (N-L) measure of super-efficiency, and find that this DDF-based approach cannot be directly changed to incorporate integer requirement. Actually, the DDF (directional distance function) requires that the inputs decrease at the same rate as outputs increase to reach the DEA frontier, which becomes problematic when inputs and outputs are integers under the concept of super-efficiency. To address the above problem, Chen et al. (2013b) assume different changing rates for inputs and outputs of the DMU under evaluation to reach the frontier constructed by the remaining DMUs. In doing this, their DDF-based integer super-efficiency avoids suffering from the infeasibility problem under VRS.

Suppose that part of the inputs and outputs are constrained to integer values, and denoted the subsets of integer-valued, real-valued inputs and outputs by $I^I$, $I^{NI}$, $O^I$ and $O^{NI}$, respectively. For any integer-restricted measure, its reference target with respect to the efficient frontier is also supposed to be an integer. Therefore, the N-L super-efficiency measure considering integer data is presented by Chen et al. (2013b)

as

$$
\begin{aligned}
\max \quad & \beta_k \\
s.t. \quad & y_{rk} + \beta_k y_{rk} = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj} - s_r^+, r \in O^{NI} \\
& x_{ik} - \beta_k x_{ik} = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij} + s_i^-, i \in I^{NI} \\
& \tilde{x}_{ik} - s_i^I = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij}, i \in I^I \\
& \tilde{y}_{rk} + s_r^I = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj}, r \in O^I \\
& \tilde{x}_{ik} + s_i^- = x_{ik} - \beta_k x_{ik}, i \in I^I \\
& \tilde{y}_{rk} - s_r^+ = y_{rk} + \beta_k y_{rk}, r \in O^I \\
& \tilde{x}_{ik} \in Z_+, i \in I^I \\
& \tilde{y}_{rk} \in Z_+, r \in O^I \\
& \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n, j \neq k \\
& s_r^+ \geq 0, s_i^- \geq 0, r \in O^I \cup O^{NI}, i \in I^I \cup I^{NI} \\
& s_i^I \geq 0, i \in I^I, s_r^I \geq 0, r \in O^I \\
& \beta_k \textit{ free in sign}
\end{aligned}
\tag{14.41}
$$

However, when the integer restriction is taken into account, it is very likely that inputs cannot decrease at the same rate as outputs increase. This fact will lead to erroneous results for some DMUs.

Chen et al. (2013b) settle the above problem by assigning different rates $\beta_x$ and $\beta_y$ to inputs and outputs, respectively. If DMU $k$ is efficient, inputs should be augmented and outputs should be contracted, which implies $\beta_x \leq 0$ and $\beta_y \leq 0$. On the other hand, if DMU $k$ is inefficient, inputs should be contracted and outputs should be augmented, which means $\beta_x \geq 0$ and $\beta_y \leq 0$. These two situations can be incorporated by enforcing $\beta_x \beta_y \geq 0$. Based on the above analysis, Chen et al. (2013b)

modify model (14.41) into

$$
\begin{aligned}
\max \quad & \beta_x + \beta_y \\
s.t. \quad & y_{rk} + \beta_y y_{rk} = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj} - s_r^+, r \in O^{NI} \\
& x_{ik} - \beta_x x_{ik} = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij} + s_i^-, i \in I^{NI} \\
& \tilde{x}_{ik} - s_i^I = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j x_{ij}, i \in I^I \\
& \tilde{y}_{rk} + s_r^I = \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j y_{rj}, r \in O^I \\
& \tilde{x}_{ik} + s_i^- = x_{ik} - \beta_x x_{ik}, i \in I^I \\
& \tilde{y}_{rk} - s_r^+ = y_{rk} + \beta_y y_{rk}, r \in O^I \\
& \tilde{x}_{ik} \in Z_+, i \in I^I \\
& \tilde{y}_{rk} \in Z_+, r \in O^I \\
& \sum_{\substack{j=1 \\ j \neq k}}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n, j \neq k \\
& s_r^+ \geq 0, s_i^- \geq 0, r \in O^I \cup O^{NI}, i \in I^I \cup I^{NI} \\
& s_i^I \geq 0, i \in I^I, s_r^I \geq 0, r \in O^I \\
& \beta_x \beta_y \geq 0
\end{aligned}
\tag{14.42}
$$

Note that model (14.42) is non-linear due to the constraint of $\beta_x \beta_y \geq 0$. Chen et al. (2013b) transform this constraint into the following set of linear constraints by introducing two binary integer variables, $w$ and $z$.

$$
\begin{aligned}
& -M(1-w) \leq \beta_x \leq Mw \\
& -Mz \leq \beta_y \leq M(1-z) \\
& w + z = 1 \\
& w \in \{0, 1\}, z \in \{0, 1\}
\end{aligned}
$$

where $M$ is a sufficiently large number. Chen et al. (2013b) point out that $w = 1$ and $z = 0$ signify $\beta_x \geq 0$ and $\beta_y \geq 0$, and $w = 0$ and $z = 1$ signify $\beta_x \leq 0$ and $\beta_y \leq 0$, respectively. Therefore, $\beta_x \beta_y \geq 0$ could be replaced with the above set of

linear constraints plus the two binary integer variables, and non-linear model (14.42) becomes a mixed integer linear programming problem.

Since $\beta_x^* \leq 1$ and $\beta_y^* \geq -1$, Chen et al. (2013b) define $\frac{1-\beta_x^*}{1+\beta_y^*}$ as the super-efficiency score. The higher the score is, the more efficient a DMU under evaluation is. When a DMU under evaluation is inefficient, this score is between 0 and 1. When a DMU under evaluation is efficient, this score is greater than 1. When $\beta_y^* = -1$, the score diverges to infinity and the super-efficiency score of infinity is allowed in Chen et al. (2013b).

## 14.7 Conclusions

This chapter introduces the concept of super-efficiency and various super-efficiency measures, especially under the assumption of VRS. Besides the infeasibility issue which is caused by the convexity constraint in VRS models, zero data can be problematic in any super-efficiency approaches. For example, Zhu (1996) shows that the CRS super-efficiency model becomes infeasible when an efficient DMU has zero inputs. Lee and Zhu (2012) claim that either the conventional or the modified VRS super-efficiency models, such as the two-stage procedures provided by Cook et al. (2009) and Lee et al. (2011) with an attempt to address infeasibility, will become infeasible when zero data are present. They thus extend the work of Lee et al. (2011) to make the revised model feasible when zero data exist in inputs.

**Acknowledgements** This research is partially funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, China. Dr. Juan Du thanks the support by the National Natural Science Foundation of China (Grant No. 71471133).

## References

Andersen P, Petersen NC (1993) A procedure for ranking efficient units in data envelopment analysis. Manage Sci 39(10):1261–1264

Chambers RG, Chung Y, Färe R (1996) Benefit and distance functions. J Econ Theory 70:407–419.

Charnes A, Cooper WW, Seiford LM, Stutz J (1982) A multiplicative model for efficiency analysis. Socio-Econ Plan Sci 16(5):223–224

Charnes A, Cooper WW, Thrall RM (1991) A structure for classifying and characterizing efficiency and inefficiency in data analysis. J Product Anal 2:197–237

Charnes A, Cooper WW (1962) Programming with linear fractional functional. Naval Research Logistics Quarterly 15:333–334

Chen Y (2004) Ranking efficient units in DEA. Omega 32:213–219

Chen Y (2005) Measuring super-efficiency in DEA in the presence of infeasibility. Eur J Oper Res 161:545–551

Chen Y, Liang L (2011) Super-efficiency DEA in the presence of infeasibility: one model approach. Eur J Oper Res 213:359–360

Chen C-M, Du J, Huo JZ, Zhu J (2012) Undesirable factors in integer-valued DEA: evaluating the operational efficiencies of city bus systems considering safety records. Decis Support Syst 54(1):330–335

Chen Y, Du J, Huo JZ (2013a) Super-efficiency based on a modified directional distance function. Omega 41:621–625

Chen Y, Djamasbi s, Du J, Lim S (2013b) Integer-valued DEA super-efficiency based on directional distance function with an application of evaluating mood and its impact on performance. Int J Prod Econ 46(2):550–556

Cook WD, Liang l, Zha Y, Zhu J (2009) A modified super-efficiency DEA model for infeasibility. J Oper Res Soc 60:276–281

Du J, Liang L, Zhu J (2010) A slacks-based measure of super-efficiency in data envelopment analysis: a comment. Eur J Oper Res 204:694–697

Du J, Chen C-M, Chen Y, Cook WD, Zhu J (2012) Additive super-efficiency in integer-valued data envelopment analysis. Eur J Oper Res 218(1):186–192

Kuosmanen T, Kazemi Matin R (2009) Theory of integer-valued data envelopment analysis. Eur J Oper Res 192:658–667

Lee H-S, Zhu J (2012) Super-efficiency infeasibility and zero data in DEA. Eur J Oper Res 216: 429–433

Lee H-S, Chu CW, Zhu J (2011) Super-efficiency DEA in the presence of infeasibility. Eur J Oper Res 212:141–147

Lin RY, Chen ZP (2014) Super-efficiency measurement under variable return to scale: an approach based on a new directional distance function. Journal of the Operational Research Society, in press

Liu WB, Meng W, Li XX, Zhang DQ (2010) DEA models with undesirable inputs and outputs. Ann Oper Res 173:177–194

Lovell CAK, Rouse APB (2003) Equivalent standard DEA models to provide super-efficiency scores. J Oper Res Soc 54:101–108

Lozano S, Villa G (2006) Data envelopment analysis of integer-valued inputs and outputs. Comput Oper Res 33(10):3004–3014

Ray SC (2008) The directional distance function and measurement of super-efficiency: an application to airlines data. J Oper Res Soc 59(6):788–797

Seiford LM, Zhu J (1999) Infeasibility of super-efficiency data envelopment analysis models. Infor 37(2):174–187

Thrall RM (1996) Duality, classification and slacks in DEA. Ann Oper Res 66:109–138

Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509

Tone K (2002) A slacks-based measure of super-efficiency in data envelopment analysis. Eur J Oper Res 143:32–41

Zhu J (1996) Robustness of the efficient DMUs in data envelopment analysis. Eur J Oper Res 90:451–460

# Chapter 15
# DEA Models with Undesirable Inputs, Intermediates, and Outputs

**Zhongbao Zhou and Wenbin Liu**

**Abstract** In real applications involving the use of Data Envelopment Analysis (DEA) models, undesirable inputs and outputs have been frequently encountered and addressed, e.g., via data transformation. These studies were scattered in the literature and often confined to some particular applications. In this paper, we present a systematic investigation concerning the building of DEA models. First, we describe the desirability of inputs and outputs, as well as the disposability assumptions in the presence of undesirable inputs and outputs. Next we construct a number of DEA models with different disposability assumptions and performance measures for the case of single-stage DEA. Next, we try to systematically investigate two-stage DEA models with undesirable inputs, intermediates and outputs. Particularly, we utilize the free-disposal axioms to construct the production possibility sets and the corresponding DEA models with undesirable inputs, intermediates, and outputs.

**Keywords** Data envelopment analysis · Two-stage systems · Undesirable variables · Extended strongly free disposability · Weakly free disposability · Production possibility set

## 15.1 Introduction

Since the introduction of DEA in 1978, it has been widely used in efficiency analysis of many business and industry applications. Excellent literature surveys can be found in, for instance, Seiford (1996) and Cooper et al. (2004). The best-known DEA models are the CCR model (Charnes et al. 1978), the BCC model (Banker et al. 1984), the Additive model (Charnes et al. 1985), and the Cone Ratio model (Charnes

W. Liu (✉) · Z. Zhou
School of Business Administration, Hunan University, 410082 Changsha, China
e-mail: W.B.Liu@kent.ac.uk

Z. Zhou
e-mail: ZhongbaoZhou@gmail.com

W. Liu
Kent Business School, University of Kent, CT2 7PE Canterbury, UK

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_15

et al. 1989). These DEA models were all formulated for desirable inputs and outputs; however, there frequently exist undesirable inputs and/or outputs in real applications.

In DEA literature, extensive research already exists concerning applications with un-desirable inputs and/or outputs. There is a useful summary in Liu et al. (2010) for the case of single-stage DEA. Although extensive literature review is not a primary task of this work, a portion of the existing approaches are briefly summarized as follows:

An intuitive reaction is to apply certain transformations. Translation $f(U) = -U + \beta$ is the most widely used one (e.g., Ali and Seiford 1990; Pastor 1996; Scheel 2001; Seiford and Zhu 2002). However, it is well-known that not only ranking but also classification may depend on β. Another widely used one is $f(U) = -U$, the so-called ADD approach suggested by Koopmans (1951); the undesirable inputs or outputs will become desirable after this transformation. However, the data may subsequently become negative, and it is not straightforward to define efficiency scores for negative data. It is useful to realize that the additive models are able to handle negative data; these models will be discussed subsequently. The approaches based on data-transformation may unexpectedly produce adverse results as discussed in Liu and Sharp (1999). Nonlinear transformations, such as the multiplicative inverse: $f(U) = 1/U$ (e.g., Golany and Roll 1989; Lovell et al. 1995), can also be used. Being a nonlinear transformation, this transformation's behaviors are even more complicated (Scheel 1998). Thus, how to properly select a suitable transformation is highly case-dependent.

There also exist many approaches that can avoid data transformation. For example, one may regard undesirable inputs as desirable outputs, or undesirable outputs as desirable inputs; see Liu and Sharp (1999) for an initial attempt to formulate this method. This approach is an attractive method in studying operational efficiency for single-stage DEA due to its simplicity and elegance, although it changes the physical input-output relationship, especially in the case of two-stage DEA. We will further extend this approach in this work and discuss its relationship with other approaches.

Our investigation focuses on theoretical aspects of these issues. Our main idea is to examine these aspects within the general framework proposed in Liu et al. (2010), where the free disposability and the possible production sets are extensively used to address undesirable variables. The principal objective of this paper is to discuss desirability of inputs/outputs, disposability assumptions, and production possibility sets and to construct a number of DEA models for single-stage and two-stage systems in the presence of undesirable measures. This approach leads to a unified framework of DEA models with undesirable measures.

## 15.2 Single-Stage DEA Models with Undesirable Variables

### *15.2.1 Desirability and Disposability*

In single-stage DEA models, we normally assume that we know which variables are desirable and which are not. In two-stage DEA models, the meanings of desirable can be controversial or disputable. Thus here we have to formally define them. It is also

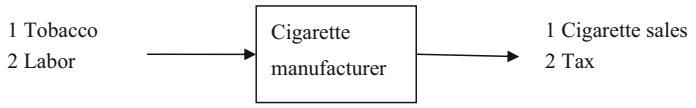



**Fig. 15.1** Cigarette manufacturer example

interesting to note that desirability of variables affects its disposability. For example, normally we can only freely dispose extra desirable inputs or outputs. Therefore we here discuss these two concepts in the same section.

#### 15.2.1.1 Desirability Determination

1. **Desirability of outputs**

Desirability of inputs and outputs has been used in an ad hoc fashion in the DEA literature for single-stage systems. However for rigorously dealing with two-stage DEA models with undesirable intermediates later on, it is essential to firstly clarify the precise definition of desirability.

We believe that the desirability of outputs is often determined by the Decision Maker (DM) in real applications. What the DM hopes to produce as much as possible are desirable outputs and otherwise they are undesirable outputs in our framework. Thus, the types of outputs reflect the subjective judgments of the DM. For example, in the cigarette manufacturer example in Fig. 15.1, if the DM is cigarette manufacturers, they will hope to sell cigarettes as many as possible while to pay tax as little as possible in order to obtain more profits. Therefore, cigarette sales is a desirable output and the tax is an undesirable output. However, if the DM is a government agency with the interests of the whole nation, they will prefer smaller sales but higher tax in order to protect public health and reduce medical expenses due to smoking. In this case, the cigarette sale is an undesirable output and the tax is a desirable output.

2. **Desirability of inputs**

After determining the desirability of outputs, we argue that the desirability of inputs should be defined according to the intrinsic production mechanisms. If the increase of an input will not reduce the desirable outputs, then it is desirable. If its increase will not increase the desirable outputs, then it is classified as undesirable (because the purpose of a practical system is to obtain the desirable outputs, undesirable outputs are not considered in determining the types of inputs). For example, a power plant may produce waste gases at the time of producing electricity. However we cannot classify the fuel as an undesirable input because it produces the waste gases. In the post office example in Fig. 15.2, if the post office is the DM, the amount of correctly delivered letters is a desirable output while the amount of wrongly delivered letters is an undesirable output. The increase of the letters with correct addresses will increase the amount of correctly delivered letters, so the amount of the letters with correct

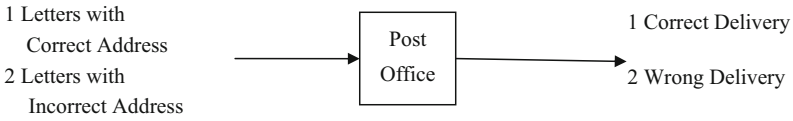



**Fig. 15.2** Post office example

addresses is a desirable input. Similarly, we can determine that the amount of the letters with incorrect addresses is an undesirable input.

#### 15.2.1.2 Disposability Assumptions in DEA

Assume that there are n decision-making units (DMUs) to be evaluated. Let $X_j$ and $Y_j$ denote the inputs and outputs of $DMU_j$ with $j = 1, 2, \ldots, n$. The Production Possibility Set (PPS) is one of the building blocks of a DEA model (see Liu et al. (2006) for the other blocks).

A PPS contains all of the realizable DMUs for identifying the best DMUs in a DEA model, although some of the DMUs may not, in fact, exist. In the DEA theory, the DMUs are referred to as "virtual" DMUs and are also included in the comparisons. If a DMU $(X_j, Y_j)$ is found to be the "best" in the PPS using the Pareto preference then it is considered to be efficient. In the standard DEA models several assumptions are made on the PPS, such as convexity and no-free-lunch. The most relevant property here is the disposability, which states as **"free disposal"**.

1. **Extended strongly free disposability**

The property of strongly free disposability holds if the absorption of any additional amounts of inputs without any reduction in outputs is always possible. Let $P$ be the Production Possibility Set, the assumption can be stated:

$$\text{if } (X, Y) \in P \text{ and } W \geq X, Z \leq Y, \text{then } (W, Z) \in P.$$

Let us note that such free disposal can only hold up to some extent in practice as W cannot be infinitely large-if so eventually one will not be able to disposal it freely. Assuming the strong disposal, convexity and the minimum span, then the standard PPS spanned from the inputs and the outputs has the following form for desirable inputs and outputs.

$$PPS = \left\{ (X, Y) : X \geq X(\lambda) = \sum_{j=1}^{n} \lambda_j X_j,\ Y \leq Y(\lambda) = \sum_{j=1}^{n} \lambda_j Y_j, \lambda \in S \right\} \quad (15.1)$$

where $S = \{\lambda_j \geq 0, j = 1,2,\ldots n\}$ or $S = \{\lambda_j \geq 0, \sum_{j=1}^{n} \lambda_j = 1\}$ in the DEA literature.

To handle undesirable inputs or outputs satisfactorily, one needs to extend the strongly free disposability. There seems to exist several possible ways.

Here we directly extend the above strongly free disposability via using the same statement but with the preferences adopted for undesirable. Formally this extended strongly free disposability can be stated as:

Let $(X, Y) = (X^D, X^U, Y^D, Y^U) \in P$ be desirable and undesirable inputs and outputs respectively, if $W^D \geq X^D$, $W^U \leq X^U$ and $Z^D \leq Y^D$, $Z^U \geq Y^U$, then $(W^D, W^U, Z^D, Z^U) \in P$.

There are many practical situations where such free disposability can be observed. Take a post-office for instance, letters with correct addresses are good inputs but those with in- correct addresses are bad ones. Therefore one can produce a given output with more good inputs and fewer bad inputs. Most electricity generators have pollution control systems, such as equipments to reduce sulfur dioxide in their production processes. Thus undesirable outputs like sulfur dioxide can be "freely" increased, at least to some extent, by shutting down these pollution control systems. Similar examples can be found in service sectors where the desirable and undesirable outputs are numbers of served customers and received complaints respectively. If there are plenty of customers, then the extended strongly free disposability holds as it is possible to freely increase complaints without reducing numbers of serviced customers. It will be seen below that many existing DEA models in fact use this type of extended strongly free disposability to handle undesirable variables. With the extended strongly free disposability, the corresponding PPS with convexity reads:

$$PPS = \left\{ (X^D, X^U, Y^D, Y^U) : X^D \geq \sum_{j=1}^{n} \lambda_j X_j^D \, , X^U \leq \sum_{j=1}^{n} \lambda_j X_j^U, \right.$$
$$\left. Y^D \leq \sum_{j=1}^{n} \lambda_j Y_j^D, Y^U \geq \sum_{j=1}^{n} \lambda_j Y_j^U, \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0 \right\}. \tag{15.2}$$

It is clear that the above PPS can be equivalently expressed via regarding the undesirable inputs (outputs) as desirable outputs (inputs) respectively, and then applying the standard strongly free disposability. Therefore this conclusion provides a theoretical foundation to the approach of exchanging undesirable variables with desirable ones, which will be examined in more detail later on.

Below we firstly show that many existing DEA models have assumed the extended strongly free disposability. For instance, in many cases assuming the strongly free disposability for transferred variables is just to assume the extended strongly free disposability for the original variables. Let us further elaborate that the model in Seiford and Zhu (2002) actually assumed extended strongly free disposability and used the above PPS in the original variables, where they used the transformation $\overline{Y}_j^U = -Y_j^U + W$, with $Y_j^U < W$. Then they assumed the standard Strong Free

Disposability and convexity. Thus the PPS with the new variables reads:

$$\left\{(X,Y)\colon X \geq \sum_{j=1}^{n}\lambda_j X_j, Y^D \leq \sum_{j=1}^{n}\lambda_j Y_j^D, \overline{Y}^U \leq \sum_{j=1}^{n}\lambda_j \overline{Y}_j^U, \sum_{j=1}^{n}\lambda_j = 1, \lambda_j \geq 0\right\}. \tag{15.3}$$

Then back to the original variables via $\overline{Y}_j^U = -Y_j^U + W$, the PPS reads:

$$\left\{(X,Y)\colon X \geq \sum_{j=1}^{n}\lambda_j X_j, Y^D \leq \sum_{j=1}^{n}\lambda_j Y_j^D, Y^U \geq \sum_{j=1}^{n}\lambda_j Y_j^U, \sum_{j=1}^{n}\lambda_j = 1, \lambda_j \geq 0\right\}. \tag{15.4}$$

Thus one may say that this model is in fact based on the extended strongly free disposability. Here the convexity plays a key role in deriving the equivalence of PPS if $W \neq 0$. Therefore it follows that the above equivalence holds for ADD transformation. On the other hand, if outputs are desirable but some of them are negative, then one may first apply ADD to change them into undesirable but positive variables, and then use extended strongly free disposability.

2. **Weakly free disposability**

The basic idea of weakly free disposability for outputs is that the undesirable outputs may not be reduced alone but may be reduced with a proportional reduction of certain desirable outputs. This idea was first developed by Shephard, where all inputs (outputs) must be increased (decreased) with the same percentage (Shephard 1970). This notion is different from extended strongly free disposability. Take the coal plant, for instance: the desirable and undesirable outputs are the electricity and carbon dioxide produced by burning coal, respectively. Weakly free disposability implies that a fixed percent reduction in carbon dioxide is possible if accompanied by the same percent reduction in the output of electricity provided the inputs remain unchanged. There are many studies addressing weakly free disposability for outputs, especially in banking and environment performance evaluation, such as Färe et al. (2005).

One example of weakly free disposability is formally stated as: undesirable outputs are weakly disposal

$$\text{if } (Y^D, Y^U) \in P(X) \text{ and } 0 \leq \alpha \leq 1, \text{then } (\alpha Y^D, \alpha Y^U) \in P(X)$$

It is easy to derive the weakly free disposability for inputs.

The recent research by (Kuosmanen 2005; Kuosmanen and Kazemi Matin 2011; Podinovski and Kuosmanen 2011) argues that the correct implementation of the weakly free disposability axiom requires the use of different abatement factors for each observed activity. In the following sections, we will adopt the classic definition of weakly free disposability.

The production possibility set with weakly free disposability can be expressed as

$$\left\{(X, Y^D, Y^U) \left| \sum_{j=1}^{n} \lambda_j X_j \le X, \sum_{j=1}^{n} \lambda_j Y_j^D = \alpha Y^D, \sum_{j=1}^{n} \lambda_j Y_j^U = \alpha Y^U, \alpha \ge 1, \lambda_j \in S \right.\right\} \tag{15.5}$$

In practice, there exist hybrid situations where certain outputs are weakly free disposal while other outputs are strongly free disposal (e.g., Färe and Grosskopf 2004; Ray 2004), or some inputs are weakly free disposal and other inputs are weakly free disposal, while all outputs are strongly free disposal.

The extended strongly free disposability and the weakly free disposability are independent. **Whether one should assume an extended strongly free disposability or weakly free disposability in a DEA model mostly depends on the nature of the applications it handles**. Taking the service example above, for instance, if the market has already become very competitive then it is no longer possible to increase complaints freely; in this case, one should consider a weakly free disposability instead. However, in this paper, unless otherwise stated, we will always assume extended strongly free disposability.

### *15.2.2 DEA Models with Undesirable Inputs/Outputs for Single-Stage Systems*

1. **Slacks-based DEA models**

In this section we examine slacks-based DEA models. For the case with undesirable inputs and outputs, we will assume that the inputs and output of *j-th* unit can be decomposed into

$$X = \begin{pmatrix} X_j^{DI} \\ X_j^{UI} \end{pmatrix}, \; Y_j = \begin{pmatrix} Y_j^{DO} \\ Y_j^{UO} \end{pmatrix},$$

with *{DI}, {UI}, {DO}, {UO}* being fixed index sets independent of $j$, such that $X_j^{DI}$, $Y_j^{DO}$ are desirable inputs and outputs, and $X_j^{UI}$, $Y_j^{UO}$ are undesirable inputs and outputs. For instance, $DI = \{1,2\}$, $UI = \{3,4,\ldots,m\}$, $DO = \{1,2,3\}$, $UO = \{4,5,\ldots,s\}$ so that $|DI| = 2$, $|UI| = m-2$, $|DO| = 3$, $|UO| = s-3$. For example, we assume $m = s = 5$ below,

$$X = \begin{pmatrix} x_1^{DI} \\ x_2^{DI} \\ x_3^{UI} \\ x_4^{UI} \\ x_5^{UI} \end{pmatrix}, \; Y = \begin{pmatrix} y_1^{DO} \\ y_2^{DO} \\ y_3^{UO} \\ y_4^{UO} \\ y_5^{UO} \end{pmatrix},$$

In such a case, let

$$X = \begin{pmatrix} X^{DI} \\ X^{UI} \end{pmatrix}, \; Y = \begin{pmatrix} Y^{DO} \\ Y^{UO} \end{pmatrix}.$$

We will assume the extended strongly free disposability. Note that maximizing desirable outputs (undesirable inputs) and minimizing undesirable outputs (desirable inputs) can be achieved by maximizing the correspondent slack measurements. Next, the Additive DEA model with measure weights reads:

$$\begin{aligned} \max \quad & w_{DI}^t s^{DI} + w_{UI}^t s^{UI} + w_{DO}^t s^{DO} + w_{UO}^t s^{UO}, \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j Y_j^{DO} - s^{DO} = Y_0^{DO} \\ & \sum_{j=1}^{n} \lambda_j Y_j^{UO} + s^{UO} = Y_0^{UO} \\ & \sum_{j=1}^{n} \lambda_j X_j^{DI} + s^{DI} = X_0^{DI} \\ & \sum_{j=1}^{n} \lambda_j X_j^{UI} - s^{UI} = X_0^{UI}, \\ & s^{DI}, s^{UI}, s^{DO}, s^{UO} \geq 0, \lambda \in S \end{aligned} \tag{15.6}$$

where $w_{DI}, w_{UI}, w_{DO}, w_{UO}$ are (strictly) positive weight vectors. Next, $DMU_0$ is efficient if and only if the maximum is zero.

However, the above DEA model cannot produce efficiency scores directly. For the desirable nonnegative inputs and outputs, one can use the Tone (2001) formula:

$$\begin{aligned} \min \quad & \rho = \frac{1 - \frac{1}{m}\sum_{i=1}^{m} s_i^- / x_{i0}}{1 + \frac{1}{s}\sum_{r=1}^{s} s_r^+ / y_{r0}}. \\ s.t. \quad & X_0 = \sum_{j=1}^{n} \lambda_j X_j + s^- \\ & Y_0 = \sum_{j=1}^{n} \lambda_j Y_j - s^+ \\ & \lambda \in S, s^- \geq 0, s^+ \geq 0. \end{aligned} \tag{15.7}$$

The model was shown to be units invariant and the scores to be between [0, 1]. Later we will see that the division of $X_0$, $Y_0$ for the slacks defined as above may need to be changed. For the case where there are undesirable variables, but they are nonnegative, the above DEA model can be readily extended as follows:

$$
\begin{aligned}
\min \quad & \rho = \frac{1 - \frac{1}{|DI|+|UO|}\left(\sum s_i^{DI}/x_{i0}^{DI} + \sum s_i^{UO}/y_{i0}^{UO}\right)}{1 - \frac{1}{|DO|+|UI|}\left(\sum s_r^{DO}/y_{r0}^{DO} + \sum s_i^{UO}/y_{i0}^{UO}\right)} \\
s.t. \quad & \sum_{j=1}^{n} \lambda_j Y_j^{DO} - s^{DO} = Y_0^{DO} \\
& \sum_{j=1}^{n} \lambda_j Y_j^{UO} + s^{UO} = Y_0^{UO}, \\
& \sum_{j=1}^{n} \lambda_j X_j^{DI} + s^{DI} = X_0^{DI} \\
& \sum_{j=1}^{n} \lambda_j X_j^{UI} - s^{UI} = X_0^{UI}, \\
& s^{DI}, s^{UI}, s^{DO}, s^{UO} \geq 0, \lambda \in S
\end{aligned}
\tag{15.8}
$$

It is clear this DEA model has units invariant, and the scores are between [0, 1]. However, this model is not translation invariant. It follows from our discussions above that regarding the undesirable inputs and outputs as desirable outputs and inputs and then applying the strongly free disposability will lead to the same models. If we assume the weakly free disposability instead, we can have the following model:

$$
\begin{aligned}
\min \quad & \rho = \frac{1 - \frac{1}{|DI|+|UO|}\left(\sum s_i^{DI}/x_{i0}^{DI} + \sum s_i^{UO}/y_{i0}^{UO}\right)}{1 + \frac{1}{|DO|+|UI|}\left(\sum s_i^{DO}/y_{i0}^{DO} + \sum s_i^{UI}/x_{i0}^{UI}\right)} \\
s.t. \quad & \sum_{j=1}^{n} \lambda_j X_j^{DI} + s^{DI} = X_0^{DI} \\
& \sum_{j=1}^{n} \lambda_j X_j^{UI} - s^{UI} = X_0^{UI} \\
& \sum_{j=1}^{n} \lambda_j Y_j^{DO} - s^{DO} = Y_0^{DO} \\
& \sum_{j=1}^{n} \lambda_j Y_j^{UO} - s^{UO} = Y_0^{UO} \\
& \sum_{j=1}^{n} Y_j^{DO} \lambda_j = \alpha Y_0^{DO} \\
& \sum_{j=1}^{n} Y_j^{UO} \lambda_j = \alpha Y_0^{UO} \\
& \alpha \geq 1, \lambda \in S, s^{DI}, s^{UI} \geq 0
\end{aligned}
\tag{15.9}
$$

Where the negative data for inputs or outputs is non-trivial and deserves study, the standard measure used above can be negative. Often people still prefer to use negative data in some applications, see Liu and Sharp (1999) and Sharp et al. (2006) for some discussions. Below we assume that all the inputs and outputs are desirable but can be negative. In Silva Portela et al. (2004), the SP Range was introduced, and the Range Directional Model was formulated as

$$\begin{aligned} \min \quad & \beta \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j X_j \le X_0 - \beta P_0^- \\ & \sum_{j=1}^{n} \lambda_j Y_j \ge Y_0 + \beta P_0^+ \\ & \beta \ge 0, \lambda \in S, \end{aligned} \tag{15.10}$$

where the SP Range is defined by

$$P_0^- = X_0 - Z_i, \text{ with } Z_i = \min_j x_{ij}, P_0^+ = W_r - Y_0, \text{ with } W_r = \max_j y_{rj}$$

to handle negative data. Next, $1-\beta$ gives efficiency scores. This idea can be used to handle negative data. In fact, one only needs to replace the $X_0$, $Y_0$ in the Tone's formula with $P_0^-$, $P_0^+$ to have the following DEA model:

$$\begin{aligned} \min \quad & \rho = \frac{1 - \frac{1}{m}\sum_{i=1}^{m} s_i^- / p_{i0}^-}{1 + \frac{1}{s}\sum_{r=1}^{s} s_r^+ / p_{r0}^+} \\ s.t. \quad & X_0 = \sum_{j=1}^{n} \lambda_j X_j + s^- \\ & Y_0 = \sum_{j=1}^{n} \lambda_j Y_j - s^+ \\ & \lambda \in S, s^- \ge 0, s^+ \ge 0. \end{aligned} \tag{15.11}$$

When $p_{i0}^-$, $p_{r0}^+$ are zero, the corresponding terms will be dropped from the numerator/denominator, respectively. It can be shown that the above measure is in the range [0,1] (see Liu et al. 2006). Therefore, the efficiency measure in Model (15.11) is in the range [0,1]. It is clear that this DEA model is not only units invariant but also translation invariant. The above model is applicable to the case where all inputs and outputs are desirable but may be negative. Furthermore, it is clear that the general Model (15.12) can be similarly modified so that it can handle either desirable and undesirable or positive and negative data.

## 2. DEA models with radial measurement

Now assume that all the components of inputs and outputs are positive. For the general case we again decompose the inputs and outputs into desirable and undesirable parts:

$$X_j = \begin{pmatrix} X_j^{DI} \\ X_j^{UI} \end{pmatrix}, Y_j = \begin{pmatrix} Y_j^{DO} \\ Y_j^{UO} \end{pmatrix}, X = \begin{pmatrix} X^{DI} \\ X^{UI} \end{pmatrix}, Y = \begin{pmatrix} Y^{DO} \\ Y^{UO} \end{pmatrix}$$

If we wish to use a single ratio to measure the radial extension or contraction for both desirable and undesirable parts of inputs or outputs, then we may have to address DEA models with nonlinear objective functions, such as $\theta + 1/\theta$. However, it is possible to measure desirable outputs and undesirable inputs with the radial measure by assuming the extended strongly free disposability regarding the undesirable inputs as desirable outputs. From this point of view, we can derive DEA models of radial type for undesirable inputs and outputs, as follows:

$$\begin{aligned} \max \quad & \theta + \varepsilon \left( \left| s^{DI} \right| + \left| s^{UI} \right| + \left| s^{DO} \right| + \left| s^{UO} \right| \right) \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j Y_j^{DO} - s^{DO} = \theta Y_0^{DO}, \sum_{j=1}^{n} \lambda_j X_j^{UI} - s^{UI} = \theta X_0^{UI}, \\ & \sum_{j=1}^{n} \lambda_j Y_j^{UO} + s^{UO} = Y_0^{UO}, \sum_{j=1}^{n} \lambda_j X_j^{DI} + s^{DI} = X_0^{DI}, \\ & s^{DI}, s^{UI}, s^{DO}, s^{UO} \geq 0, \lambda \in S, \theta \geq 1. \end{aligned} \tag{15.12}$$

Similarly, we can write down the following input-oriented DEA model with undesirable inputs and/or outputs:

$$\begin{aligned} \min \quad & \theta - \varepsilon \left( \left| s^{DI} \right| + \left| s^{UI} \right| + \left| s^{DO} \right| + \left| s^{UO} \right| \right), \\ s.t. \quad & \sum_{j=1}^{n} Y_j^{DO} \lambda_j - s^{DO} = Y_0^{DO}, \sum_{j=1}^{n} X_j^{UI} \lambda_j - s^{UI} = X_0^{UI} \\ & \sum_{j=1}^{n} Y_j^{UO} \lambda_j + s^{UO} = \theta Y_0^{UO}, \sum_{j=1}^{n} X_j^{DI} \lambda_j + s^{DI} = \theta X_0^{DI}, \\ & s^{DI}, s^{UI}, s^{DO}, s^{UO} \geq 0, \lambda \in S, 0 \leq \theta \leq 1. \end{aligned} \tag{15.13}$$

Now we will discuss the approaches used in Seiford and Zhu (2002). In their model, all the inputs are assumed to be desirable; however, there are undesirable outputs. These researchers first used the output transformation $\overline{Y}_j^U = -Y_j^U + w$ and, subsequently, the strongly free disposability with the radial measure to derive the

model:

$$\begin{aligned} \max \quad & \theta \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}, i = 1, \ldots, m, \\ & \sum_{j=1}^{n} \lambda_j \bar{y}_{rj}^{U} \geq \theta \bar{y}_{r0}^{U}, r = 1, \ldots, l, \\ & \sum_{j=1}^{n} \lambda_j y_{rj}^{D} \geq \theta y_{r0}^{D}, r = 1, \ldots, s, \\ & \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n. \end{aligned} \tag{15.14}$$

We know that this model uses the extended strongly free disposability, as discussed before. Often, decision-makers are more interested in the desirable outputs such that we may wish only to explicitly measure the desirable outputs in the above model; that is to say, we maximize the performance measure of desirable outputs (like total electricity generated) while asking the undesirable ones (like pollution) under control. Next, we can simply regard the undesirable outputs as desirable inputs and then have the following DEA model:

$$\begin{aligned} \max \quad & \theta \\ s.t. \quad & \sum_{j=1}^{n} \lambda_j x_{ij} \leq x_{i0}, i = 1, \ldots, m, \\ & \sum_{j=1}^{n} \lambda_j y_{rj}^{U} \leq y_{r0}^{U}, r = 1, \ldots, l, \\ & \sum_{j=1}^{n} \lambda_j y_{rj}^{D} \geq \theta y_{r0}^{D}, r = 1, \ldots, s, \\ & \sum_{j=1}^{n} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, n. \end{aligned} \tag{15.15}$$

Model (15.15) can be obtained directly from Model (15.14) by dropping the radial measures for the undesirable outputs.

### 3. **DEA models with directed-distance measurement**

Finally we discuss models using the directional distance. Let us still assume all the inputs are desirable for simplicity. Using the directional distance used in Färe and Grosskopf (2004), it is now possible to use linear measurements to measure both

desirable and undesirable variables. For example, assuming the extended strongly free disposability and CRS, one then has the following DEA model:

$$\begin{aligned} \max \quad & \theta \\ s.t. \quad & \sum_{j=1}^{n} Y_j^{DO}\lambda_j \geq Y_0^{DO} + \theta G^{DO}, \sum_{j=1}^{n} Y_j^{UO}\lambda_j \leq Y_0^{UO} - \theta G^{UO}, \\ & \sum_{j=1}^{n} X_j\lambda_j \leq X_0, \lambda \geq 0, \theta \geq 0, \end{aligned} \tag{15.16}$$

where $G^{DO}$ and $G^{UO}$ are selected references. Using the weakly free disposability discussed and the subsequent PPS instead, the model becomes

$$\begin{aligned} \max \quad & \theta \\ s.t. \quad & \sum_{j=1}^{n} Y_j^{DO}\lambda_j \geq Y_0^{DO} + \theta G^{DO}, \sum_{j=1}^{n} Y_j^{UO}\lambda_j = Y_0^{UO} - \theta G^{UO}, \\ & \sum_{j=1}^{n} X_j\lambda_j \leq X_0, \lambda \geq 0, \theta \geq 0, \end{aligned} \tag{15.17}$$

which replaces the inequality for undesirable outputs in Model (15.16) with the equality.

In practice, extra assumptions, such as null-joint of desirable and undesirable variables, may be needed (see Färe and Grosskopf (2004) for the details). If one uses $\theta$ and $\frac{1}{\theta}$ to measure performance of the desirable and undesirable outputs, respectively, and assumes extended strongly free disposability, then one will have the nonlinear model in Färe et al. (1989).

## 15.3 Two-Stage DEA Models with Undesirable Variables

In recent years, many researchers constructed various models for evaluating the efficiencies of two-stage systems. Färe (1991), Färe and Whittaker (1995), and Färe and Grosskopf (1996) advanced the frontier model and evaluated the efficiencies of 137 dairy farms in the United States. These researchers found that the resolution of the model is higher than the traditional DEA models. Liang et al. (2008) built a centralized model and leader-follower model from the perspective of game theory, in which the system efficiency is decomposed into the product of the subsystems' efficiencies; they also proposed a relatively fair scheme of efficiency decomposition. Kao and Hwang (2008) established a relational model and carefully handled the intermediate measures. Cook et al. (2010) proved that the centralized model and relational model are equivalent to the Frontier model. Zhou et al. (2013a) developed a Nash bargaining game model to obtain fair efficiency decompositions for the centralized model while

keeping the overall efficiency unchanged under this circumstance. All of these studies are multiplier models in nature. On the other hand, Färe and Grosskopf (2000) constructed production possibility sets and, consequently, built envelopment models for network DEA. Chen and Zhu (2004) built the production frontier for two-stage system and developed an integration DEA model in the envelopment form. Chen and Yan (2011) proposed three models: centralized, mixed, and decentralized production possibility sets according to the internal operation modes of two-stage systems; they established the corresponding DEA models in the envelopment form. Zhou et al. (2013b) further studied the production possibility set and performance evaluation models in supply chain DEA. Tone and Tsutsui (2009) developed a network DEA model based on the radial WSBM measurement, which can evaluate the efficiency of the system more reasonably because the importance of the subsystem is taken into account. Although some multiplier models are shown to be the dual of envelopment models, their relationships are much less clear in the contents of Network DEA (see Chen et al. (2013) for the discussions of the advantages and disadvantages of the two approaches).

However, most of the existing studies only consider desirable inputs and outputs in nature. In actual production activities, undesirable inputs and undesirable outputs may exist. There are many real applications with undesirable inputs and/or outputs. Liu et al. (2010) investigated the existing treatment on undesirable inputs/outputs for single-stage DEA using a free-disposability framework. Only a few studies focus on systems with network structures and undesirable inputs/outputs. Kordrostami and Amirteimoori (2005) considered a multistage system and took into account the undesirable factors (which can also be intermediate measures) with a minus sign in the computation of the virtual inputs and virtual outputs of a multiplier formulation. Hua and Bian (2008) extended the approach to a more general network of processes, not necessarily in series. In both papers, a multiplier DEA form is used. Fukuyama and Weber (2010) evaluated the performance of Japanese banks using a slacks-based network model where some final outputs are undesirable. Lozano et al. (2013) proposed a directional distance approach to address network DEA problems where the processes may generate not only desirable but also undesirable final outputs. The proposed approach is applied to the problem of modeling and benchmarking airport operations. Chen et al. (2012) used multi-activity network data envelopment analysis to appraise the performance of incineration plants in Taiwan. The respective efficiencies of the waste treatment and electricity generation are assessed in a unified framework.

However, emerging applications call for more systematic investigations for two-stage and network DEA with undesirable variables. For instance, depending on its operating model, whether an intermediate variable is desirable or undesirable can be questionable for a particular two-state system. Moreover, most of the existing studies only consider the final outputs to be undesirable. The characteristics of initial inputs and intermediate measures have not been investigated yet. As mentioned, there may even exist inconsistencies in deciding the types of intermediate measures. In this paper, we try to provide a more systematic investigation by following the idea of using free-disposability, which is the key element for the theoretical study of the

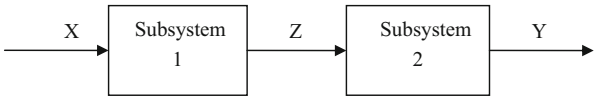



**Fig. 15.3** Two-stage systems

DEA models with undesirable variables. Therefore, we will still use envelopment framework in our study, although we are aware of the potential problems, such as division efficiency, in this approach, especially in the case of variable return to scale, see Chen et al. (2013) for more details. In this work, we build the production possibility sets and construct the corresponding envelopment DEA models with undesirable factors.

### *15.3.1 Desirability of Inputs and Outputs in Two-Stage Systems*

The two-stage system is shown in Fig. 15.3, where the whole system is composed of two subsystems connected in series. All of the outputs of subsystem 1 are the only inputs of subsystem 2. Based on the definitions of the types of inputs and outputs in Sec. 15.2, there are two viewpoints to determine the types of initial inputs, intermediate input-outputs, and final outputs in two-stage systems.

1. **Subsystem viewpoint.** According to the discussion in Sec. 15.2, at first, the types of outputs $Z$ and outputs $Y$ are determined by subsystem 1 and subsystem 2, respectively. Next, the types of inputs $X$ and inputs $Z$ are defined according to the inherent operating mechanisms of subsystem 1 and subsystem 2, respectively.
2. **System viewpoint.** The DM of the whole system defines the types of final outputs $Y$, then define the types of intermediate measures $Z$ and initial inputs $X$ in sequence according to the inherent operating mechanisms of subsystem 2 and subsystem 1.

There may exist two situations while determining the types of initial inputs, intermediate input-outputs, and final outputs.

**Consistency** The definitions of the types of inputs as well as the types of outputs are the same using both ways. In the banking system example shown in Fig. 15.4, we use the subsystem-definition approach at first. For subsystem 1, revenue and profit are desirable outputs, non-performing loans is an undesirable output, and employee, assets, and operating expenses are desirable inputs. For subsystem 2, market value and earnings per share are desirable outputs, revenue volatility is an undesirable output, revenues and profit are desirable inputs, and non-performing loans is an undesirable input. When we adopt the system-definition approach for the whole system, market value and earnings per share are desirable outputs, and revenue and profit are desirable inputs in subsystem 2 and the desirable outputs in subsystem 1. Non-performing loans is an undesirable input in subsystem 2 and an undesirable

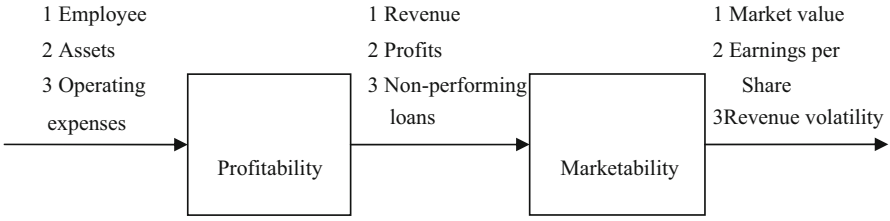



**Fig. 15.4** Banking example

output in subsystem 1. Note that the types of all indexes are the same no matter which approach is adopted.

**Inconsistency** Inconsistency means that the types of input-outputs defined by the two approaches are different. For subsystem 2, the definition of the types of inputs and outputs will always be consistent with the types by system-definition approach because the outputs of subsystem 2 are the final outputs of the whole system. So the difference of the definitions only exists in subsystem 1 or the whole system. As shown in Fig. 15.5, in the production system of power plant-fertilizer plant, $SO_2$ emissions is an undesirable output and coal and labor are desirable inputs of power plant when we use subsystem-definition method. Fertilizer production is a desirable output and power generation and $SO_2$ emissions are desirable inputs of the fertilizer plant as well as the desirable outputs of power plants, and coal and labor are desirable inputs of power plants when we use system-definition method. However, if we use the sub-system approach, the definition of the type of the $SO_2$ emissions defined by the subsystem-definition method is different from that of the system-definition method. The type of the $SO_2$ emissions is defined as undesirable by subsystem 1; however, it is defined as desirable by the overall system. Clearly there is another case that an intermediate measure is treated as desirable by subsystem 1 while undesirable by the whole system.

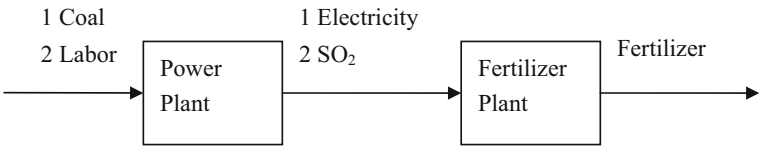



**Fig. 15.5** Power-Fertilizer plants example

### 15.3.2 Production Possibility Sets of Two-Stage Systems with Undesirable Variables

Before evaluating the efficiencies of two-stage systems, we need to establish the rational production possibility sets. Production possibility set includes not only the actual decision-maker units but also virtual units. Let us note that for two-stage production processes, the disposability situations may be different for the intermediate measures as outputs of the first stage or as inputs of the second stage. For example, the intermediate measures could have undesirable elements (like pollutants), thereby satisfying the weakly free disposability. However, these measures are all desirable inputs of the second stage, thereby satisfying the strongly free disposability as the inputs of the second stage. As another example, imagine the production of the first stage is in a small city with plenty free spaces, and thus its outputs can be assumed to be freely disposal. However, the second stage may be at a large city with very limited spaces; then the inputs of the second stage (the outputs of the first stage) may not be freely disposal. Thus, when constructing PPS for two-stage DEA models, we may assume different disposability for the intermediate variables depending on whether they are regarded as inputs or outputs of different stages.

Suppose that there are $n$ DMUs to be evaluated and that for the j-th DMU, $X_j = (X_j^{DI\,T}, X_j^{UI\,T})^T$ are the initial inputs, and $X_j^{DI} = (x_{1j}, \ldots, x_{m^D j})^T$, $X_j^{UI} = (x_{1j}, \ldots, x_{m^U j})^T$ represent the desirable and undesirable inputs of $X_j$, respectively. $Z_j = (Z_j^{DODI\,T}, Z_j^{UOUI\,T}, Z_j^{UODI\,T}, Z_j^{DOUI\,T})^T$ are the intermediate measures; that is, the outputs of subsystem 1, as well as the only inputs of subsystem 2. $Z_j^{DODI} = (z_{1j}, \ldots, z_{q^{DD} j})^T$ represent the desirable outputs of subsystem 1 and the desirable inputs of subsystem 2, $Z_j^{UOUI} = (z_{1j}, \ldots, z_{q^{UU} j})^T$ are the undesirable outputs of subsystem 1, and undesirable inputs of subsystem 2, and $Z_j^{UODI} = (z_{1j}, \ldots, z_{q^{UD} j})^T$ are the undesirable outputs of subsystem 1 and desirable inputs of subsystem 2. $Z_j^{DOUI} = (z_{1j}, \ldots, z_{q^{DU} j})^T$ are the desirable outputs of subsystem 1 and the undesirable inputs of subsystem 2. Finally, $Y_j = (Y_j^{DO\,T}, Y_j^{UO\,T})^T$ are the final outputs of the whole system; $Y_j^{DO} = (y_{1j}, \ldots, y_{s^D j})^T$, $Y_j^{UO} = (y_{1j}, \ldots, y_{s^U j})^T$ represent the desirable and undesirable outputs, respectively.

In this work we follow the idea initiated in Färe and Grosskopf (2000), and further extended in the work of Chen and Yan (2011) and Tone and Tsutsui (2009). Furthermore, we combine the idea in Liu et al. (2010)to construct the PPS and models of the network DEA when there may exist undesirable inputs, intermediate input-outputs, and outputs.

1. **Production possibility sets in the consistent case**

Usually, the production possibility set of a two-stage system is defined as

$$P = \{(X, Y) : (X, Z) \in P_1, (Z, Y) \in P_2\} \tag{15.18}$$

where $P_1$ and $P_2$ are the production possibility set of the first and second stages, respectively. This equation means $X$ is used to produce $Z$ in the first stage, and all of the products of the first stage $Z$ are used to produce $Y$ in the second stage. However, we think it is more flexible and beneficial for a systematical study to define it as:

$$P = \{(X, Z, W, Y) : (X, Z) \in P_1, (W, Y) \in P_2, (Z, W) \in \wedge\}. \tag{15.19}$$

This PPS indicates that $X$ is used to produce $Z$ in the first stage, and $W$ is used to produce $Y$ in the second stage, where $Z$ and $W$ satisfy some type of relationship$\wedge$. For example, the set $\wedge = \{(Z, W) : Z = W\}$ means that all the products of the first stage must be used by the second stages, while the $\wedge = \{(Z, W) : Z \geq W\}$ means that some products of the first stage may can be freely disposed at the second stages. Set $\wedge = \{(Z, W) : Z = \tau W, \tau \geq 1\}$ means that the products of the first stage must be used by the second stages proportionally. Of course, some more complex situations may be constructed according to the real production mechanisms.

The key point to construct two-stage production possibility sets with undesirable inputs and outputs is how to handle intermediate measures. When the types of intermediate measures are determined consistently, for the desirable part of the intermediate measures, the outputs of subsystem 1 should not be less than the inputs of subsystem 2. For the undesirable part of the intermediate measures $Z$, the outputs of subsystem 1 should not be greater than the inputs of subsystem 2. Therefore, we can construct the production possibility set below.

Under the assumption of constant return to scale (CRS), if all the initial inputs, intermediate measures, and final outputs satisfy the extended strongly free disposability, then the production possibility set can be expressed as:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \leq x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \geq x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \geq z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} \leq z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} \leq w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} \geq w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \geq y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \leq y^{UO} \\ z^{DODI} \geq w^{DODI}, w^{UOUI} \geq z^{UOUI}, \\ \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \end{array} \right\} \tag{15.20}$$

As explained in Liu et al. (2010), the extended strongly free disposability does hold for some situations and is assumed in many existing DEA models for undesirable variables. It is easy to find that many existing PPS with undesirable variables used in literatures, such as (Fukuyama and Weber 2009; Fukuyama and Mirdehghan 2012; Huang et al. 2014; Wang et al. 2014) are just special cases of (15.20) when all initial

inputs and intermediate measures are desirable while only some final outputs are undesirable.

The variable return to scale (VRS) is considerably more complicated to study for the two-stage DEA, as discussed in Chen et al. (2013). Of course if both the subsystems are VRS, then the overall system should also be VRS. However, even if one stage is CRS and another is VRS, the overall system still should be VRS. Thus it seems that the convexity constraints $\sum_{j=1}^{n} \lambda_j = 1$ and $\sum_{j=1}^{n} \mu_j = 1$ are only one possible way to impose VRS in the PPS. In this work we are not in the position to investigate the case of VRS fully. Instead we will only consider the case where both of the subsystems are VRS, as assumed in Tone and Tsutsui (2009), Lewis and Sexton (2004), Azizi and Matin (2010).

Under the CRS assumption, if all the initial inputs, intermediate measures, and final outputs satisfy the weakly free disposability, then the production possibility set can be expressed as:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} = \phi x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} = \phi x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = \alpha z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \alpha z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = \beta w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = \beta w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} = \varphi y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} = \varphi y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\ \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \\ \alpha \geq 1, \varphi \geq 1, \tau \geq 1, 0 \leq \beta \leq 1, 0 \leq \phi \leq 1 \end{array} \right\} \tag{15.21}$$

In this case, $\alpha, \beta, \varphi, \phi$ can actually be removed, the above PPS is equivalent to:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} = x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} = x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} = y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} = y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\ \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \\ \tau \geq 1 \end{array} \right\} \tag{15.22}$$

Assuming VRS for both subsystems, if all the initial inputs, intermediate measures, and final outputs satisfy the weakly free disposability, then the production possibility set can be expressed as:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} = \phi x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} = \phi x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = \alpha z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \alpha z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = \beta w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = \beta w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} = \varphi y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} = \varphi y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\ \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \\ \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \\ \alpha \geq 1, \varphi \geq 1, \tau \geq 1, 0 \leq \beta \leq 1, 0 \leq \phi \leq 1 \end{array} \right\} \tag{15.23}$$

Note that $\alpha, \beta, \varphi, \phi$ in (15.23) cannot be removed now. This property applies to the models due to the convexity constraints $\sum_{j=1}^{n} \lambda_j = 1$ and $\sum_{j=1}^{n} \mu_j = 1$. As discussed in the literature, the weakly free disposability is mostly observed in applications in environment studies, such as production of electricity with carbon dioxide. However, often the initial inputs (and/or outputs) are, in fact, desirable or, more generally, satisfy the extended strongly free disposability. These cases are studied below.

Under the CRS assumption, if initial inputs and final outputs satisfy the extended strongly free disposability, while intermediate measures satisfy the weakly free disposability, then the production possibility set can be expressed as:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \leq x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \geq x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \geq y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \leq y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\ \tau \geq 1, \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \end{array} \right\} \tag{15.24}$$

To our best knowledge, the PPS of this type is new in the literature, although the PPS in (Maghbouli et al. 2014) under the CRS assumption is similar with the above

PPS, the intensive variables in (Maghbouli et al. 2014) are the same for all initial inputs, intermediate measures, and final outputs. The above PPS can be used in many situations where intermediate measures follow weakly free disposability. For example, we can use the above PPS to analyze the airport example in (Maghbouli et al. 2014) under the CRS assumption.

Assuming VRS for both subsystems, if initial inputs and final outputs satisfy the extended strongly free disposability, while intermediate measures satisfy the weakly free disposability, then the production possibility set can be expressed as

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = \alpha z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \alpha z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = \beta w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = \beta w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\ \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1 \\ \lambda_j, \mu_j \ge 0, j = 1, \ldots, n \\ \tau \ge 1, \alpha \ge 1, 0 \le \beta \le 1 \end{array} \right\} \tag{15.25}$$

Of course, other cases can be similarly discussed.

2. **Production possibility sets in the inconsistent case**

To our best knowledge, situations below appear frequently in real applications (such as the power-fertilizer plant discussed in Sect. 15.3.1), but are never discussed in the literature.

There are two additional subclasses: (i) some intermediate outputs are undesirable for the subsystem 1, but are desirable inputs for subsystem 2;(ii) some intermediate outputs are desirable for the subsystem 1, but are undesirable inputs for subsystem 2. Clearly, if assuming that all these inconsistent measures satisfy the extended strongly free disposability, then one can infer that there will exist no linkage between these variables. For example, if $Z$ is a desirable output of subsystem 1, then any $W1 < Z$ is possible to produce by assuming the strongly free disposability. However, if it is also an undesirable input for subsystem 2, then any $W2 < Z$ is a possible input. Next, there exist no linkage between $W1$ and $W2$. Thus we should only assume weakly free or non-free disposability for those variables. We now further discuss the following useful cases.

Under the CRS assumption, if all the inconsistent variables satisfy non-free disposability, and consistent ones satisfy the extended strongly free disposability, then

the production possibility set can be expressed as:

$$\left\{ \left( \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \\ z^{DOUI} \\ z^{UODI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \\ w^{DOUI} \\ w^{UODI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \right) \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI}, \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \ge z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} \le z^{UOUI}, \\ \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} = z^{DOUI}, \sum_{j=1}^{n} \lambda_j Z_j^{UODI} = z^{UODI}, \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} \le w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} \ge w^{UOUI}, \\ \sum_{j=1}^{n} \mu_j Z_j^{DOUI} = w^{DOUI}, \sum_{j=1}^{n} \mu_j Z_j^{UODI} = w^{UODI}, \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO}, \\ z^{DODI} \ge w^{DODI} \ge 0, w^{UOUI} \ge z^{UOUI} \ge 0, \\ z^{UODI} = w^{UODI} \ge 0, w^{DOUI} = z^{DOUI} \ge 0, \\ \lambda_j, \mu_j \ge 0, j = 1, \ldots, n \end{array} \right\} \tag{15.26}$$

As mentioned before, the PPS of this type is the first instance of its discussion in literature. The situation does exist in reality. For example, in the above power-fertilizer plant example, $SO_2$emission is an inconsistent intermediate measure and electricity is a consistent one. Consequently, we can assume strongly free disposability for electricity, while non-free disposability for $SO_2$ emissions.

Under the CRS assumption, if all the initial inputs and final outputs satisfy the extended strongly free disposability and intermediate measures satisfy the weakly free disposability, then the production possibility set can be expressed as

$$\left\{ \left( \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \\ z^{DOUI} \\ z^{UODI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \\ w^{DOUI} \\ w^{UODI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \right) \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UODI} = z^{UODI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} = z^{DOUI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UODI} = w^{UODI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} = w^{DOUI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UODI} = \tau w^{UODI} \\ z^{DOUI} = \tau w^{DOUI}, z^{UOUI} = \tau w^{UOUI} \\ \tau \ge 1 \\ \lambda_j, \mu_j \ge 0, j = 1, \ldots, n \end{array} \right\} \tag{15.27}$$

If both subsystems are VRS, the production possibility set can be expressed as

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \\ z^{DOUI} \\ z^{UODI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \\ w^{DOUI} \\ w^{UODI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = \alpha z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UODI} = \alpha z^{UODI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} = \alpha z^{DOUI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \alpha z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} = \beta w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UODI} = \beta w^{UODI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} = \beta w^{DOUI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \beta w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO} \\ z^{DODI} = \tau w^{DODI}, z^{UODI} = \tau w^{UODI} \\ z^{DOUI} = \tau w^{DOUI}, z^{UOUI} = \tau w^{UOUI} \\ \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1 \\ \lambda_j, \mu_j \ge 0, j = 1, \ldots, n \\ \tau \ge 1, \alpha \ge 1, 0 \le \beta \le 1 \end{array} \right\} \tag{15.28}$$

As mentioned before, the PPS of this type is also the first instance of its discussion in the literature.

The above production possibility sets are only some special cases. According to the actual production mechanism in an application, we can construct the production possibility sets, with different assumptions on returns to scale and free disposability.

### *15.3.3 Two-Stage DEA Models with Undesirable Variables*

As summarized in Liu et al. (2010), several approaches can be used in handling undesirable variables. One of the most frequently used of these approaches is data transformation. We first examine this approach for the consistent case.

As in Seiford and Zhu (2002), by transforming $\bar{x}_{ij}^{DI} = M - x_{ij}^{UI}$, $\bar{y}_{rj}^{DO} = G - y_{rj}^{UO}$, $\bar{z}_{qj}^{DODI} = F - z_{qj}^{UOUI}$, where $M > \max_{j=1,\ldots,n}(x_{ij}^{UI})$, $F > \max_{j=1,\ldots,n}(z_{qj}^{UOUI})$, $G > \max_{j=1,\ldots,n}(y_{rj}^{UO})$, the data are $\bar{X}_j^{DI} = (\bar{x}_{1j}^{DI}, \ldots, \bar{x}_{mj}^{DI})^T$, $\bar{Y}_j^{DO} = (\bar{y}_{1j}^{DO}, \ldots, \bar{y}_{rj}^{DO})^T$, $\bar{Z}_j^{DODI} = (\bar{z}_{1j}^{DODI}, \ldots, \bar{z}_{tj}^{DODI})^T$.

Now assuming VRS for both subsystems, and the strongly free disposability for initial inputs, intermediate measures, and final outputs, the following input oriented

two-stage DEA model can be constructed:

$$
\begin{aligned}
\min \quad & \theta \\
s.t. \quad & \sum_{j=1}^{n} \lambda_j X_j^{DI} \le \theta X_0^{DI}, \sum_{j=1}^{n} \lambda_j \bar{X}_j^{DI} \le \theta \bar{X}_0^{DI} \\
& \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \ge z^{DODI}, \sum_{j=1}^{n} \lambda_j \bar{Z}_j^{DODI} \ge \bar{z}^{DODI} \\
& \sum_{j=1}^{n} \mu_j Z_j^{DODI} \le w^{DODI}, \sum_{j=1}^{n} \mu_j \bar{Z}_j^{DODI} \le \bar{w}^{DODI} \\
& \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge Y_0^{DO}, \sum_{j=1}^{n} \mu_j \bar{Y}_j^{DO} \ge \bar{Y}_0^{DO} \\
& z^{DODI} \ge w^{DODI} \ge 0, \bar{z}^{DODI} \ge \bar{w}^{DODI} \ge 0 \\
& \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \ge 0 j = 1, \ldots, n
\end{aligned}
\tag{15.29}
$$

As explained in Liu et al. (2010), this approach is equivalent to the assumption of the extended strongly free disposability. In this study, we wish to explore whether this finding is still true for the two-stage systems.

With the transformed data, the PPS for the above two-stage DEA model with the VRS and strongly free disposability reads:

$$
\left\{ \begin{pmatrix} x^{DI} \\ \bar{x}^{DI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ \bar{z}^{DODI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ \bar{w}^{DODI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ \bar{y}^{DO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j \bar{X}_j^{DI} \le \bar{x}^{DI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \ge z^{DODI}, \sum_{j=1}^{n} \lambda_j \bar{Z}_j^{DODI} \ge \bar{z}^{DODI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} \le w^{DODI}, \sum_{j=1}^{n} \mu_j \bar{Z}_j^{DODI} \le \bar{w}^{DODI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j \bar{Y}_j^{DO} \ge \bar{y}^{DO} \\ z^{DODI} \ge w^{DODI}, \bar{z}^{DODI} \ge \bar{w}^{DODI}, \\ \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \ge 0 \quad j = 1, \ldots, n \end{array} \right\}
\tag{15.30}
$$

Next, due to the VRS assumption, back to the original variables the PPS now reads:

$$\left\{ \begin{pmatrix} x^{DI} \\ x^{UI} \end{pmatrix}, \begin{Bmatrix} z^{DODI} \\ z^{UOUI} \end{Bmatrix}, \begin{Bmatrix} w^{DODI} \\ w^{UOUI} \end{Bmatrix}, \begin{pmatrix} y^{DO} \\ y^{UO} \end{pmatrix} \middle| \begin{array}{l} \sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI} \\ \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \ge z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} \le z^{UOUI} \\ \sum_{j=1}^{n} \mu_j Z_j^{DODI} \le w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} \ge w^{UOUI} \\ \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO} \\ z^{DODI} \ge w^{DODI}, w^{UOUI} \ge z^{UOUI}, \\ \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \ge 0 \quad j = 1, \ldots, n \end{array} \right\} \tag{15.31}$$

where $x_{ij}^{UI} = M - \bar{x}_{ij}^{DI}$, $y_{rj}^{UO} = G - \bar{y}_{rj}^{DO}$, $z_{qj}^{UOUI} = F - \bar{z}_{qj}^{DODI}$.

This is exact (15.20) with VRS assumption for both of the subsystems. Thus, data transformation approach is clearly justified for the consistent case. However, if there exists inconsistency for intermediate measures, then it may be difficult to use them. For example, assume $Z_j^{UODI}$ are the undesirable outputs of subsystem 1, we then need to use data transformation; however, $Z_j^{UODI}$ are also the desirable inputs of subsystem 2, and thus do not need data transformation. Thus, the data transformation method needs special care in the case of inconsistency.

Now let us examine another commonly used approach: regard undesirable inputs (undesirable outputs) as desirable outputs (desirable inputs). The advantage of this method is that there is no change in production possibility set. However, this explanation is not valid for a two-stage DEA model because the intensity variables $\lambda_j, \mu_j$ now are different in the two stages. Formally one can still use this idea (to maximize bad inputs and minimize bad outputs) to have the following input orientated two-stage DEA model, assuming the CRS and the extended strongly free disposability for both of the subsystems:

$$\begin{array}{ll} \min & \theta \\ s.t. & \sum_{j=1}^{n} \lambda_j X_j^{DI} \le \theta X_0^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge X_0^{UI} \\ & \sum_{j=1}^{n} \lambda_j Z_j^{DODI} \ge Z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} \le Z^{UOUI} \\ & \sum_{j=1}^{n} \mu_j Z_j^{DODI} \le W^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} \ge W^{UOUI} \\ & \sum_{j=1}^{n} \mu_j Y_j^{DO} \ge Y_0^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le \theta Y_0^{UO} \\ & Z^{DODI} \ge W^{DODI} \ge 0, W^{UOUI} \ge Z^{UOUI} \ge 0 \\ & \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \ge 0 j = 1, \ldots, n \end{array} \tag{15.32}$$

Below we will build two-stage DEA models using the PPS in the section above for the consistent case. We will only illustrate some examples.

1. **Slacks-based DEA model**

Using PPS (15.20) and slack measurement for input and outputs, we have the following non-oriented two-stage DEA model:

$$\min \quad \rho = \frac{1 - \dfrac{\sum_{a=1}^{m^{DI}} s_{Xa}^{DI} \Big/ x_{a0}^{DI} + \sum_{b=1}^{s^{UO}} s_{Yb}^{UO} \Big/ y_{b0}^{UO} + \sum_{c=1}^{q^{DODI}} s_{Zc}^{DODI} \Big/ z_{c0}^{DODI} + \sum_{d=1}^{q^{UOUI}} s_{Wd}^{UOUI} \Big/ z_{d0}^{UOUI}}{m^{DI} + s^{UO} + q^{DODI} + q^{UOUI}}}{1 + \dfrac{\sum_{e=1}^{s^{DO}} s_{Xe}^{DO} \Big/ y_{e0}^{DO} + \sum_{f=1}^{m^{UI}} s_{Yf}^{UI} \Big/ x_{f0}^{UI} + \sum_{g=1}^{q^{DODI}} s_{Wg}^{DODI} \Big/ z_{g0}^{DODI} + \sum_{h=1}^{q^{UOUI}} s_{Zh}^{UOUI} \Big/ z_{h0}^{UOUI}}{s^{DO} + m^{UI} + q^{DODI} + q^{UOUI}}}$$

$$\begin{aligned}
s.t. \quad & \sum_{j=1}^{n} \lambda_j X_j^{DI} + S_X^{DI} = X_0^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} - S_X^{UI} = X_0^{UI} \\
& \sum_{j=1}^{n} \lambda_j Z_j^{DODI} - S_Z^{DODI} = Z_0^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} + S_Z^{UOUI} = Z_0^{UOUI} \\
& \sum_{j=1}^{n} \mu_j Z_j^{DODI} + S_W^{DODI} = Z_0^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} - S_W^{UOUI} = Z_0^{UOUI} \\
& \sum_{j=1}^{n} \mu_j Y_j^{DO} - S_Y^{DO} = Y_0^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} + S_Y^{UO} = Y_0^{UO} \\
& S_X^{DI}, S_X^{UI}, S_Y^{DO}, S_Y^{UO}, S_Z^{DODI}, S_Z^{UOUI}, S_W^{DODI}, S_W^{UOUI} \geq 0; \\
& \sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \geq 0 j = 1, \ldots, n
\end{aligned} \tag{15.33}$$

2. **DEA model with radial measurement**

Using PPS (15.22) and radial measurement for input and outputs, we have the following input-oriented two-stage DEA model:

$$\begin{aligned}
\min \quad & \theta \\
& \sum_{j=1}^{n} \lambda_j X_j^{DI} = \theta x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} = x^{UI} \\
& \sum_{j=1}^{n} \lambda_j Z_j^{DODI} = z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = z^{UOUI} \\
s.t. \quad & \sum_{j=1}^{n} \mu_j Z_j^{DODI} = w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = w^{UOUI} \\
& \sum_{j=1}^{n} \mu_j Y_j^{DO} = y^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} = y^{UO} \\
& z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\
& \lambda_j, \mu_j \geq 0, j = 1, \ldots, n \\
& \tau \geq 1
\end{aligned} \tag{15.34}$$

3. **DEA model with directed-distance measurement**

Under the assumption of PPS (15.25), and if the direction vector is specified as $g = (g_D, -g_U)$, we can formulate the following output-oriented DEA model to handle undesirable variables, where the efficiency is defined as $1 - \beta$:

$$
\begin{aligned}
&\max \quad \beta \\
&s.t. \\
&\sum_{j=1}^{n} \lambda_j X_j^{DI} \le x^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} \ge x^{UI} \\
&\sum_{j=1}^{n} \lambda_j Z_j^{DODI} = \alpha z^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} = \alpha z^{UOUI} \\
&\sum_{j=1}^{n} \mu_j Z_j^{DODI} = r w^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} = r w^{UOUI} \\
&\sum_{j=1}^{n} \mu_j Y_j^{DO} \ge y^{DO} + \beta g_D, \sum_{j=1}^{n} \mu_j Y_j^{UO} \le y^{UO} - \beta g_U \\
&z^{DODI} = \tau w^{DODI}, z^{UOUI} = \tau w^{UOUI} \\
&\sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1 \\
&\lambda_j, \mu_j \ge 0, j = 1, \ldots, n \\
&\tau \ge 1, \alpha \ge 1, 0 \le r \le 1
\end{aligned}
\tag{15.35}
$$

Below we will build two-stage DEA models to address inconsistencies in intermediate measures.

1. **Slacks-based DEA model**

It is clear that the SBM model proposed by Tone (2001) can be directly used to handle undesirable variables, even in the case of inconsistency. Without loss of generality, we assume the VRS for both subsystems and that all of the inconsistent variables satisfy non-free disposability, and consistent ones satisfy the extended strongly free disposability. If all the slacks of initial inputs, intermediate measures, and final

outputs are included, then the following hybrid model can be constructed.

$$\min \quad \rho = \frac{1 - \dfrac{\sum_{a=1}^{m^{DI}} \frac{s_{Xa}^{DI}}{x_{a0}^{DI}} + \sum_{b=1}^{s^{UO}} \frac{s_{Yb}^{UO}}{y_{b0}^{UO}} + \sum_{c=1}^{q^{DODI}} \frac{s_{Zc}^{DODI}}{z_{c0}^{DODI}} + \sum_{d=1}^{q^{UOUI}} \frac{s_{Wd}^{UOUI}}{z_{d0}^{UOUI}} + \sum_{u=1}^{q^{UODI}} \frac{s_{Zu}^{UODI}}{z_{u0}^{UODI}}}{m^{DI} + s^{UO} + q^{DODI} + q^{UOUI} + q^{UODI}}}{1 + \dfrac{\sum_{e=1}^{s^{DO}} \frac{s_{Xe}^{DO}}{y_{e0}^{DO}} + \sum_{f=1}^{m^{UI}} \frac{s_{Yf}^{UI}}{x_{f0}^{UI}} + \sum_{g=1}^{q^{DODI}} \frac{s_{Wg}^{DODI}}{z_{g0}^{DODI}} + \sum_{h=1}^{q^{UOUI}} \frac{s_{Zh}^{UOUI}}{z_{h0}^{UOUI}} + \sum_{v=1}^{q^{DOUI}} \frac{s_{Zv}^{DOUI}}{z_{v0}^{DOUI}}}{s^{DO} + m^{UI} + q^{DODI} + q^{UOUI} + q^{DOUI}}}$$

$$s.t. \quad \sum_{j=1}^{n} \lambda_j X_j^{DI} + S_X^{DI} = X_0^{DI}, \sum_{j=1}^{n} \lambda_j X_j^{UI} - S_X^{UI} = X_0^{UI}$$

$$\sum_{j=1}^{n} \lambda_j Z_j^{DODI} - S_Z^{DODI} = Z_0^{DODI}, \sum_{j=1}^{n} \lambda_j Z_j^{UOUI} + S_Z^{UOUI} = Z_0^{UOUI}$$

$$\sum_{j=1}^{n} \lambda_j Z_j^{UODI} + S_Z^{UODI} = Z_0^{UODI}, \sum_{j=1}^{n} \lambda_j Z_j^{DOUI} + S_Z^{DOUI} = Z_0^{DOUI}$$

$$\sum_{j=1}^{n} \mu_j Z_j^{DODI} + S_W^{DODI} = Z_0^{DODI}, \sum_{j=1}^{n} \mu_j Z_j^{UOUI} - S_W^{UOUI} = Z_0^{UOUI}$$

$$\sum_{j=1}^{n} \mu_j Z_j^{UODI} + S_Z^{UODI} = Z_0^{UODI}, \sum_{j=1}^{n} \mu_j Z_j^{DOUI} + S_Z^{DOUI} = Z_0^{DOUI}$$

$$\sum_{j=1}^{n} \mu_j Y_j^{DO} - S_Y^{DO} = Y_0^{DO}, \sum_{j=1}^{n} \mu_j Y_j^{UO} + S_Y^{UO} = Y_0^{UO}$$

$$S_X^{DI}, S_X^{UI}, S_Y^{DO}, S_Y^{UO}, S_Z^{DODI}, S_Z^{UOUI}, S_W^{DODI}, S_W^{UOUI} \geq 0; \quad S_Z^{UODI}, S_Z^{DOUI} \ free\ in\ sign$$

$$\sum_{j=1}^{n} \lambda_j = 1, \sum_{j=1}^{n} \mu_j = 1, \lambda_j, \mu_j \geq 0 j = 1, \ldots, n \tag{15.36}$$

By using the Charnes-Cooper transformation, the above model can be transformed to a linear model.

2. **DEA model with radial measurement**

Under the same assumption of Model (15.33), if we only want to measure the desirable initial inputs and desirable final outputs by adopting radial measurement, then we can construct the following non-oriented DEA model:

$$\min \quad \rho = \frac{\theta}{\phi}$$

$$s.t. \quad \sum_{j=1}^{n} \lambda_j x_{i_1 j}^{DI} \leq \theta x_{i_1 0}^{DI}, \sum_{j=1}^{n} \lambda_j x_{i_2 j}^{UI} \geq x_{i_2 0}^{UI}, \quad i_1 = 1, \ldots, m^{DI}, i_2 = 1, \ldots, m^{UI}$$

$$\sum_{j=1}^{n} \lambda_j z_{d_1 j}^{DODI} \geq z_{d_1}^{DODI}, \sum_{j=1}^{n} \lambda_j z_{d_2 j}^{UOUI} \leq z_{d_2}^{UOUI}, \quad d_1 = 1, \ldots, q^{DODI}, d_2 = 1, \ldots, q^{UOUI}$$

$$\sum_{j=1}^{n} \mu_j z_{d_1 j}^{DODI} \leq w_{d_1}^{DODI}, \sum_{j=1}^{n} \mu_j z_{d_2 j}^{UOUI} \geq w_{d_2}^{UOUI}, \quad d_1 = 1, \ldots, q^{DODI}, d_2 = 1, \ldots, q^{UOUI}$$

$$\begin{aligned}
&\sum_{j=1}^{n}\lambda_j z_{d_3 j}^{UODI}=\sum_{j=1}^{n}\mu_j z_{d_3 j}^{UODI},\quad d_3=1,\ldots,q^{UODI}\\
&\sum_{j=1}^{n}\lambda_j z_{d_4 j}^{DOUI}=\sum_{j=1}^{n}\mu_j z_{d_4 j}^{DOUI},\quad d_4=1,\ldots,q^{DOUI}\\
&\sum_{j=1}^{n}\mu_j y_{r_1 j}^{DO}\geq\varphi y_{r_1 0}^{DO},\sum_{j=1}^{n}\mu_j y_{r_2 j}^{UO}\leq y_{r_2 0}^{UO},\quad r_1=1,\ldots,s^{DO},r_2=1,\ldots,s^{UO}\\
&z_{d_1 j}^{DODI}\geq w_{d_1 j}^{DODI}\geq 0,w_{d_2 j}^{UOUI}\geq z_{d_2 j}^{UOUI}\geq 0,\quad d_1=1,\ldots,q^{DODI},d_2=1,\ldots,q^{UOUI},j=1,\ldots,n\\
&\sum_{j=1}^{n}\lambda_j=1,\sum_{j=1}^{n}\mu_j=1,\lambda_j,\mu_j\geq 0,j=1,\ldots,n
\end{aligned}\tag{15.37}$$

Similarly, readers may construct DEA models to measure undesirable initial inputs and/or final outputs.

3. **DEA model with directed-distance measurement**

Finally, we can adopt the directed distance approach in the inconsistency case. Under the same assumption with model (15.33), and if the direction vector is specified as $g=(g_D,-g_U)$, we can formulate the following output-oriented DEA model to handle undesirable variables, where the efficiency is defined as $1-\beta$:

$$\begin{aligned}
\max\quad & \beta\\
s.t.\quad & \sum_{j=1}^{n}\lambda_j X_j^{DI}\leq X_0^{DI},\sum_{j=1}^{n}\lambda_j X_j^{UI}\geq X_0^{UI}\\
& \sum_{j=1}^{n}\lambda_j Z_j^{DODI}\geq z^{DODI},\sum_{j=1}^{n}\lambda_j Z_j^{UOUI}\leq z^{UOUI}\\
& \sum_{j=1}^{n}\lambda_j Z_j^{UODI}=z^{UODI},\sum_{j=1}^{n}\lambda_j Z_j^{DOUI}=z^{DOUI}\\
& \sum_{j=1}^{n}\mu_j Z_j^{DODI}\leq w^{DODI},\sum_{j=1}^{n}\mu_j Z_j^{UOUI}\geq w^{UOUI}\\
& \sum_{j=1}^{n}\mu_j Z_j^{UODI}=w^{UODI},\sum_{j=1}^{n}\mu_j Z_j^{DOUI}=w^{DOUI}\\
& \sum_{j=1}^{n}\mu_j Y_j^{DO}\geq Y_0^{DO}+\beta g_D,\sum_{j=1}^{n}\mu_j Y_j^{UO}\leq Y_0^{UO}-\beta g_U\\
& z^{DODI}\geq w^{DODI}\geq 0,w^{UOUI}\geq z^{UOUI}\geq 0,\\
& z^{UODI}=w^{UODI}\geq 0,z^{DOUI}=w^{DOUI}\geq 0\\
& \sum_{j=1}^{n}\lambda_j=1,\sum_{j=1}^{n}\mu_j=1,\lambda_j,\mu_j\geq 0 j=1,\ldots,n
\end{aligned}\tag{15.38}$$

## 15.4 Conclusions

In this paper we discuss many general approaches of DEA to handle undesirable inputs, intermediates, and outputs. The main rule is to combine the disposability assumptions and the performance metrics. We first discuss desirability to rigorously define desirable variables, and then disposability to extend the standard strongly free disposability to the cases where undesirable variables present. Then, for the single-stage case, we show that assuming extended strongly free disposability is equivalent to treating undesirable inputs and outputs as desirable outputs and inputs while assuming the standard strongly free disposability in forming the PPS. We then show it is possible to construct possible production sets by combining different disposal assumptions for the two-stage case. By combining these blocks with different performance measurements, we are able to provide a unified presentation of several classes of DEA models with undesirable inputs, intermediates, and outputs both for the single-stage and two-stage cases.

**Acknowledgement** This research is supported by the National Natural Science Foundation of China (No. 71371067, 71201158), Chinese Postdoctoral Science Foundation, Hunan Provincial Foundation for Social Sciences of China (No. 09YBB073).

## References

Ali A, Seiford LM (1990) Translation invariance in data envelopment analysis. Oper Res Lett 10:403–405

Azizi R, Matin RK (2010) Two-stage production systems under variable returns to scale technology: a DEA approach. J Ind Eng 5:67–71

Banker RD, Charnes A, Cooper WW (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 30:1078–1092

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econom 30:91–107

Charnes A, Cooper WW, Wei QL, Huhng ZM (1989) Cone ratio data envelopment analysis and multi-objective programming. Int J Syst Sci 20:1099–1118

Chen Y, Yan H (2011) Network DEA model for supply chain performance evaluation. Eur J Oper Res 213(1):147–155

Chen Y, Zhu J (2004) measuring information technology's indirect impact on firm performance. Inf Technol Manag 5(1):9–22

Chen P-C, Chang C-C, Yu M-M, Hsu S-H (2012) Performance measurement for incineration plants using multi-activity network data envelopment analysis: the case of Taiwan. J Environ Manag 93:95–103

Chen Y, Cook W, Kao C, Zhu J (2013) Network DEA pitfalls: divisional efficiency and frontier projection under general network structures. Eur J Oper Res 226:505–515

Cook WD, Liang L et al (2010) Measuring performance of two-stage network structures by DEA: a review and future perspective. Omega 38(6):423–430

Cooper WW, Seiford LM, Thanassaulis E, Zanakis SH (2004) DEA and its uses in different countries. Eur J Oper Res 154:337–344

Färe R (1991) Measuring Farrell efficiency for a firm with intermediate inputs. Acad Econ Pap 19:329–340
Färe R, Grosskopf S (1996). Productivity and intermediate products: a frontier approach. Econ Lett 50(1):65–70
Färe R, Grosskopf S (2000) Network DEA. Socio-Econ Plan Sci 34(1):35–49
Färe R, Grosskopf S (2004) Modelling undesirable factors in efficiency evaluation: comments. Eur J Oper Res 157(1):242–245
Färe R, Whittaker G (1995) An intermediate input model of dairy production using complex survey data. J Agric Econ 46(2):201–213
Färe R, Grosskopf S, Lovell CAK, Pasurka C (1989) Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach. Rev Econ Stat 71:90–98
Färe R, Grosskopf S, Noh D-W, Weber W (2005) Characteristics of a polluting technology: theory and practice. J Econom 126:469–492
Fukuyama H, Mirdehghan SM (2012) Identifying the efficiency status in network DEA. Eur J Oper Res 220:85–92
Fukuyama H, Weber WL (2009) A directional slacks-based measure of technical inefficiency. Socio-Econ Plan Sci 43:274–287
Fukuyama H, Weber WL (2010) A slacks-based inefficiency measure for a two-stage system with bad outputs. Omega 38:398–409
Golany B, Roll Y (1989) An application procedure for DEA. Omega 17:237–250
Hua Z, Bian Y (2008) Performance measurement for network DEA with undesirable factors. Int. J. Manage. Decis. Making 9:141–153
Huang C-w, Ho FN, Chiu Y-h (2014) Measurement of tourist hotels′ productive efficiency, occupancy, and catering service effectiveness using a modified two-stage DEA model in Taiwan. Omega 48:49–59
Kao C, Hwang S-N (2008) Efficiency decomposition in two-stage data envelopment analysis: an application to non-life insurance companies in Taiwan. Eur J Oper Res 185:418–429
Koopmans TC (1951) Analysis of production as an efficient combination of activities. In: Koopmans TC (ed) Activity analysis of production and allocation, Cowles commission. Wiley, New York, pp 33–97
Kordrostami S, Amirteimoori A (2005) Un-desirable factors in multi-component performance measurement. Appl Math Comput 171:721–729
Kuosmanen T (2005) Weak disposability in nonparametric production analysis with undesirable outputs. Am J Agric Econ 87:1077–1082
Kuosmanen T, Kazemi Matin R (2011) Duality of weakly disposable technology. Omega 39:504–512
Lewis HF, Sexton TR (2004) Network DEA: efficiency analysis of organizations with complex internal structure. Comput Oper Res 31:1365–1410
Liang L, Cook WD et al (2008) DEA models for two-stage processes: game approach and efficiency decomposition. Nav Res Logist 55(7):643–653
Liu WB, Sharp J (1999) DEA models via goal programming. In: Westerman G (ed) Data envelopment analysis in the public and private sector. Deutscher Universitats-Verlag, Wiesbaden
Liu WB, Sharp J, Wu ZM (2006) Preference, production and performance in data envelopment analysis. Ann Oper Res 145:105–127
Liu WB, Meng W, Li XX, Zhang DQ (2010) DEA models with undesirable inputs and outputs. Ann Oper Res 173:177–194
Lovell CAK, Pastor JT, Turner JA (1995) Measuring macroeconomic performance in the OECD: a comparison of European and non-European countries. Eur J Oper Res 87:507–518
Lozano S, Gutiérrez E, Moreno P (2013) Network DEA approach to airports performance assessment considering undesirable outputs. Appl Math Model 37(4):1665–1676
Maghbouli M, Amirteimoori A, Kordrostami S (2014) Two-stage network structures with undesirable outputs: a DEA based approach. Measurement 48:109–118
Pastor JT (1996) Translation invariance in data envelopment analysis. Ann Oper Res 66:93–102

Podinovski VV, Kuosmanen T (2011) Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. Eur J Oper Res 211:577–585
Ray SC (2004) Data envelopment analysis: theory and techniques for economics and operations research. Cambridge University, Cambridge
Scheel H (1998). Negative data and undesirable outputs in DEA. Working paper in EURO Summer Institute
Scheel H (2001) Undesirable outputs in efficiency evaluation. Eur J Oper Res 132:400–410
Seiford LM (1996) Date envelopment analysis: evolution of the state-of-the-art (1978–1998). J Product Anal 7:99–137
Seiford LM, Zhu J (2002) Modeling undesirable factors in efficiency evaluation. Eur J Oper Res 142:16–20
Sharp J, Meng W, Liu WB (2006) A modified slacks based measure model for data envelopment analysis with natural negative outputs and inputs. J Oper Res Soc 58:1672–1677
Shephard RW (1970) Theory of cost and production functions. Princeton University, Princeton
Silva Portela MCA, Thanassoulis E, Simpson G (2004) Negative data in DEA: a directional distance approach applied to bank branches. J Oper Res Soc 55:1111–1121
Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509
Tone K, Tsutsui M (2009) Network DEA: a slack-based measure approach. Eur J Oper Res 197(1):243–252
Wang K, Huang W, Wu J, Liu Y-N (2014) Efficiency measures of the Chinese commercial banking system using an additive two-stage DEA. Omega 44:5–20
Yu MM (2004) Measuring physical efficiency of domestic airports in Taiwan with undesirable outputs and environmental factors. J Air Transp Manag 10:295–303
Zhou Z, Sun L, Yang W, Liu W, Ma C (2013a) A bargaining game model for efficiency decomposition in the centralized model of two-stage systems. Comput Ind Eng 64:103–108
Zhou ZB, Wang M, Ding H, Ma CQ, Liu WB (2013b) Further study of production possibility set and performance evaluation model in supply chain DEA. Ann Oper Res 206:585–592

# Chapter 16
# Frontier Differences and the Global Malmquist Index

**Mette Asmild**

**Abstract** This chapter reviews different ways of comparing the efficiency frontiers for subgroups within a data set, specifically program efficiency, the metatechnology (or technology gap) ratio and the global frontier difference index. The latter is subsequently used to define a global Malmquist index, as well as in an alternative decomposition of the traditional Malmquist index which also considers socalled favourability and favourability change components, indicating whether individual observations are located in favourable positions in the production space based on the extent of frontier shifts they observe. The various approaches are illustrated in an empirical case of Ghanaian banks.

**Keywords** Frontier differences · Program efficiency · Metatechnology ratio · Technology gap ratio · Global Malmquist index · Malmquist index · Favourability index · Favourability change index

## 16.1 Introduction

It is often interesting to compare subgroups of observations within a data set. The subgroups can relate either to different time periods, typically for the same observations, or more generally to distinct groups of observations, like observations from different countries or with different underlying characteristics, like organizational forms. Through such comparisons it can be determined whether efficiency improves over time, are higher in one country than in another, or under one organizational form rather than another (e.g. investor owned firms vis-à-vis cooperatives).

An obvious approach to comparing subgroups might be to compare e.g. the average efficiency scores in the different groups. But if the efficiency scores that are compared are all simply measured relative to a pooled frontier, then the comparison does not distinguish between what Charnes et al. (1981) denote managerial vis-à-vis program efficiencies, that is, there is no distinction between differences between the group-specific frontiers and in the efficiencies relative to those frontiers. Here we

M. Asmild (✉)
IFRO, University of Copenhagen, Rolighedsvej 25, 1958 Frederiksberg C, Denmark
e-mail: meas@ifro.ku.dk

© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9_16

mainly focus on the former, in order to examine whether the characteristics of one subgroup provides better production possibilities than another, e.g. whether regulatory reforms have led to improved possibilities or whether the regulatory regime in one country provides superior production possibilities to that in another. Different ways of comparing the frontiers for subgroups of observations have been proposed in the literature, including program efficiency, metatechnology ratios and the global frontier difference index. The latter can furthermore be used to define the global Malmquist index of Asmild and Tam (2007). These will all be discussed in the remainder of this chapter.

## 16.2 Program Efficiency

One of the very first applications of DEA can be found in Charnes et al. (1981), which extends the work from the seminal DEA paper by Charnes et al. (1978) by, amongst other things, adding an empirical example. Where the 1978 paper actually uses the problem context of efficiency assessment of schools participating in the so-called "Follow Through" program as part of the motivation for the theoretical approach proposed in the paper (and even briefly mentions the idea of comparing program efficiency, as well as managerial efficiency, of subgroups of schools), the actual empirical analysis of this is provided in the 1981 paper.

In Charnes et al. (1981), two subgroups of observations (or DMUs) are considered: Schools participating in the Program Follow Through (PFT) experiment and those not participating, Non-Follow Through (NFT). Managerial efficiency is defined as the efficiency of a school relative to the frontier for its own subgroup (either PFT or NFT), and program efficiency subsequently defined as the efficiency of the schools relative to a (pooled) frontier constructed from the schools from both subgroups, after all the schools have first been made managerially efficient, that is, projected onto their group-specific frontiers.

To formalize, let $x_{ij}^{t}$ denote the consumption of input i by DMU j (in the previous, school j), where $i = 1, \ldots, m$ indicates the m inputs considered in the efficiency assessment, $j = 1, \ldots, n$ is the set of n observed DMUs and the superscript t denotes that the DMU belongs to subgroup t, $t = 1, \ldots, T$. Similarly $y_{rj}^{t}$ denotes the production of output r by DMU j, belonging to subgroup t and with $r = 1, \ldots, s$ indicating the s outputs included in the analysis.

Determining the input oriented managerial efficiency of Charnes et al. (1981) for $DMU_0^{t'}$ under the constant returns to scale (CRS) assumption is done by solving the following linear programming problem:

$$
\begin{aligned}
& & ME_0^{t'} &= \text{Min } \theta \\
\text{s.t.} & & \Sigma_j \lambda_j x_{ij}^{t'} &\leq \theta\, x_{i0}^{t'}, i = 1, \ldots, m \\
& & \Sigma_j \lambda_j y_{rj}^{t'} &\geq y_{r0}^{t'}, r = 1, \ldots, s \\
& & \lambda_j &\geq 0, \forall j \in t' \qquad (16.1)
\end{aligned}
$$

In program (16.1) can be seen that if the DMU under analysis, $DMU_0$, belongs to subgroup t', then it is only compared to other DMUs also belonging to subgroup t' in

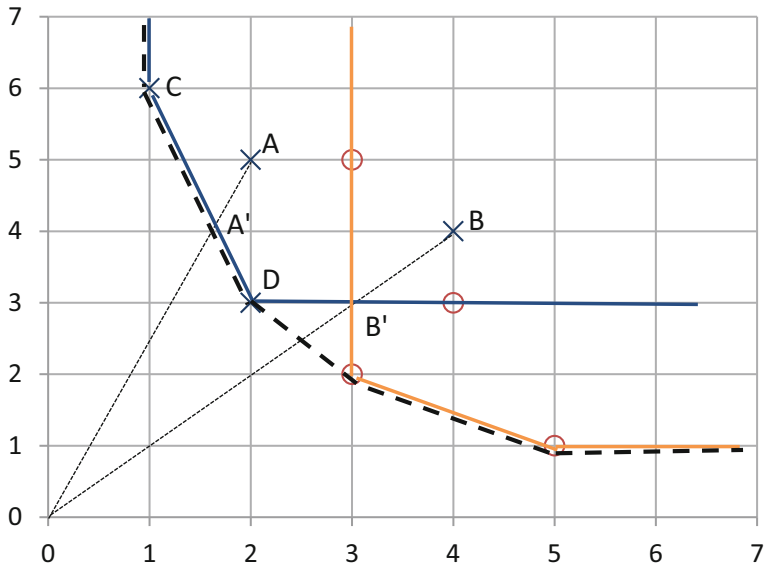



**Fig. 16.1** Illustration of managerial and program efficiencies (two inputs, fixed output)

order to assess its managerial efficiency, i.e. managerial efficiency is the efficiency of a DMU relative to the frontier of the subgroup it belongs to.

The (input oriented) program efficiency of Charnes et al. (1981) for $DMU_0^{t'}$ can now be estimated using the following program:

$$PE_0^{t'} = \text{Min } \theta$$

s.t.

$$\begin{aligned} \Sigma_j \lambda_j x_{ij} &\leq \theta(ME_0^{t'} x_{i0}^{t'}), i = 1, \ldots, m \\ \Sigma_j \lambda_j y_{rj} &\geq y_{r0}^{t'}, r = 1, \ldots, s \\ \lambda_j &\geq 0, \forall j \end{aligned} \tag{16.2}$$

In program (16.2) we see that the DMU under analysis is first made managerially efficient, by multiplying its input values with its (input oriented) managerial efficiency score, after which this transformed DMU is compared to a frontier constructed from all DMUs, that is, no longer just those belonging to its own subgroup.

It should be noted that the above is easily modified to the output orientation and/or to the case of variable returns to scale (VRS), see e.g. Cooper et al. (2004).

This 2-stage approach of first estimating managerial efficiency and then program efficiency is illustrated for a 2-input, fixed output case in Fig. 16.1 above.

In Fig. 16.1, the observations A, B, C & D, indicated by x's, all belong to one subgroup whereas the remaining observations, indicated by o's, belong to another subgroup. When assessing the managerial efficiency, we note that observation A and B are managerially inefficient, whereas observation C and D are managerially

efficient. In order to estimate program efficiency, observation A and B must first be projected onto the frontier for their own subgroup, indicated by the dark solid line, resulting in the transformed points A' and B'. Those are subsequently projected onto the pooled frontier enveloping all observations from both subgroups, indicated by the dotted line. Thus we see that observation A is managerially inefficient but program efficient, whereas observation B is both managerially and program inefficient.

## 16.3 Metafrontier Analysis and the Metatechnology Ratio

The concept of metafrontier analysis was introduced by Battese et al. (2004), albeit within the setting of Stochastic Frontier Analysis (SFA), but the subsequent paper, Battese et al. (2008), also considers the DEA version. The main idea of metafrontier analysis in DEA is defining a frontier enveloping the observations from a number of subgroups. Efficiency is then calculated relative to both the metafrontier and to the frontier of the subgroup the observation belongs to, and the ratio of these two efficiency scores is referred to as the metatechnology ratio (or technology gap ratio, or best-practice gap). This ratio indicates the distance between the frontier for the subgroup and the metafrontier, from the point of view of the observation under analysis, and thus is exactly the same as the program efficiency score of Charnes et al. (1981).

The metafrontier and program efficiency analyses in DEA rely on the use of a pooled- or meta-frontier, which assumes convexity, not just within the subgroups but also between subgroups. The latter is potentially problematic as it implies that when efficiency is measured relative to the metafrontier, this might be done relative to a benchmark constructed from observations from different subgroups which may make the interpretation and appropriateness of the benchmark, and resulting efficiency score, questionable. For example, in the Charnes et al. (1981) case a benchmark can be constructed from a combination of PFT and NFT schools, making the understanding of what best practice is, and what an inefficient school has to do to improve performance and reach the benchmark, somewhat unclear.

The interpretation of the metatechnology ratios (program efficiency scores) is generally along the lines of how far the frontier for the subgroup to which the observation under analysis belongs is behind the overall best practice indicated by the pooled- or meta-frontier, for the observation in question. These observation specific scores are typically then used to subsequently compare the subgroups in order to facilitate conclusions about the superiority of one subgroup's frontier relative to another.

Providing an overall assessment of which subgroup is superior requires a comparison of the distributions of the metatechnology ratios (program efficiency scores) between the subgroups. These are often compared using the non-parametric (Wilcoxon) Mann-Whitney rank statistic as suggested by Brockett and Golany (1996) or, for more than two groups, using the Kruskal-Wallis test, see e.g. Sueyoshi and

Aoki (2001). For a critical view of the use of these two approaches see also Simpson (2007). Alternatively, if one subgroup first-order stochastically dominates another (based on comparisons of the cumulative density functions of the efficiency scores) then the overall conclusion is straightforward (see e.g. Asmild et al. 2012).

## 16.4 Global Frontier Difference Index

If the purpose of an analysis is to provide overall conclusions about whether one subgroup is superior to another, then using the global frontier shift, more appropriately referred to as the global frontier difference, index of Asmild and Tam (2007) might be the solution. When the standard DEA based Malmquist index of Färe et al. (1994) is decomposed, a frontier shift component is provided for each DMU, which is calculated as the geometric mean of the frontier shifts associated with the DMUs location in each of the two time periods in question. These DMU-specific frontier shifts are then typically aggregated, using geometric means, and the resulting value interpreted as the overall shift of the frontier, with a value larger than 1 indicating an overall improvement of the frontier etc.

The global frontier difference index directly provides an overall measure of the difference between two frontiers, which can be frontiers for the same observations in two different time periods, in which case it is a global version of the standard frontier shift component of the DEA based Malmquist index (Färe et al. 1994) or, more generally, between any two subgroups in the data set. In the former case, the standard frontier shift can only be calculated for observations for which data are available for both of the time periods under analysis. The global index, however, utilizes the locations of all the observations in the data set to estimate the global difference, including observations from other time periods, and is furthermore not limited to balanced panel data. As shown in Asmild and Tam (2007) this provides a more accurate estimation of the overall shift. Furthermore, the global frontier difference index can estimate differences between the frontiers for any two subgroups within the data set and is thus not limited to considering shifts over time.

Formally, let the (input oriented CRS) efficiency score for $DMU_0^t$ relative to the frontier for subgroup t' be denoted by $\theta_0^{t,t'}$ and estimated by

$$\begin{aligned} \theta_0^{t,t'} &= \text{Min } \theta \\ \text{s.t.} \quad \Sigma_j \lambda_j x_{ij}^{t'} &\le \theta x_{i0}^{t}, i = 1, \ldots, m \\ \Sigma_j \lambda_j y_{rj}^{t'} &\ge y_{r0}^{t}, \quad r = 1, \ldots, s \\ \lambda_j &\ge 0, \forall j \in t' \end{aligned} \tag{16.3}$$

Next the global frontier difference index, between subgroups t' and t'', can be calculated as

$$TC^G(t', t'', X, Y) = \left( \frac{\prod_j \theta_j^{t,t'}}{\prod_j \theta_j^{t,t''}} \right)^{1/n} \tag{16.4}$$

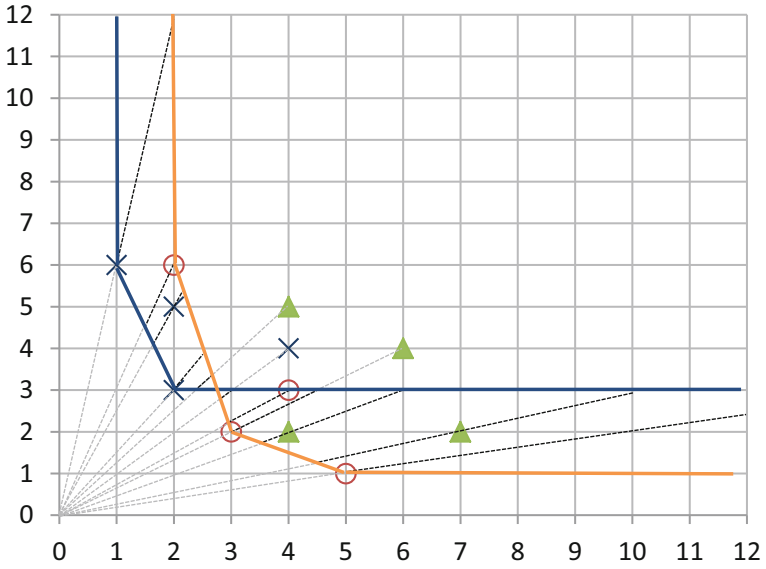



**Fig. 16.2** Illustration of the global frontier difference (two inputs, fixed output)

where X denotes the matrix of all the input vectors in the data set ($x_{ij}^t, i = 1, \ldots, m$; $j = 1, \ldots, n, \forall t$), and Y similarly the matrix of output vectors ($y_{rj}^t, r = 1, \ldots, s$; $j = 1, \ldots, n, \forall t$).

As can be seen from equation (16.4), the global frontier difference index between t' and t" is, in effect, calculated as the geometric mean of the efficiency scores for all observations in the data set estimated relative to the frontier for t' divided by the geometric means of the efficiency scores for all observations estimated relative to the frontier for t". It is here worth noting that the efficiency scores for all n observations in the data set are included in the calculation, and not just those observations belonging to t' and t". Furthermore t' and t" can denote any two subgroups in the data set, including but not limited to different time periods.

The global frontier difference index is illustrated in Fig. 16.2 above. Figure 16.2 shows a data set with observations belonging to 3 different subgroups, and where we want to estimate the global frontier difference index between the subgroup whose observations are indicated by x's and the subgroup indicated by o's. It is here important to note that also the observations belonging to the third subgroup (indicated by triangles) influence the global frontier index since, for all observations, their distance to one frontier over the distance to the other frontier factors into the calculations, as indicated by the darker dotted lines in Fig. 16.2.

An empirical example of the use of the global frontier difference index can be found in Paton et al. (2007), investigating the competitiveness of the UK electronics sector.

**Table 16.1** Global frontier differences between electronics firms (SIC30) in the UK and in the US. Source: Paton et al. (2007)

| SIC30 | UK-USA frontier difference | % inefficient units where UK frontier better than USA (%) |
|---|---|---|
| 1995 | 1.07 | 76.66 |
| 1996 | 1.04 | 71.76 |
| 1997 | 1.04 | 65.60 |
| 1998 | 0.90 | 21.23 |
| 1999 | 1.02 | 50.34 |
| 2000 | 0.43 | 0.88 |
| 2001 | 0.66 | 1.57 |
| 2002 | 0.61 | 0.51 |
| 2003 | 0.69 | 0.17 |
| 2004 | 0.68 | 0.00 |

Table 16.1 above shows the estimated global frontier differences between the UK and the US, within each of the years 1995–2004, where it should be noted that electronic firms from other countries, including Japan, are also included in the analysis and thus contribute to the estimating of the values shown. The results show how the UK frontier is superior to that of the US in the first half of the study period, but with this pattern changing such that it is the US frontier that is superior to the UK frontier in the latter half of the study period. The second column in Table 16.1 also highlights the fact that the global frontier difference is an overall measure, and even though one frontier overall is better than the other, this does not mean that the former is consistently better, as the frontiers may intersect. Specifically we see that in the earlier years, where the UK overall has the best frontier, many of the inefficient observations in the data set actually project onto sections of the frontier where it is the US frontier that is the best. We also see that for the last year of the study, the US frontier overall is superior to that of the UK and all inefficient units are, in fact, projected onto parts of the frontier where the US frontier is the best, which could indicate that the UK frontier is nested within the US frontier or, in other words, that the US frontier consistently dominates the UK frontier.

## 16.5 The Global Malmquist Index

The global frontier shift (or difference) index defined above is furthermore one of the components of the global Malmquist index of Asmild and Tam (2007), noting that this is different from the global Malmquist index proposed by Pastor and Lovell (2005), and the subsequent modifications of the latter by e.g. Oh (2010), Tohidi et al. (2012) and Tohidi and Razavyan (2013). The global Malmquist index of Pastor

and Lovell (2005) is, in fact, measured relative to a metafrontier, thus assuming convexity between all time periods in the data set (and the index for any time period is furthermore sensitive to the inclusion of additional time periods in the overall data set). Alternatively, the biennial index by Pastor et al. (2011), only assumes convexity between the two time periods for which a specific Malmquist index is calculated, though at the loss of transitivity (circularity).

That this is a global index again means that the index directly measures overall productivity change for the set of DMUs, rather than individual measures which then have to be aggregated subsequently in order to provide conclusions about the data set as a whole. As this is still a Malmquist productivity change index, the frontier difference component now relates specifically to subgroups which are different time periods, but the frontier shift component is calculated from all observations in the data set (not just those in the two time periods for which the shift is estimated) and furthermore does not necessarily require balanced panels (even though the original specification in Asmild and Tam (2007) seem to indicate so). Like the traditional Malmquist index, the global Malmquist index comprises a (global) frontier shift component as well as a (global) efficiency change component. The former is defined in (16.4) above, the latter as follows:

$$EC^G(t',t'') = \frac{\left(\prod_{j\in t''} \theta_j^{t'',t''}\right)^{1/|t''|}}{\left(\prod_{j\in t'} \theta_j^{t',t'}\right)^{1/|t'|}} \tag{16.5}$$

where |t| denotes the cardinality of (number of observations in) subgroup t.

The global Malmquist productivity change index is now defined as the product of the global frontier shift index (16.4) and the global efficiency change index (16.5), i.e.

$$\begin{aligned} MI^G(t',t'',X,Y) &= TC^G(t',t'',X,Y)\ x\ EC^G(t',t'') \\ &= \left(\frac{\prod_j \theta_j^{t,t'}}{\prod_j \theta_j^{t,t''}}\right)^{1/n} \frac{\left(\prod_{j\in t'} \theta_j^{t',t'}\right)^{1/|t'|}}{\left(\prod_{j\in t''} \theta_j^{t'',t''}\right)^{1/|t''|}} \end{aligned} \tag{16.6}$$

To illustrate the global Malmquist index and its components we utilize the electricity data set of Pastor and Lovell (2005), where 93 US electricity generating firms are observed in each of four years (1977, 1982, 1987, 1992), all using three inputs to generate a single output. We here, however, utilize an input oriented analysis (though under the maintained CRS assumption this does not really matter).

Calculating the global frontier shift component and the global efficiency change component and combining them into the global Malmquist index provides the following results:

By looking at the global frontier shift components in the second column of Table 16.2 above, we observe that the frontier worsened substantially between 1977 and 1982, worsened somewhat between 1982 and 1987 but then improved from 1987

**Table 16.2** Global frontier shifts, global efficiency changes and the global Malmquist index

| | Global frontier shift | Global efficiency change | Global Malmquist index |
|---|---|---|---|
| 1977–1982 | 0.520 | 1.163 | 0.605 |
| 1982–1987 | 0.807 | 1.088 | 0.878 |
| 1987–1992 | 1.157 | 0.930 | 1.076 |

to 1992. The actual values can be compared to the aggregated values for the best practice gaps in Pastor and Lovell (2005), that measure technical change and which are 0.589, 0.977 and 1.118 respectively. The latter show how the annual frontiers are moving relative to the metafrontier and are aggregated over the observations belonging to the two time periods considered. The global frontier shift directly measures the distances between the frontiers for the two time periods, i.e. without resorting to the use of a metafrontier, and those distances between the two frontiers in question are aggregated across all observations in the data set. The global efficiency changes are (except for round-off errors in either of the software packages employed) identical to those of Pastor and Lovell (2005), so the differences between the two versions of the global Malmquist index (the one presented here which directly defines global measures, and the aggregation of the observation specific values of the Pastor and Lovell (2005) index, which are 0.685, 1.064 and 1.039 respectively) comes from the differences in how the frontier shift is estimated.

## 16.6 Using the Global Frontier Shift Index in a Decomposition of the Standard Malmquist Index

Alternatively, the global frontier shift component can be used in a decomposition of the standard Malmquist index, considering so-called favourability components that relate to whether individual observations observe more or less frontier shift than the global (overall) shift (see Asmild and Ohene-Asare. 2012).

The standard Malmquist index of Färe et al. (1994) can be estimated from the DEA scores defined in (16.3) as

$$MI(x_0^{t'}, y_0^{t'}, x_0^{t''}, y_0^{t''}, X, Y) = \left( \frac{\theta_0^{t'',t'} \theta_0^{t'',t''}}{\theta_0^{t',t'} \theta_0^{t',t''}} \right)^{1/2} \tag{16.7}$$

and can be decomposed into

$$MI(x_0^{t'}, y_0^{t'}, x_0^{t''}, y_0^{t''}, X, Y) = \underbrace{\frac{\theta_0^{t'',t''}}{\theta_0^{t',t'}}}_{\text{EC}} \times \underbrace{\left(\frac{\prod_j \theta_j^{t,t'}}{\prod_j \theta_j^{t,t''}}\right)^{1/n}}_{\text{TC}}$$
$$\times \underbrace{F^{t',t''}\left(x_0^{t'}, y_0^{t'}, X, Y\right)}_{\text{FI}} \times \underbrace{\left[\frac{F^{t',t''}\left(x_0^{t''}, y_0^{t''}, X, Y\right)}{F^{t',t''}\left(x_0^{t'}, y_0^{t'}, X, Y\right)}\right]^{1/2}}_{\text{FCI}} \tag{16.8}$$

where

$$F^{t',t''}\left(x_0^t, y_0^t, X, Y\right) = \left(\frac{\theta_0^{t,t'}}{\theta_0^{t,t''} \times TC^G(t', t'', X, Y)}\right) \tag{16.9}$$

The first element of the decomposition in (16.8) is the standard efficiency change component (EC) of the Malmquist index (for $DMU_0$ between t' and t") and the second element is the global frontier shift, or technical change (TC), component defined in (16.4). What is new is the third and the fourth element (together with the fact that this is a decomposition of the standard Malmquist index), which are the socalled favourability index (FI) and favourability change index (FCI) respectively. The favourability index indicates the favourability of the original location of $DMU_0$ $(x_0^{t'}, y_0^{t'})$, in the sense of how big a frontier shift is observed by this DMU relative to the global shift, where an index value larger than 1 indicates that the frontier shift from the point of view of $(x_0^{t'}, y_0^{t'})$ is larger than the global shift etc. The favourability change component, in turn, indicates the change in favourability obtained by $DMU_0$ by moving from the old location $(x_0^{t'}, y_0^{t'})$ to the new location $(x_0^{t''}, y_0^{t''})$ where a value larger than one indicates that the new location observes a larger frontier shift than the old location.

## 16.7 Empirical Example: Ghanaian Banks

Finally an empirical example illustrating the use of the various concepts presented above is provided, which utilizes the data from Asmild and Ohene-Asare (2014), where different subgroups of Ghanaian banks are compared. The total of 21 banks can be divided into three subgroups based on ownership: State banks (3), domestic banks (9) and foreign banks (9). Each bank is observed in three years: 2006, 2007 and 2008. The efficiency analysis considers three inputs (fixed assets, labour and deposits) and three outputs (loans, other earning assets and corporate social responsibility expenses). For further discussion of the data and this modelling set-up please refer to Asmild and Ohene-Asare (2014).

First managerial- and program efficiencies are calculated (c.f. Eq. 16.1 and 16.2) and shown in Table 16.3 above. For simplicity and due to the small sample size, especially within the subgroups, we for now pool the observations over time, that

**Table 16.3** Average managerial- and program efficiencies

| All banks, pooled 06–08 | Mean managerial efficiency | Mean program efficiency |
|---|---|---|
| State banks | 1.05 | 1.61 |
| Domestic banks | 1.32 | 1.11 |
| Foreign banks | 1.28 | 1.10 |

**Table 16.4** Individual efficiency scores for state owned banks

| State owned banks, pooled 06–08 | Managerial efficiency | Program efficiency | Efficiency relative to meta-frontier | Ratio meta-frontier/managerial |
|---|---|---|---|---|
| ADB06 | 1.00 | 1.00 | 1.00 | 1.00 |
| GCB06 | 1.15 | 1.39 | 1.59 | 1.39 |
| NIB06 | 1.12 | 2.01 | 2.26 | 2.01 |
| ADB07 | 1.09 | 1.62 | 1.76 | 1.62 |
| GCB07 | 1.05 | 1.51 | 1.59 | 1.51 |
| NIB07 | 1.00 | 2.08 | 2.08 | 2.08 |
| ADB08 | 1.00 | 1.54 | 1.54 | 1.54 |
| GCB08 | 1.00 | 1.38 | 1.38 | 1.38 |
| NIB08 | 1.03 | 1.93 | 1.98 | 1.93 |
| MEAN | 1.05 | 1.61 | 1.69 | 1.61 |

is, across the three years of the study period. The efficiency is measured as output-oriented and assuming CRS, and the scores presented below are the output expansion factors ( $\geq 1$, with 1 indicating a technically efficient unit).

Looking at the average managerial efficiency scores in the first column of Table 16.3 above we observe that the state banks on average are the most managerially efficient, that is, closer to their group-specific frontier than the other two groups of domestic and foreign banks, which both show substantially more intra-group variation. Considering next the program efficiency in the second column it is clear, that the state owned banks are less program efficient than the other two groups of banks. In other words, the state owned banks are generally located close to their own frontier, but this is the worst of the frontiers. The variation within the other two groups means that while their respective frontiers are better than that for the state owned banks, the observations are on average further away from these frontiers.

To understand the relationship between program efficiency and the metatechnology ratio (technology gap ratio, best-practice gap), consider the individual efficiency scores within the subgroup of state owned banks, shown in Table 16.4 above:

In Table 16.4, the managerial efficiencies in the first column is where the efficiencies of the state owned banks are assessed relative to their group specific frontier, that is, the frontier spanned by the state owned banks alone. The program efficiencies in the second column are found by first projecting the state owned banks on to their

**Table 16.5** Efficiency scores for all observations relative to the three group frontiers

| | Eff. relative to state frontier | Eff. relative to domestic frontier | Eff. relative to foreign frontier |
|---|---|---|---|
| Geomean | 0.58 | 1.20 | 1.23 |

group specific frontier, i.e. eliminating the managerial efficiency, and then assessing efficiency relative to the meta-frontier constructed from all banks in the sample. In the third column are show the efficiency scores for the original observations relative to the meta-frontier and in the last column are shown the metatechnology ratios, which are the ratios of the scores relative to the meta-frontier and relative to the group frontier, and we note how these are, in fact, identical to the program efficiency scores.

Next consider the global frontier differences between the frontiers for the three subgroups, with the data still pooled over time (2006–2008). First calculate the efficiency scores for all observations but relative to the frontiers for each of the three subgroups in turn. The geometric means of these scores are shown in Table 16.5 above.

We can then calculate the global frontier difference between e.g. the state bank frontier and the domestic frontier as the ratio of the geometric mean of the scores relative to the state frontier over the geometric mean of the scores relative to the domestic bank frontier. Thus we get that:

Global frontier difference state-domestic = 0.48
Global frontier difference state-foreign = 0.47
Global frontier difference domestic-foreign = 0.97

The values above implies that the state bank frontier is around half as good as both the domestic and the foreign bank frontiers, meaning that observations located on the state frontier will, on average, only be around 50 % efficient relative to either of those two frontier. The frontiers for the domestic and the foreign banks are, on average, equally good. It should here be noted that these are mean considerations, and there are substantial variations across the individual observations. For example, one state owned bank is actually located on the meta-frontier, and therefore there will be (small) segments where the state frontier is superior to the other two frontiers. Similarly, whilst the global frontier difference between the foreign and the domestic banks is very small (as indicated by the index value close to 1), the differences for individual observations vary between the foreign frontier being almost twice as good as the domestic frontier in some locations, but the domestic frontier being more than three times as good as the foreign frontier in other locations.

In order to estimate the global Malmquist index and the favourability decomposition of the standard Malmquist index, we need to consider changes over time. Therefore, instead of pooling the data set over time as in the previous, we now consider the three time periods separately.

Calculating the global Malmquist index involves calculating the global frontier shift between the time periods (2006–2007 and 2007–2008) as well as the global efficiency changes between, as shown in Table 16.6 above.

**Table 16.6** Global Malmquist index and its components

| | Global frontier shift | Global efficiency change | Global Malmquist index |
|---|---|---|---|
| 2006–2007 | 1.07 | 0.95 | 1.01 |
| 2007–2008 | 0.94 | 1.04 | 0.97 |

**Table 16.7** Favourability decomposition of the standard Malmquist index

| 2006/2007 | MI | EC | GFS | FI | FCI |
|---|---|---|---|---|---|
| State banks | 0.87 | 0.83 | 1.07 | 0.81 | 1.20 |
| Domestic banks | 1.12 | 1.01 | 1.07 | 0.92 | 1.13 |
| Foreign banks | 0.84 | 0.93 | 1.07 | 0.70 | 1.21 |
| 2007/2008 | MI | EC | GFS | FI | FCI |
| State banks | 1.06 | 1.30 | 0.94 | 0.86 | 1.02 |
| Domestic banks | 0.95 | 1.01 | 0.94 | 0.92 | 1.09 |
| Foreign banks | 0.94 | 0.99 | 0.94 | 0.94 | 1.07 |

The global frontier shifts show that the frontier, on average, improves (by 7 %) between 2006 and 2007 but then worsens between 2007 and 2008. Conversely, the banks are, on average falling further behind the frontier from 2006 to 2007 but then catching up with the frontier between 2007 and 2008. This is not surprising, as the improving frontier makes it more likely that individual observations are falling behind and similarly that a receding frontier makes it more likely for observations to catch up. The product of the two components provide the global Malmquist index and we observe that, overall, the banks improve their productivity a little between 2006 and 2007 but then the productivity worsens (more) between 2007 and 2008.

Finally consider the favourability decomposition of the standard Malmquist index (Eq. 16.8 and 16.9), where the (geometric) mean values of the various components for each of the three subgroups are shown in Table 16.7 above.

In Table 16.7 we observe that the domestic banks on average observed productivity improvements between 2006 and 2007, coming from an almost neutral efficiency change, global frontier improvement and whilst being located in an unfavourable location in 2006 (relative to where the main frontier improvement is taking place), the observations were in 2007 located a lot more favourably (relative to the past frontier improvement). The state and foreign bank groups both, on average, experienced productivity decrease from 2006 to 2007, in spite of the globally improving frontier. This was caused by being unable to keep up with the improving frontier ("negative" efficiency change) and being located in unfavourable locations in 2006 albeit moving to more favourable locations in 2007.

Similarly, considering the bottom half of Table 16.7 we now see the state banks showing productivity improvements, mainly caused by efficiency changes. Domestic and foreign banks have productivity decreases caused by a combination of neutral efficiency change, a worsening frontier and unfavourable locations in 2007 albeit moving to better locations in 2008 (relative to where the largest frontier shift happened previously).

## 16.8 Conclusion

This chapter has reviewed different approaches to estimating differences between the efficiency frontiers for different subgroups within a data set. This idea dates back to one of the very first DEA papers and applications by Charnes et al. (1981) which introduced the notion of program efficiency, and which was subsequently relaunched as the as the metatechnology ratio (or technology gap ratio, or best-practice gap) by Battese et al. (2004, 2008). Where the program efficiency scores are specific to the individual observations and therefore have to subsequently be aggregated in some way in order to provide conclusions about the overall differences between the frontiers for the subgroups, the global frontier difference index by Asmild and Tam (2007) directly provides an overall measure for the differences between any two subgroups of observation within a data set. Furthermore, by utilizing all the observations in the data set, and not just the ones belonging to either of the two groups that are being compared, it also provides a more accurate estimate of the difference between the frontiers. Additionally, if the two frontiers being compared are for the same observations observed in different time periods, then the global frontier difference (or shift) can be used to define a global Malmquist index, which directly measures the overall productivity changes, rather than observation-specific changes which subsequently have to be aggregated. And finally, the global frontier shift index can be used in an alternative decomposition of the traditional Malmquist index, which includes a socalled favourability index, indicating whether the original location of the observation under analysis observed a frontier shift that is smaller or larger than the global shift, as well as a favourability change index indicating the change in favourability obtained by the DMU moving to its new location.

The program efficiency measure and the global frontier difference index have interesting interpretations in terms of whether one subgroup provides superior production possibilities to those for another subgroup. If the groups are different time periods, then the measures assess whether or not the possibilities improve over time. But they can also be used more generally to compare different types of observations, and thereby for example examine whether the regulatory regime in one country provides better production possibilities than that in another country, or whether a certain organizational form or management strategy is superior to another in terms of the effects on the resulting production possibilities.

The use of the global frontier shift index in a decomposition of the standard Malmquist index, including also favourability and favourability change components, may provide a valuable tool since it facilitates potentially interesting conclusions

about whether individual or groups of observations are located in, or moving towards, favourable positions, relative to where the frontier is improving most. This might be useful for e.g. policy recommendations, relating for example to which (types or groups of) DMUs are best able to take advantage of technological improvements.

## References

Asmild M, Ohene-Asare K (2014) Considering favourability indexes as part of the Malmquist index, MSAP working paper 01/2014

Asmild M, Tam F (2007) Estimating global frontier shifts and global Malmquist indices. J Prod Anal 27:137–148

Asmild M, Bogetoft P, Nielsen K (2012) Are high labour costs destroying the competitiveness of Danish dairy farmers? Evidence from an international benchmarking analysis, MSAP working paper 01/2012

Battese GE, Rao DSP, O'Donnell CJ (2004) A metafrontier production function for estimation of technical efficiencies and technology potentials for firms operating under different technologies. J Prod Anal 21:91–103

Battese GE, Rao DSP, O'Donnell CJ (2008) Metafrontier frameworks for the study of firm-level efficiencies and technology ratios. Empir Econ 34:231–255

Brockett PL, Golany B (1996) Using rank statistics for determining programmatic efficiency differences in data envelopment analysis. Manage Sci 42(3):466–472

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2:429–444

Charnes A, Cooper WW, Rhodes E (1981) Evaluating program and managerial efficiency: an application of data envelopment analysis to program follow through. Manage Sci 27(6):668–697

Cooper WW, Seiford LM, Zhu J (eds) (2004) Handbook on Data Envelopment Analysis. Taylor & Francis, Kluwer Academic Publishers, Norwell, US

Färe R, Grosskopf S, Lindgren B, Roos P (1994) Productivity developments in Swedish hospitals: a Malmquist output index approach. In Charnes A, Cooper WW, Lewin AY, Seiford LM (eds) Data envelopment analysis: theory, methodology and applications. Kluwer Academic Publishers, Boston

Oh D (2010) A global Malmquist-Luenberger productivity index. J Prod Anal 34(3):183–197

Pastor JT, Lovell CAK (2005) A global Malmquist productivity index. Econ Lett 88(2):266–271

Pastor JT, Asmild M, Lovell CAK (2011) The Biennial Malmquist Productivity Change Index. Socio-Econ Plan Sci 45:10–15

Paton D, Swann GMP, Thompson S, Girma S, Asmild M, Hanley A (2007) Competitiveness in the UK electronics sector. Report for the UK Department of Trade and Industry: Industry Economics and Statistics Division, May 2007

Simpson G (2007) A cautionary note on methods of comparing programmatic efficiency between two or more groups of DMUs in data envelopment analysis. J Prod Anal 28:141–147

Sueyoshi T, Aoki S (2001) A use of a nonparametric statistic for DEA frontier shift: the Kruskal and Wallis rank test. Omega Int J Manage Sci 29:1–18

Tohidi G, Razavyan S (2013) A circular global profit Malmquist productivity index in data envelopment analysis. Appl Math Model 37(1–2):216–227

Tohidi G, Razavyan S, Tohidnia S (2012) A global cost Malmquist productivity index using data envelopment analysis. J Oper Res Soc 63(1):72–78

# Index



© Springer Science+Business Media New York 2015

J. Zhu (ed.), *Data Envelopment Analysis,* International Series in Operations Research & Management Science 221, DOI 10.1007/978-1-4899-7553-9