Journal of Modern Applied Statistical Methods

Volume 12 | Issue 1 Article 17

5-1-2013

An Alternative Approach to Reduce Dimensionality in Data Envelopment Analysis

Grace Lee Ching Yap
The University of Nottingham Malaysia Campus, Selangor Darul Ehsan, Malaysia

Wan Rosmanira Ismail
Universiti Kebangsaan Malaysia, Selangor Darul Ehsan, Malaysia

Zaidi Isa
Universiti Kebangsaan Malaysia, Selangor Darul Ehsan, Malaysia

An Alternative Approach to Reduce Dimensionality in Data Envelopment Analysis

Cover Page Footnote
The authors would like to thank L. P. Teo for constructive advice.

Journal of Modern Applied Statistical Methods
May 2013, Vol. 12, No. 1, 128-147

1538 – 9472/13/$95.00

An Alternative Approach to Reduce Dimensionality in Data Envelopment Analysis

Grace Lee Ching Yap
The University of Nottingham Malaysia Campus,
Selangor Darul Ehsan, Malaysia

Wan Rosmanira Ismail Zaidi Isa
Universiti Kebangsaan Malaysia,
Selangor Darul Ehsan, Malaysia

Principal component analysis reduces dimensionality; however, uncorrelated components imply the existence of variables with weights of opposite signs. This complicates the application in data envelopment analysis. To overcome problems due to signs, a modification to the component axes is proposed and was verified using Monte Carlo simulations.

Key words: Data envelopment analysis, principal component analysis, redundancy analysis, Monte Carlo simulation.

Introduction

Data envelopment analysis (DEA), first introduced by Charnes, et al. (1978), serves as a tool for relative performance evaluation and benchmarking among decision making units (DMUs) with common inputs and outputs. In many circumstances, researchers may be faced with too many variables (inputs and outputs) involved in a performance measure: This will distort the discerning power of the analysis if the number of observations cannot be increased accordingly due to the curse of dimensionality (Daraio, et al., 2007). There are several approaches to increasing discrimination between observations. Based on reviews by Angulo-Meza and Lins (2002) and Podinovski and Thanassoulis (2007), the most popular approaches used are super efficiency (Andersen & Petersen, 1993) and cross-efficiency (Doyle & Green, 1994; Green, et al., 1996; Sexton, et al., 1986). These approaches do not attempt to reduce dimensionality but, by using complete information, they involve additional procedures to rank the observations. Conversely, to increase discrimination, researchers may consider keeping a reasonable dimensionality in a DEA model. Dyson, et al. (2001) indicated that the number of observations must be at least $2p \times q$ where $p \times q$ is the product of the number of inputs and outputs; thus, practitioners should be parsimonious in numbers of inputs and outputs. Although it is tempting to omit correlated variables in order to increase discrimination, Dyson, et al. (2001) showed that omitting even highly correlated variables could have a significant effect on computed efficiency scores.

Several approaches address issues of determining relevant variables, including: aggregates (Simar & Wilson, 2001), variable reduction (VR) (Jenkins & Anderson, 2003), principal component analysis (PCA-DEA) (Alder & Golany, 2001, 2002; Alder & Yazhemsky, 2010; Ueda & Hoshiai 1997), efficiency contribution measure (ECM) (Pastor, et al., 2002) and regression-based test (RB) (Ruggiero, 2005). These approaches were compared and reviewed by Sirvent, et al. (2005), Alder & Yazhemsky (2010) and Nataraja & Johnson (2011). Their analyses showed that the aggregates method requires the longest run time and its performance is not satisfactory. ECM

Grace Yap is an Assistant Professor of Applied Mathematics at the University of Nottingham Malaysia. Her research interests include data analysis, efficiency analysis and time series. Email her at: grace.yap@nottingham.edu.my. Wan Rosmanira Ismail is a Senior Lecturer in the Faculty of Science and Technology in the School of Mathematical Sciences. Email her at: wrismail@ukm.my. Zaidi Isa is a Professor in the Faculty of Science and Technology in the School of Mathematical Sciences. Email him at: zaidiisa@ukm.my.

performs moderately well under most scenarios, but it requires a long run time. The performance of RB is not as good as ECM, but its run time is significantly shorter than that of ECM. RB performs worst when variables are highly correlated; this is due to misspecification because the correlated variables would not be identified as part of the production process. Under such a scenario, PCA-DEA outperforms the other methods because it considers all original variables in the form of principal components. Most importantly, PCA-DEA involves the smallest run time due to its non-iterative characteristic. Unfortunately, PCA-DEA may not work well when data are high dimensional, meaning that some variables with weak correlation are included in the dataset. Under such a condition, these variables may cloud the principal components' dominant attributes and, consequently, the efficiency estimation is corrupted. This problem becomes less severe as the correlation between variables increases. Thus, it may be concluded that PCA-DEA is preferable when all variables are known to be relevant, and performance improves as the correlation between variables increases. In addition, PCA-DEA is robust to sample size.

Alternative to principal components, Kao, et al. (2011) proposed independent components to be used as new variables in a DEA model. The independent components are generated from independent component analysis (ICA) which is viewed as an extension of PCA in the sense that it not only de-correlates the data, but it also reduces high order statistical dependencies (Lee, 1998). However, ICA does not overcome the problem of PCA-DEA. Because PCA is popular due to its undemanding nature to reduce the dimensionality, this study focuses on the use of principal components in DEA.

PCA reorients multivariate data so that the first few dimensions account for as much of the information as possible. To be uncorrelated to each other amongst the principal components, the underlying eigenvectors must be orthogonal. This implies the existence of variables with opposite signs within a principal component because the principal components are constructed based on a mixture of positive and negative weights due to the eigenvectors. This research finds that these principal components are not suitable to replace the original variables in a DEA model as they violate the disposability assumption, consequently, meaningful efficiency estimates may not be feasible. In addition, the existence of positive and negative weights within a principal component may give rise to the problem of unboundedness in the linear program of a DEA model that uses principal components as input and/or output variables.

Although available literature does not report such a problem caused by principal components, the possibility exists for obtaining an unbounded feasible region due to the effect of positive and negative weights in the constraints of a linear program. To avoid these problems, this article proposes modifying the weights to form the principal components. As modifications to the principal components may misrepresent the original dataset, a procedure that leads to a minimal alteration is sought. The viability of such modification will be justified via a redundancy analysis whereby the proportion of explained variation in an original dataset is examined. To ascertain the motivation of such modification, the accuracy of this proposed method will be compared with the results of the standard DEA.

Reviews on Data Envelopment Analysis Model and Principal Component Analysis: Data Envelopment Analysis (DEA) Model

Data envelopment analysis (DEA) is a non-parametric method of measuring the efficiency of a decision making unit (DMU) with multiple inputs and outputs without pre-defining a production function. Following standard economic theory, the production set must be a set that contains all the input-output correspondences that are feasible in principle. The framework is similar to that in Daraio and Simar (2007), Kneip, et al. (1998), Kneip, et al. (2008) and Simar and Wilson (1998; 2000a). To illustrate, let there be a vector of p inputs, $x \in \mathrm{R}_+^p$ and a vector of q outputs $y \in \mathrm{R}_+^q$. The production set may be defined as:

$$\psi = \{(x, y) \in \mathrm{R}_+^{p+q} \mid x \text{ that can produce } y\}. \tag{2.1}$$

Specifically, the production set is assumed to be closed and strictly convex (Shephard, 1970; Fare, 1998), with the assumption of monotonicity of technology both inputs and outputs are strongly disposable. This can be described as:

$$\begin{gathered}\text{If } (x,y)\in\psi, \text{ then for any} \\ (x',y') \text{ such that } x'\geq x \text{ and} \\ y'\leq y,\ (x',y')\in\psi\end{gathered} \tag{2.2}$$

Consequently, the DMUs that are relatively efficient will lie on the production frontier. In the input orientation, the production frontier $\partial X(y)$ is defined as:

$$\partial X(y)=\{x\,|\,(x,y)\in\psi,\ (ex,y)\notin\psi,\ \forall\ 0<e<1\}. \tag{2.3}$$

Based on the efficient front of the production set, the Debreu-Farrell input measure of efficiency can be computed in a radial direction orthogonal to y, defined as follows:

$$e(x,y)=\inf\{e\,|\,(ex,y)\in\psi, e>0\} \tag{2.4}$$

In practice, with the strong disposability and constant returns-to-scale assumptions, the DEA estimator of ψ is the conical hull of the free disposal hull of an observed sample with inputs $\mathbf{X}=[x_i]$ and outputs $\mathbf{Y}=[y_i], i=1,\cdots,n$, x_i where y_i is the column vectors of p inputs and q outputs. The DEA estimator of ψ is given by

$$\hat{\psi}=\{(x,y)\,|\,y\leq\mathbf{Y}\lambda,\quad x\geq\mathbf{X}\lambda,\quad \lambda\geq 0\} \tag{2.5}$$

where λ = column vector of n non-negative variables.

The measure of efficiency is estimated using a linear programming model:

$$\hat{e}(x,y)= \min\left\{\begin{matrix} e>0|y=\mathbf{Y}\lambda-s_y,\ ex=\mathbf{X}\lambda+s_x, \\ \lambda,\ s_y,\ s_x\geq 0\end{matrix}\right\} \tag{2.6}$$

where s_y = column q-vector of output slack variables and s_x = column p-vector of input excess variables.

It is observed that the mechanism underlying this method depends largely on the constraints imposed on the model. When there are too many constraints, desirable solutions might be ruled out. In the context of DEA, this might lead to the problem of overestimating the efficiencies due to sparsity bias (Smith, 1997; Pedraja-Chaparro, et al. 1999). To avoid this problem, Simar and Wilson (2000b) suggested that the number of DMUs must increase exponentially with the addition of variables. Based on their bootstrap results, there must be at least 25 DMUs involved for the case of single input and output. For the same scenario, more than 100 DMUs are needed to have an almost exact confidence interval of the efficiency estimator. Unfortunately, this is almost impossible to achieve as large samples are generally not available in practice. This illustrates the need for discrimination improving methodologies. Because DEA is a non-parametric method, the principal component analysis (PCA) seems to be a good choice and this method has been proposed by some researchers (Ueda & Hoshiai, 1997; Alder & Golany, 2001, 2002; Alder & Yazhemsky, 2010). However, noting that PCA might violate the assumption of non-negative data in DEA, possible approaches to improve the construct of principal components for the use in DEA must be sought.

Reviews on Data Envelopment Analysis Model and Principal Component Analysis: Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical technique that reorients a dataset so that the first few dimensions account for as much information as possible. These dimensions are represented by the principal components, which are in the form of uncorrelated weighted linear combinations of the original variables that capture the maximum variance. The uncorrelated property is imposed in order to rule out the possibility of overlapped variation. These weights can be found by Eigen-decomposition, where the correlation matrix of the original set

of variables is taken as the basis for PCA. To illustrate, let there be p original standardized variables $\tilde{x}_i$ of size $n \times 1$, i =1, …, p with the matrix $\mathbf{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p]$. The correlation matrix of these variables is a $p \times p$ matrix **R**. The decomposition of the correlation matrix **R** is

$$\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathrm{T}} = \begin{bmatrix} v_1\, v_2 \cdots v_p \end{bmatrix} \begin{bmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_p \end{bmatrix} \begin{bmatrix} v_1\, v_2 \cdots v_p \end{bmatrix}^T \tag{2.7}$$

where v_j = j^{th} eigenvector of size $p \times 1$, j = 1, …, p and β_j = j^{th} eigenvalue that corresponds to v_j eigenvector, j = 1, …, p.

Note that the eigenvalues represent the explained variation the principal components, thus, they are arranged such that $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_p \geq 0$. The corresponding principal components $\mathbf{K} = [\gamma_j]^T$, with γ_j being the column vector of j principal component, j = 1, …, p are constructed based on the weights obtained from the eigenvectors:

$$\mathbf{K} = \mathbf{V}^T\mathbf{X}$$

i.e.,

$$\begin{aligned} \gamma_1 &= v_{11}\tilde{x}_1 + v_{21}\tilde{x}_2 + \cdots + v_{p1}\tilde{x}_p \\ \gamma_2 &= v_{12}\tilde{x}_1 + v_{22}\tilde{x}_2 + \cdots + v_{p2}\tilde{x}_p \\ &\vdots \\ \gamma_p &= v_{1p}\tilde{x}_1 + v_{2p}\tilde{x}_2 + \cdots + v_{pp}\tilde{x}_p \end{aligned} \tag{2.8}$$

where v_{ij} = i^{th} entry of j^{th} eigenvector, i, j = 1, …, p.

For the purpose of dimension reduction, Kaiser's rule is typically followed to choose the principal components whose eigenvalues are greater than 1; otherwise, an elbow in the Scree plot may be identified to determine the number of components to be retained. In the context of DEA, Adler and Yazhemsky (2010) noted that it is ideal to drop the principal components one-by-one until a reasonable level of discrimination is achieved or until the principal components capture at least 80% of the variance of the original data. These principal components are then used to replace the targeted inputs or outputs in the DEA model. Adapting to the additive DEA model with constant returns-to-scale (CRS) of Charnes, et al. (1985), a mixture of original data and principal components may be used to arrive at the additive model as described by Adler and Yazhemsky (2010). Equivalently, the model can be written in the form of input oriented, CRS, radial linear program as in equation (2.6).

Contrast Variables in Principal Components

Because the eigenvectors are orthogonal, there must be a mixture of positive and negative entries v_{ij}, i, j = 1, …, p within them. To illustrate, consider the first eigenvector to be $v_1 = \langle v_{11}\, v_{21} \cdots v_{p1} \rangle^T$. Even if v_1 has all positive entries, note that in order to be orthogonal to v_1, the second eigenvector $v_2 = \langle v_{12}\, v_{22} \cdots v_{p2} \rangle^T$ must satisfy the equation:

$$\begin{aligned} & v_1 \cdot v_2 = 0 \\ & \text{i.e.,} \\ & v_{11}v_{12} + v_{21}v_{22} + \cdots + v_{p1}v_{p2} = 0 \end{aligned} \tag{2.9}$$

Thus, it is straightforward to conclude that $v_2 = \langle v_{12}\, v_{22} \cdots v_{p2} \rangle^T$ consists of a mixture of positive and negative entries, for example $v_{12}, v_{22} > 0$ and $v_{32}, \cdots, v_{p2} < 0$.

For the corresponding principal component $\gamma_2 = v_{12}\tilde{x}_1 + v_{22}\tilde{x}_2 + \cdots + v_{p2}\tilde{x}_p$, the variables $\tilde{x}_1$ and $\tilde{x}_2$ are in contrast with the other variables $\tilde{x}_3$, … , $\tilde{x}_p$ as $\tilde{x}_1$ and $\tilde{x}_2$ correlate positively with γ_2 but $\tilde{x}_3$, … , $\tilde{x}_p$ correlate negatively with γ_2. To use principal components in a DEA model it is good to avoid variables with counter effect within a principal component. To simplify the label, the group of variables that capture a smaller portion of sum of squared loadings (SSL) of a principal

component are called contrast variables. Particularly for this illustration, the proportion of SSL for $\{\tilde{x}_1, \tilde{x}_2\}$ in γ_2 is $\text{SSL}_{\gamma_2(+)} = \frac{v_{12}^2 + v_{22}^2}{v_{12}^2 + v_{22}^2 + \ldots + v_{p2}^2}$. Thus, if $\text{SSL}_{\gamma_2(+)} < \frac{1}{2}$, then $\tilde{x}_1$ and $\tilde{x}_2$ are the contrast variables in γ_2, and they are to be avoided in the construct of γ_2.

In a very unfortunate (and unlikely) case if $\text{SSL}_{\gamma_2(+)} = \frac{1}{2}$, the contrast variables may be classified to the group $\{\tilde{x}_1, \tilde{x}_2\}$ or $\{\tilde{x}_3, \cdots, \tilde{x}_p\}$ that consists of the variables that have not been labeled as contrast variables in other principal components; this is to minimize the loss of information when the components are used to replace the original variables in a DEA model. To secure orthogonality, there must be contrast variables in the subsequent principal components $\gamma_3, \cdots, \gamma_p$, and the contrast variables may be any of the original variables $\{\tilde{x}_1, \cdots, \tilde{x}_p\}$. In other words, the contrast variables cannot be identified prior to PCA and they are not the same from one principal component to another; thus, the contrast variables are classified per principal component based on the sign of the entries in the eigenvector and they are not a cluster of variables that have diverse characteristics from the other variables in the dataset as a whole.

Problems of Principal Components in DEA

With the counter effect due to contrast variables, a component score can be minimized by increasing the variables that are assigned with negative weights. Hence, it cannot be interpreted that the bigger the values of the original variables, the bigger the principal component score or vice versa. This implies that the principal components violate the free disposability assumption of a DEA model as described in equation (2.2). As a result, efficiencies cannot be meaningfully estimated because the measures of efficiency rely on estimating maximum output levels for given input levels, or alternatively, minimum input levels for given output levels (Thanassoulis, 2001). In addition, the counter effect may lead to the problem of unboundedness in the linear program. To illustrate the problem, let there be m principal components $\mathbf{K}^* = [\gamma_j]^T, j = 1, \cdots, m$ replacing all p original input variables, with the other conditions remains the same as in equation (2.6). The linear program for DMU_0 with data (x_0, y_0) is then in the form:

$$\begin{aligned} &\text{Minimize } e \\ &\text{Subject to} \\ &\mathbf{Y}\lambda - s_y = y_0 \\ &\mathbf{K}^*\lambda + \mathbf{V}^{*T} s_x = e k_0^* \\ &\lambda, s_y, s_x \geq 0 \\ &\text{where} \\ &\mathbf{V}^* = [v_1\, v_2 \cdots v_m] \\ &k_0^* = \mathbf{V}^{*T} x_0 \end{aligned} \tag{2.10}$$

Note that the constraints in terms of the principal components can be restructured as follows:

$$\begin{aligned} &\mathbf{K}^*\lambda + \mathbf{V}^{*T} s_x = e k_0^* \\ &\Rightarrow (\mathbf{V}^{*T}\mathbf{X})\lambda + \mathbf{V}^{*T} s_x = e(\mathbf{V}^{*T} x_0) \\ &\Rightarrow \mathbf{V}^{*T}(\mathbf{X}\lambda + s_x) = \mathbf{V}^{*T} e\, x_0 \end{aligned} \tag{2.11}$$

To simplify the notation, let $\mathbf{T} = [t_1\, t_2 \cdots t_p]^T = \mathbf{X}\lambda + s_x$ and $x_0 = [x_{1_0}\, x_{2_0} \cdots x_{p_0}]$. By using the notations in equation (2.8), constraints in equation (2.11) can be written as:

$$\begin{cases} \dfrac{1}{(v_{11}x_{1_0} + v_{21}x_{2_0} + \cdots + v_{p1}x_{p_0})}(v_{11}t_1 + v_{21}t_2 + \cdots + v_{p1}t_p) = e \\ \dfrac{1}{(v_{12}x_{1_0} + v_{22}x_{2_0} + \cdots + v_{p2}x_{p_0})}(v_{12}t_1 + v_{22}t_2 + \cdots + v_{p2}t_p) = e \\ \vdots \\ \dfrac{1}{(v_{1m}x_{1_0} + v_{2m}x_{2_0} + \cdots + v_{pm}x_{p_0})}(v_{1m}t_1 + v_{2m}t_2 + \cdots + v_{pm}t_p) = e \end{cases} \tag{2.12}$$

Based on equation (2.10) and the requirement that $x \in \mathbf{R}_+^p$, note that $t_k \geq 0$, $k = 1, ..., p$. Thus, when all the weights v_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, m$ in equation (2.12) are the same sign (positive or negative), the linear program

produces a meaningful solution because the feasible region is bounded (≥ 0), and an optimal e^* can be obtained to minimize the objective function in equation (2.10). However, when there are positive and negative weights within a constraint, the problem of unboundedness may arise. This problem occurs when there is at least a variable x_u with moderately large weights $\langle v_{u1}, v_{u2}, \cdots, v_{um} \rangle$ of which the weights are in the opposite sign with the weights of another variable x_s that are moderately large $\langle v_{s1}, v_{s2}, \cdots, v_{sm} \rangle$ giving the product:

$$(v_{uj})(v_{sj}) < 0 \quad \forall j = 1, \cdots, m \tag{2.13}$$

The effect on the constraints in equation (2.12) is illustrated by equation (2.14) shown in Figure 1.

Note that when the weights vectors $\langle v_{u1}, v_{u2}, \cdots, v_{um} \rangle$ and $\langle v_{s1}, v_{s2}, \cdots, v_{sm} \rangle$ are dominating, and equation (2.13) is met, the values on the left hand side of equation (2.14) can be made zeros (or even negative) by loading huge input excess for, $\tilde{x}_s$ and/or $\tilde{x}_u$, , namely, $s_{x(s)}$ and $s_{x(u)}$. This will inflate the magnitudes of t_u and/or t_s, and subsequently cause the feasible region to be unbounded, of which e can be made as small as possible. In other words, this gives an unbounded solution to the objective function in equation (2.10). In order to meet the free disposability assumption and to avoid the problem of unboundedness in linear program, it is crucial to ensure that the weights assigned to the variables are non-negative.

Methodology

As weights are extracted from the eigenvectors, modifications to the eigenvectors are needed to avoid the problems of contrast variables. Nonetheless, changes made to the eigenvectors may hamper the components' potential to represent the original dataset. To provoke minimal alteration to the eigenvectors, it would be good to work on the simple structure produced by a varimax rotation; that is, an orthogonal rotation of the factor axes that maximizes the variance of the squared loadings on all the variables in a factor matrix (Kaiser, 1958). As a result, each factor tends to have a few high loadings with the rest of the loadings being zero or close to zero, leading to a simple structure, where ideally each item is loaded on only one axis (Kline, 2002). Traditionally, based

Figure 1: Effect on the Constraints in Equation (2.12)

$$\begin{cases} \dfrac{1}{(v_{11}x_{1_0} + v_{21}x_{2_0} + \cdots + v_{p1}x_{p_0})}(v_{11}t_1 + \cdots + v_{s1}t_s + \cdots + v_{u1}t_u + \cdots + v_{p1}t_p) = e \\ \dfrac{1}{(v_{12}x_{1_0} + v_{22}x_{2_0} + \cdots + v_{p2}x_{p_0})}(v_{12}t_1 + \cdots + v_{s2}t_s + \cdots + v_{u2}t_u + \cdots + v_{p2}t_p) = e \\ \vdots \\ \dfrac{1}{(v_{1m}x_{1_0} + v_{2m}x_{2_0} + \cdots + v_{pm}x_{p_0})}(v_{1m}t_1 + \cdots + v_{sm}t_s + \cdots + v_{um}t_u + \cdots + v_{pm}t_p) = e \end{cases} \tag{2.14}$$

on the simple structure, only variables with loadings above a cutoff point (for example, 0.5) are interpreted (Jolliffe, 2002). Component scores computed with such simple weighting schemes often hold up better under cross-validation compared to the exact component scores (Dunteman, 1989). By having the advantage to omit the variables with small loadings, it would be possible to restructure the weighting vectors with minimal perturbation.

To start, a varimax rotation is performed on the loadings matrix in order to obtain the simple structure $\mathbf{V}^r = [v_1^r\ v_2^r \cdots v_m^r]$. From the simple structure, dominating variables can be identified, whereby the variables with high loadings exhibit strong correlations with a principal component. In order to avoid counter effects within a component, for each component axis $v_j^r, j = 1,...,m$ the variables with positive loadings should be segregated from those with negative loadings. For illustrative purposes, the groups are labeled as positive group $v_j^{r(+)}$ and negative group $v_j^{r(-)}$. The explained variation associated to each group is depicted by the corresponding SSL, that is, $\mathrm{SSL}(v_j^{r(+)})$ and $\mathrm{SSL}(v_j^{r(-)})$.

To minimize deviations from the original principal components, the group of variables that capture a bigger portion of explained variation (the one with a larger SSL) will be extracted. Variables of another group with smaller SSL are labeled as the contrast variables. These variables are relatively less significant and are subject to be dropped: this is equivalent to assigning a zero weight to each of the contrast variables. To satisfy the requirement of unit vector (Hand, 2001), these vectors are then normalized, and hence are called the modified principal directions. The absolute values of these modified principal directions are taken to form the new weights for the construction of the modified components. The modifications can be performed with MATLAB, and the steps are described in algorithmic form as:

1. Launch varimax rotation, obtain the rotational matrix, $\mathbf{\Lambda}$.

2. Obtain the rotated component axes, that is $\mathbf{V}^r = [v_1^r\ v_2^r \cdots v_m^r] = \mathbf{V}^*\mathbf{\Lambda}$.

3. Divide the entries in each vector v_j^r into two groups, one with positive sign $v_j^{r(+)}$, and another with negative sign $v_j^{r(-)}, j = 1,\ldots,m$.

4. In each vector v_j^r, identify the group that has a bigger SSL, v_{D_j} (e.g. $v_{D_j} = v_j^{r(-)}$).

5. Normalize the vectors $v_{D_j}, j = 1,\cdots,m$.

6. Take the absolute values on the principal directions formed in step (5), giving the modified axes matrix:

$$\mathbf{W} = \begin{bmatrix} \omega_{11} & \cdots & \omega_{p1} \\ \omega_{12} & \cdots & \omega_{p2} \\ \vdots & & \vdots \\ \omega_{1m} & \cdots & \omega_{pm} \end{bmatrix}$$
$$\omega_{ij} \geq 0 \text{ for } i = 1,\cdots, m, j = 1,\cdots, p \tag{3.1}$$

7. Form the modified components $\mathbf{C} = [c_1\ c_1 \cdots c_m]^T$ based on the weights in equation (3.1):

$$\begin{aligned} c_1 &= \omega_{11}x_1 + \omega_{21}x_2 + \cdots + \omega_{p1}x_p \\ c_2 &= \omega_{12}x_1 + \omega_{22}x_2 + \cdots + \omega_{p2}x_p \\ &\vdots \\ c_m &= \omega_{1m}x_1 + \omega_{2m}x_2 + \cdots + \omega_{pm}x_p \end{aligned} \tag{3.2}$$

Simply stated, this modification only involves the exclusion of a less significant group of variables. Alternatively, to avoid negative weights, other options may be considered, such as: (1) taking the squared values on the eigenvectors, or (2) taking the absolute values of the eigenvectors. The option that best fits the original dataset should capture the most amount of explained variation in the original data.

To compare the options graphically, a specific case with 3 variables that can be

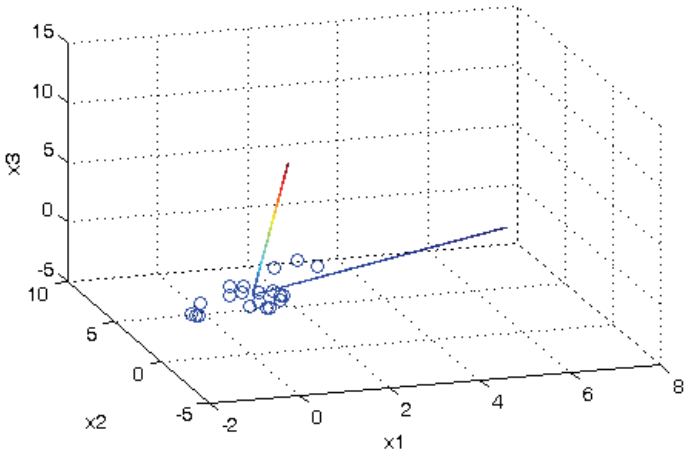
explained by two principal components is used. Figure 2(a) shows how the eigenvectors capture the distribution of the data. Using the same set of data, the modified axes from the proposed model and the other options (1) squaring the entries of the eigenvectors and (2) taking absolute values of the eigenvectors are shown in Figures 2(b), 2(c) and 2(d) respectively. Note that the proposed model gives the nearest approximation to the original eigenvectors, hence capturing almost the same amount of explained variation in the original data. To consolidate the justification, the amount of explained variation will be verified via redundancy analysis.

Justification of Modifications

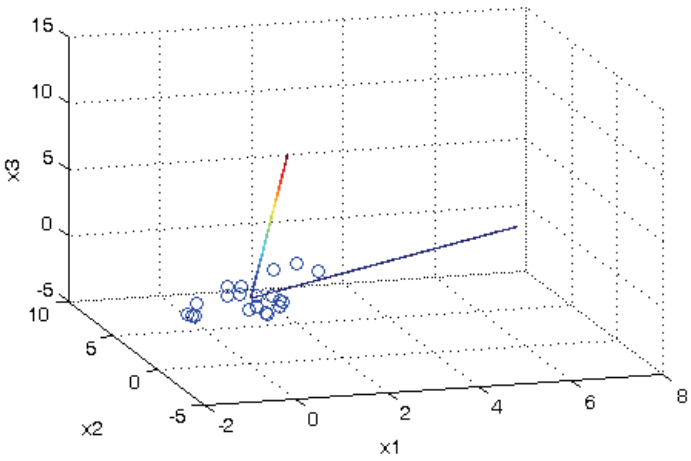
The aim of the proposed modifications is to avoid the contrast variables in principal components without much sacrifice to the ability to represent the original data. To examine this aspect, redundancy analysis (Van den Wollenberg, 1977) is used. This procedure aims to extract factors from the set of dependent variables $\tilde{\mathbf{Y}}$ that are the most predictive of the independent variables $\tilde{\mathbf{X}}$. Because interest lies in knowing how much of the variance in the original variables is explained by the modified components, let the modified components be the dependent variables, $\tilde{\mathbf{Y}}$, and the original

Figure 2: A Comparison between Eigenvectors and the Modified Directions

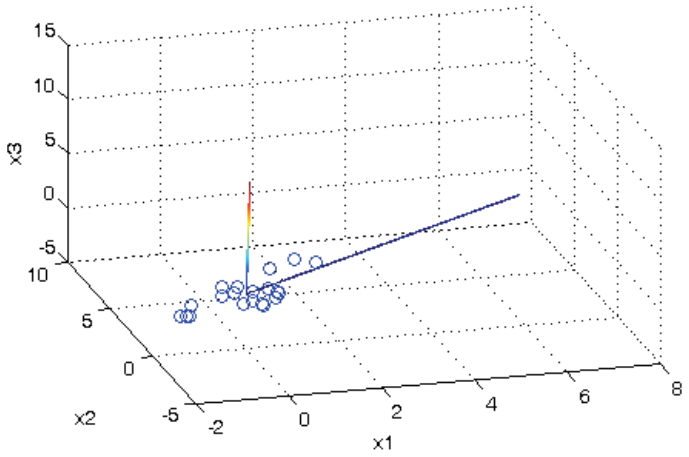
(a) Eigenvectors of Principal Components



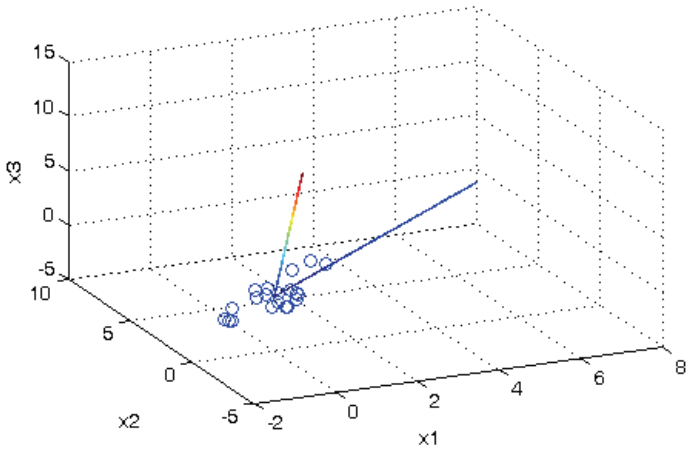
(b) Principal Directions of Proposed Modification



(c) Principal Directions of Squared Modification



(d) Principal Directions of Absolute Value Modification



variables be the independent variables, $\tilde{\mathbf{X}}$. Based on the objective of canonical correlation analysis (Hotelling, 1936), two sets of canonical variates, $\mathbf{u}_x = [u_{x_j}]$ and $\mathbf{u}_y = [u_{y_j}], j = 1, \cdots, m$ are constructed to represent $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ respectively, such that the correlation between the canonical variates, $r_j(u_{xj}, u_{yj})$, $j = 1, \cdots, m$ is maximized. Based on the canonical correlations, the proportion of variation in $\tilde{\mathbf{X}}$ being explained by $\tilde{\mathbf{Y}}$ can be computed using the redundancy index developed by (Stewart and Love 1968):

$$rd_{y \to x} = \sum_{j=1}^{m} \left(\sum_{i=1}^{p} \frac{a_{x_{i,j}}^2}{p} \right) r_j^2 \tag{3.3}$$

where $a_{x_{i,j}}$ = canonical loadings.

To compare the proposed modification to the other two options, redundancy analysis will be carried out on all the methods. The option causing the least perturbations to the eigenvectors should largely retain the proportion of explained variation, which will then be indicated by a largest redundancy.

Modified PC-DEA

After the modification that captures the largest redundancy is identified, the modified PC-DEA model can be constructed based on the modified axes and the corresponding components. To simplify the notation, assume that the proposed modification gives the largest redundancy. Thus, the modified components $\mathbf{C}$ and the modified axes $\mathbf{W}$ will be used to replace the principal components and the eigenvectors in equation (2.10). In essence, the modified PC-DEA model for DMU_0 with data (x_0, y_0) is as follows:

$$\begin{aligned} &\text{Minimize } e \\ &\text{Subject to} \\ &\mathbf{Y}\lambda - s_y = y_0 \\ &\mathbf{C}\lambda + \mathbf{W}^T s_x = e\, c_0 \\ &\lambda, s_y, s_x \geq 0 \\ &\text{where } c_0 = \mathbf{W}^T x_0 \end{aligned} \tag{3.4}$$

As shown in (3.4), the modified PC-DEA is similar to PCA-DEA, except changing the eigenvectors to the modified axes. Thus, the modified PC-DEA is suitable for the scenarios that are favorable to PCA-DEA, particularly when all the variables are known to be relevant in the production function under study. The modification can be obtained by running MATLAB codes that execute steps 1-6 described earlier. Because these steps are not heavy, the inclusion of them in a computer program would not increase the run time, and hence would preserve the strength of having the shortest run time amongst the alternatives to reduce the dimensionality. In other words, by having a better data reconstruction that avoids the problem of unboundedness in a linear program, the modified PC-DEA improves the use of principal components in a DEA model, and it offers a convenient alternative to dimension reduction.

Results

To demonstrate the problem of contrast variables within principal components in the DEA framework, the data generation process (DGP) based on the idea of Kneip, et al. (1998) and Simar and Wilson (1998, 2000a, 2001) were followed where each DMU is attached with single output efficiency and no DMU is regarded as strictly efficient. However, DEA identifies the estimates of relative efficiency. By definition, at least one DMU will be identified as relatively efficient. To mitigate the need of large sample size, it is necessary to restrict to CRS because when the boundary of the production set displays constant returns-to-scale, the DEA estimators converge faster and, hence, introduce less noise (Daraio & Simar, 2007). Each DMU_k is associated with an inefficiency index, τ_k, which is drawn independently from an inefficiency distribution. Following the criteria set by Alder and Yazhemsky (2010), a DMU is deemed relatively efficient if the simulated $e^{-\tau}$ is greater than 0.9.

To emphasize the problem of discriminatory power, consider cases with relatively many input variables compared to the number of DMUs and begin with a numerical illustration that consists of 20 DMUs that use 7

inputs to produce an output. Correlated input variables $\tilde{x}_j, j = 1, \cdots, 7$ are generated by post-multiplying a set of random numbers from a uniform distribution on the interval (0, 100) by the upper triangular Cholesky decomposition of a pre-assigned correlation matrix $\mathbf{R}_1$ with moderate pairwise correlation ($r < 0.6$). These input variables are used in a Cobb-Douglas production function $\tilde{y} = \prod_{j=1}^{7} (\tilde{x}_j)^{\frac{1}{7}}$. An inefficiency index is simulated for each DMU independently from a half normal distribution, that is, $\tau_k \sim$ HN(0,1). Under CRS, the inefficiency parameter can be assigned to either input side or the output side, as they produce the same efficiency score. In this example, the output values are calculated based on the equation $\tilde{y} = \prod_{j=1}^{7} (\tilde{x}_j)^{\frac{1}{7}} \cdot e^{-\tau}$, and the data for 20 DMUs are generated as shown in Table 1(a). The correlation matrix for the input variables is shown in Table 1(b).

To reduce dimensionality, PCA is applied to all the input variables. Four principal components were extracted in order to retain at least 80% explained variation. These components are then taken for efficiency estimations using equation (2.10). The component scores are shown in the first 4 columns of Table 2 and the eigenvectors are the first 4 rows of Table 3. From the eigenvectors, observe that the weights attached to variables $\tilde{x}_4$, $\tilde{x}_5$ and $\tilde{x}_7$ are dominant and a combination of these weights will cause the feasible region to be unbounded. To illustrate, following equation (2.12), the constraints relating to the principal components for the efficiency estimation for DMU1 are:

Table 1: Simulated Data and Correlation Matrix for Input Variables

(a) Simulated Data for $\tilde{y} = \prod_{j=1}^{7} (\tilde{x}_j)^{\frac{1}{7}} \cdot e^{-\tau}$

DMU	$\tilde{y}$	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$e^{-\tau}$
DMU1	26.260	75.793	89.197	73.386	88.115	0.201	123.52	73.210	0.728
DMU2	9.977	93.371	67.702	101.92	91.872	19.778	74.313	10.790	0.194
DMU3	48.509	15.580	74.798	80.262	97.408	36.999	143.91	17.229	0.961
DMU4	18.868	14.366	65.858	94.079	64.095	17.039	84.701	66.670	0.397
DMU5	26.997	9.258	78.416	88.645	65.644	49.054	141.87	9.075	0.630
DMU6	17.032	34.792	40.767	47.370	86.329	9.530	97.046	3.994	0.569
DMU7	5.631	37.114	16.970	53.378	37.088	65.351	73.318	10.042	0.163
DMU8	16.999	41.154	90.177	87.628	53.287	15.536	116.66	69.836	0.293
DMU9	11.283	66.556	76.222	28.593	54.421	3.057	48.587	58.638	0.319
DMU10	10.045	80.534	36.821	78.626	135.93	5.326	135.96	15.306	0.225
DMU11	11.785	24.151	25.517	61.538	85.657	15.639	115.45	13.299	0.328
DMU12	15.525	64.394	33.103	105.25	102.50	41.122	84.464	54.637	0.243
DMU13	8.922	86.984	30.712	86.334	102.49	7.062	96.267	14.978	0.211
DMU14	7.937	32.015	51.944	60.220	103.44	15.248	129.60	23.887	0.170
DMU15	8.212	21.021	16.192	99.012	112.37	54.485	110.26	12.282	0.190
DMU16	5.920	33.092	38.637	40.868	51.107	17.277	45.590	30.691	0.169
DMU17	9.307	2.936	53.055	102.67	84.770	1.466	140.52	2.934	0.496
DMU18	13.535	70.179	87.534	120.91	77.687	5.979	109.71	4.156	0.340
DMU19	14.805	6.832	61.754	57.921	42.567	63.600	58.750	3.897	0.520
DMU20	12.697	29.560	9.457	38.286	104.42	39.826	78.962	37.364	0.328

$$\begin{cases} \dfrac{1}{(-6.635)}\begin{pmatrix} -0.342t_1 - 0.272t_2 - 0.448t_3 - 0.44t_4 \\ +0.446t_5 - 0.46t_6 - 0.067t_7 \end{pmatrix} = e \\ \dfrac{1}{(0.751)}\begin{pmatrix} 0.305t_1 + 0.421t_2 - 0.193t_3 - 0.36t_4 \\ -0.342t_5 - 0.357t_6 + 0.564t_7 \end{pmatrix} = e \\ \dfrac{1}{(1.345)}\begin{pmatrix} -0.51t_1 + 0.566t_2 + 0.314t_3 - 0.427t_4 \\ +0.183t_5 + 0.325t_6 + 0.004t_7 \end{pmatrix} = e \\ \dfrac{1}{(0.153)}\begin{pmatrix} 0.486t_1 + 0.085t_2 + 0.587t_3 - 0.191t_4 \\ +0.365t_5 - 0.405t_6 - 0.281t_7 \end{pmatrix} = e \end{cases} \quad (4.1)$$

To emphasize the problem of unboundedness, choose a point within the feasible region, that is, $t_1 = t_2 = t_3 = t_6 = 0$. At this point, equation (4.1) is simplified to

$$\begin{cases} 0.066t_4 - 0.067t_5 + 0.01t_7 = e \\ -0.48t_4 - 0.455t_5 + 0.751t_7 = e \\ -0.317t_4 + 0.136t_5 + 0.003t_7 = e \\ -1.245t_4 + 2.383t_5 - 1.839t_7 = e \end{cases} \quad (4.2)$$

Observe that if the input excesses $s_{x(4)}, s_{x(5)}$ and $s_{x(7)}$ are loaded heavily, for example $s_{x(4)} = s_{x(5)} = s_{x(7)} = s_b$, where s_b is a large number, the constraints will then be driven by

$$\begin{cases} 0.009s_b = e \\ -0.184s_b = e \\ -0.178s_b = e \\ -0.701s_b = e \end{cases} \quad (4.3)$$

It can be observed from equation (4.3) that the constraints related to γ_2, γ_3 and γ_4 lead to unbounded feasible region for e because e can be made as small as possible in the linear program. In the constraint related to γ_1, the input excesses are weighted with a very small positive number. Thus, this constraint can easily be made zero or negative, if $v_1^{\mathrm{T}}(\mathbf{X}\lambda)$ is negative. As a result, the PCA-DEA estimator encounters the problem of unboundedness, and this is shown in the efficiency scores obtained in column 2 of Table 4. These values are close to zero due to the setting of the lower bound of e to a zero in the linear program.

To produce non-negative data that meet the free disposability assumption in a DEA model, modifications on the eigenvectors were performed on the same set of data following the procedure suggested herein. First, the eigenvectors are rotated with a varimax rotation, giving the rotated factor axes shown in rows 5-8 of Table 3. Note that the first rotated axis v_1^r is dominated by the variables with negative

Table 1 (continued): Simulated Data and Correlation Matrix for Input Variables
(b) Correlation Matrix of Input Variables

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$
$\tilde{x}_1$	1	0.110	0.154	0.323	-0.454	-0.164	0.198
$\tilde{x}_2$	0.110	1	0.290	-0.320	-0.346	0.220	0.318
$\tilde{x}_3$	0.154	0.290	1	0.295	-0.065	0.471	-0.106
$\tilde{x}_4$	0.323	-0.320	0.295	1	-0.254	0.505	-0.142
$\tilde{x}_5$	-0.454	-0.346	-0.065	-0.254	1	-0.199	-0.256
$\tilde{x}_6$	-0.164	0.220	0.471	0.505	-0.199	1	-0.180
$\tilde{x}_7$	0.198	0.318	-0.106	-0.142	-0.256	-0.180	1

weights. Thus, variables $\tilde{x}_1, \tilde{x}_2, \tilde{x}_5$ and $\tilde{x}_7$ with positive weights that capture 17.1% of the SSL in γ_1 are classified as contrast variables in γ_1. To form an axis without the counter effect from the contrast variables, these variables are excluded, and the remaining variables $\tilde{x}_3, \tilde{x}_4$ and $\tilde{x}_6$ are used to form the normalized principal direction ω_1. This procedure is repeated for the other rotated axes $v_j^r, j = 2, 3, 4$ and the corresponding normalized principal directions $\omega_j, j = 2, 3, 4$ are produced (refer to rows 10-12 of Table 3).

This example illustrates that the contrast variables differ from one component to the other, and they cannot be identified prior to PCA. To examine if the modifications made to the eigenvectors weaken the components' ability to represent the dataset, a redundancy analysis was performed on the modified components against the original dataset. Results show that the modified components retain 82.1% explained variation of the original dataset, compared to 84.8%, captured by the principal components.

As described in the methodology, there are alternatives to avoid negative weights in eigenvectors, for example, (1) squaring the entries of the eigenvectors and (2) taking absolute values of the eigenvectors. To compare these alternatives, redundancy analyses were performed on the modified components corresponding to these methods against the original dataset using equation (3.3). The redundancy analyses show that there is a 69.0% redundancy from components obtained by option (1) and 69.5% redundancy from components obtained by option (2). This means that, although there is a drop in the amount of retained variation, the proposed modification is still the best among the other options. Hence, the components from the proposed modifications are used to replace the original variables in the DEA model for the efficiency estimation.

To illustrate the benefit gained from the dimensionality reduction due to these modified components the efficiency scores of the proposed method (modified PC-DEA) was compared to the results of the standard DEA (columns 2 and 4 of Table 4). As expected, the standard DEA suffers from overestimation. Refer to the efficiencies pre-assigned, $e^{-\tau}$ (see column 10, Table 1), DMUs 1, 5, 6, 17, 18, 19 and 20 should not be classified as efficient as being identified by the standard DEA (see column 2, Table 4). This problem is overcome by the proposed method, whereby only DMU3 is identified as efficient, reflecting the scenario as portrayed in the pre-assigned efficiencies. As such, it may be said that there is no significant loss of information due to the modified components. This example shows that the efficiency estimates obtained from the modified PC-DEA is more accurate than that of the standard DEA.

It is known that DEA is sensitive towards the dimensionality relative to the sample size and PCA is best used for dimension reduction when data are highly correlated. To generalize the findings, Monte Carlo simulations that take 100 trials were designed for each of the cases classified by these factors, that is, the dimensionality, correlation levels and the sample sizes. The data generating process is the same as described above, whereby a production function $\tilde{y} = \prod_{j=1}^{p} (\tilde{x}_j)^{\frac{1}{p}}$, where p is the number of inputs is used to simulate data with CRS. For the factor of correlation, two levels of correlations are examined; a case where variables are moderately correlated (r < 0.6); pre-assigned with a correlation matrix $\mathbf{R}_1$, and another case where variables are highly correlated (r > 0.6); pre-assigned with a correlation matrix $\mathbf{R}_2$. Random samples for both levels of correlation are generated based on the upper triangular Cholesky decomposition of $\mathbf{R}_1$ and $\mathbf{R}_2$ respectively. These cases were repeated for the sample sizes of 20, 50 and 100 (see Table 5).

Results shows that, on average, for the inputs that are highly correlated, there is 1 principal component returned for case of 4 inputs and 1.4 principal components returned for the case with 7 inputs for all the sample sizes. The sharp reduction in the dimensionality validates the use of PCA when the data that are highly correlated. For the inputs that are moderately correlated, more principal components are returned in order to capture at least 80% of explained variation. On average,

Table 2: Principal Components (γ_j) and Modified Components (c_j)

DMU	γ_1	γ_2	γ_3	γ_4	c_1	c_2	c_3	c_4
DMU1	-6.635	0.751	1.345	0.153	5.420	4.536	5.439	4.488
DMU2	-5.867	-0.876	0.495	2.702	4.371	4.784	3.270	5.423
DMU3	-5.417	-2.427	2.552	0.189	6.172	2.328	3.900	4.853
DMU4	-4.665	-0.079	2.314	0.555	3.929	2.332	4.453	4.633
DMU5	-4.680	-2.417	3.450	0.847	5.370	1.692	3.865	5.311
DMU6	-4.377	-1.700	0.523	0.024	4.646	2.510	2.124	2.862
DMU7	-1.916	-1.910	1.080	1.683	2.874	2.024	1.408	3.591
DMU8	-5.456	0.529	2.789	0.543	4.441	3.032	5.429	4.964
DMU9	-3.829	1.724	0.475	0.372	2.613	3.448	4.091	2.459
DMU10	-6.962	-2.320	-0.440	0.425	6.909	4.864	2.608	4.301
DMU11	-4.496	-2.252	0.809	-0.120	5.126	2.315	2.014	3.302
DMU12	-5.221	-1.361	0.417	1.823	4.891	4.379	3.209	5.375
DMU13	-5.917	-1.524	-0.443	1.486	5.131	4.672	2.234	4.381
DMU14	-5.392	-1.887	1.083	-0.374	5.907	2.875	3.150	3.639
DMU15	-4.570	-3.680	0.958	1.202	5.755	2.754	1.853	5.039
DMU16	-2.773	-0.130	0.541	0.593	2.492	2.181	2.336	2.483
DMU17	-5.872	-2.627	2.441	0.107	5.843	1.765	2.992	4.746
DMU18	-6.683	-1.084	2.055	2.291	4.957	3.911	4.032	6.076
DMU19	-2.000	-1.562	2.392	1.611	2.659	1.109	2.632	3.910
DMU20	-3.324	-1.954	-0.399	-0.138	4.613	2.804	1.726	2.583

Table 3: Eigenvectors (v_j), Rotated Axes (v_j^r) and Modified Principal Directions (ω_j)

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$
v_1	-0.342	-0.272	-0.448	-0.440	0.446	-0.460	-0.067
v_2	0.305	0.421	-0.193	-0.360	-0.342	-0.357	0.564
v_3	-0.510	0.566	0.314	-0.427	0.183	0.325	0.004
v_4	0.486	0.085	0.587	-0.191	0.365	-0.405	-0.281
v_1^r	0.156	0.037	-0.066	-0.575	0.380	-0.703	0.031
v_2^r	0.806	-0.046	0.121	0.299	-0.373	-0.273	0.176
v_3^r	-0.051	0.718	0.110	-0.349	-0.329	0.145	0.468
v_4^r	0.171	0.244	0.806	-0.018	0.303	0.136	-0.389
ω_1	0	0	0.072	0.632	0	0.772	0
ω_2	0.910	0	0.136	0.337	0	0	0.198
ω_3	0	0.820	0.126	0	0	0.166	0.534
ω_4	0.186	0.265	0.875	0	0.329	0.148	0

Table 4: Estimated Efficiency Scores for DEA, PCA-DEA and Modified PC-DEA

DMU	$\hat{e}$ (DEA)	$\hat{e}$ (PCA-DEA)	$\hat{e}$ (mPC-DEA)
DMU1	1	5.4E-14	0.616
DMU2	0.398	4.0E-15	0.290
DMU3	1	3.1E-15	1
DMU4	0.750	2.7E-15	0.611
DMU5	1	2.0E-16	0.766
DMU6	1	1.2E-16	0.645
DMU7	0.466	9.4E-18	0.321
DMU8	0.707	5.8E-14	0.487
DMU9	0.958	1.1E-16	0.549
DMU10	0.729	2.6E-15	0.310
DMU11	0.689	1.2E-16	0.470
DMU12	0.655	2.4E-15	0.404
DMU13	0.642	2.0E-16	0.321
DMU14	0.304	2.2E-15	0.218
DMU15	0.679	1.2E-16	0.356
DMU16	0.385	2.3E-17	0.302
DMU17	1	4.2E-15	0.253
DMU18	1	5.4E-15	0.347
DMU19	1	6.3E-18	0.708
DMU20	1	7.6E-17	0.591

Table 5: List of Monte Carlo Experiments

Experiment	Sample Size	n (inputs)	Pairwise Correlation Level
1	20	4	High ($r > 0.6$)
2	20	4	Moderate ($r < 0.6$)
3	20	7	High ($r > 0.6$)
4	20	7	Moderate ($r < 0.6$)
5	50	4	High ($r > 0.6$)
6	50	4	Moderate ($r < 0.6$)
7	50	7	High ($r > 0.6$)
8	50	7	Moderate ($r < 0.6$)
9	100	4	High ($r > 0.6$)
10	100	4	Moderate ($r < 0.6$)
11	100	7	High ($r > 0.6$)
12	100	7	Moderate ($r < 0.6$)

there are 2.7-3.0 principal components returned for the case with 4 inputs, and 3.9-4.4 principal components returned for the case with 7 inputs. To compare the information retention power, redundancy analyses between the original variables and the modified components were performed on these simulated dataset, comparing the redundancies due to the proposed method, taking squared value of eigenvectors (option 1) and taking absolute value of the eigenvectors (option 2). The results of the analyses are shown in Table 6. Note that, when there is only 1 principal component returned, there is no difference between the three options because there is only one factor axis to be considered. However, when there is more than one principal component, the redundancies captured by these options differ. As the proposed method provokes the least perturbations to the eigenvectors, it captures the most explained variation in all the cases, with reasonably low standard deviation. Referring to column 2 of Table 6, it is observed that the modified components obtained with the proposed method retain almost as much the information as in the principal components, that is, capturing at least 80% of explained variation. Thus, it may be concluded that the proposed method is the best alternative among these options to avoid negative weights in principal components because it causes the least information loss.

To compare the efficacy of the proposed method (modified PC-DEA) to the standard DEA, the efficiency estimates from the modified PC-DEA and the standard DEA were compared to the simulated efficiencies. Figure 3 illustrates the comparisons for two extreme cases, namely (a) the worst case with a sample size n = 20, 1 output and 7 moderately correlated inputs, and (b) the best case with a sample size n = 100, 1 output and 4 highly correlated inputs. Note that for both cases, the efficiency estimates from the modified PC-DEA are closer to the simulated efficiencies compared to the standard DEA.

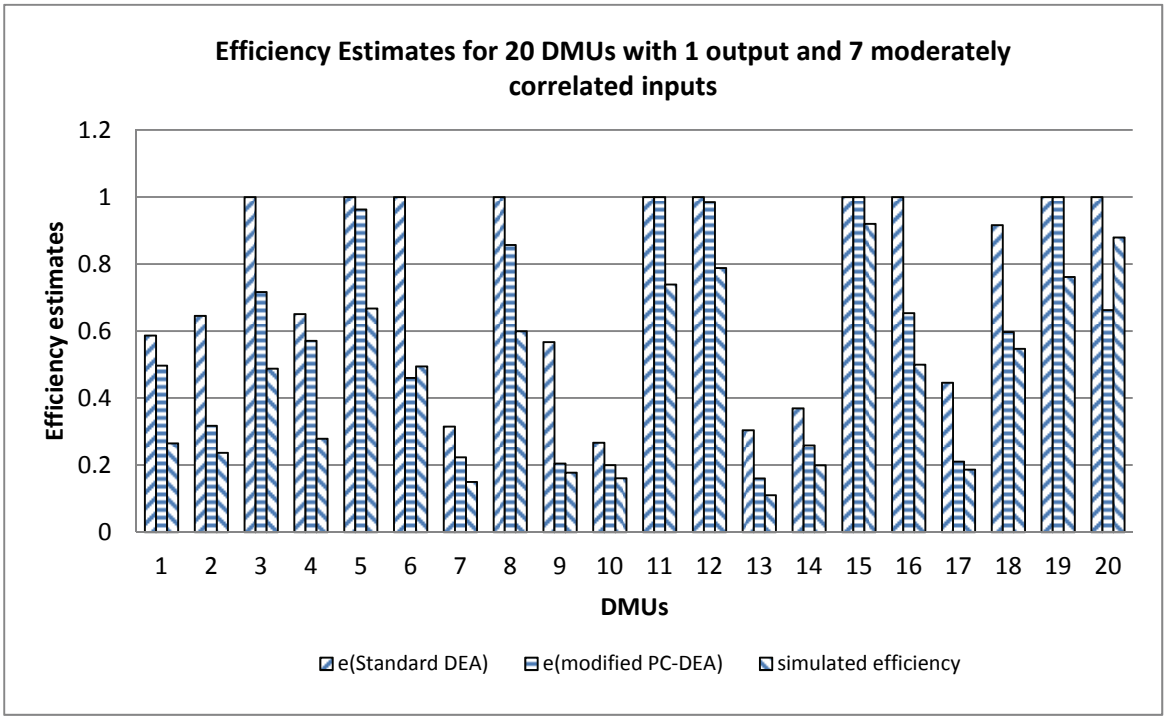
Table 6: Results of the Redundancy Analyses

Experiment	Redundancy[a] Proposed Method		Redundancy[a] Option 1[b]		Redundancy[a] Option 2[c]	
	Average	Std Dev	Average	Std Dev	Average	Std Dev
1	0.937	0.021	0.937	0.021	0.937	0.021
2	0.883	0.046	0.833	0.069	0.822	0.070
3	0.860	0.034	0.857	0.033	0.857	0.032
4	0.846	0.031	0.770	0.060	0.760	0.058
5	0.936	0.011	0.936	0.011	0.935	0.012
6	0.905	0.033	0.845	0.042	0.838	0.041
7	0.851	0.034	0.849	0.032	0.850	0.032
8	0.831	0.034	0.773	0.052	0.759	0.053
9	0.933	0.009	0.933	0.009	0.933	0.009
10	0.910	0.020	0.834	0.032	0.827	0.030
11	0.841	0.033	0.839	0.031	0.839	0.031
12	0.834	0.043	0.780	0.058	0.762	0.059

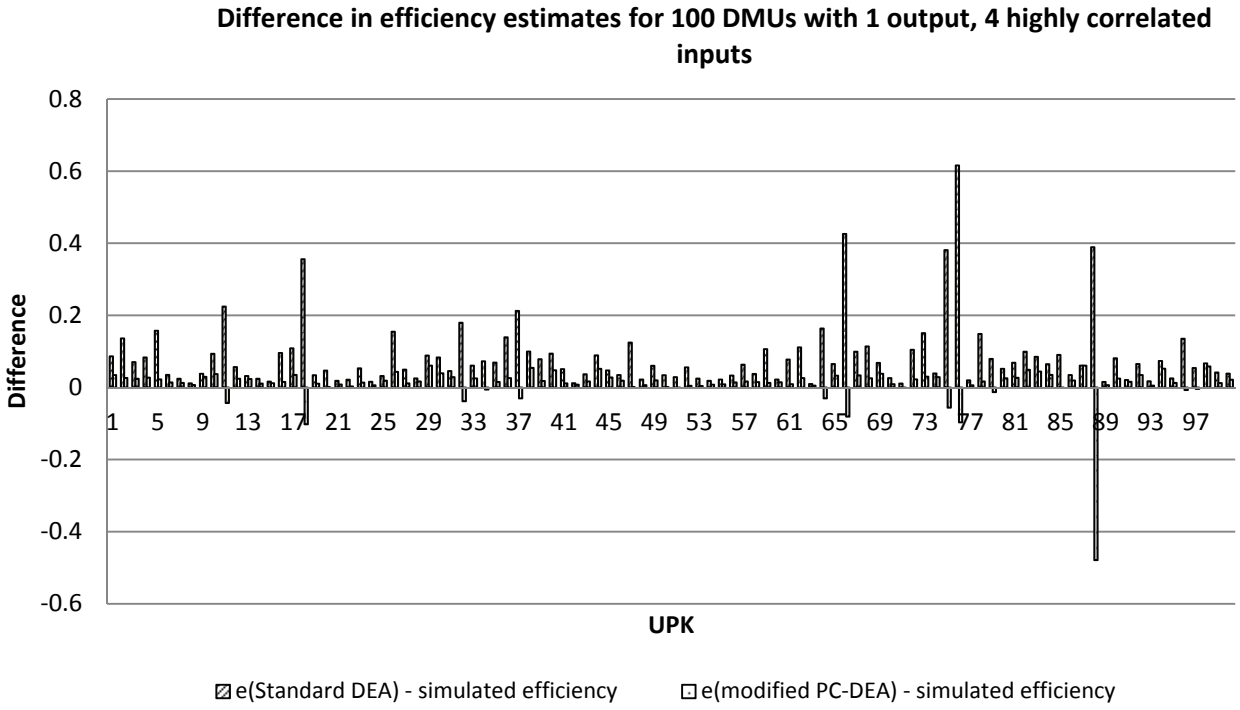
a: Redundancy between the original variables and the modified components
b: Option 1 represents the squared value of eigenvectors
c: Option 2 represents the absolute value of eigenvectors

Figure 3: Comparison of Efficiency Estimates to the Simulated Efficiencies

(a) Efficiency Estimates for 20 DMUs with 1 Output and 7 Moderately Correlated Inputs



(b) Difference in Efficiency Estimates for 100 DMUs with 1 Output and 4 Highly Correlated Inputs



To further examine the discriminatory power of the estimators, the percentages of overestimation and underestimation of each model were reckoned. An overestimation is observed when an inefficient DMU ($e^{-\tau} < 0.9$) is identified as efficient ($\hat{e} = 1$), and an underestimation occurs when an efficient DMU ($e^{-\tau} > 0.9$) is identified as inefficient ($\hat{e} < 1$). The results of the Monte Carlo simulations are shown in Table 7. Note that the standard DEA suffers from the curse of dimensionality. As expected, the worst case (Experiment 4, of which n = 20, with 1 output and 7 moderately correlated inputs) produces huge overestimation (42%). Consistent with Simar and Wilson (2000b), the increase in the sample size (from n = 20 to n = 100) does not give much ease to the overestimation problem (from 42% to 26.31%). Conversely, note that by using the modified components to replace the original variables, the problem of overestimation is reduced sharply.

For this worst case (Experiment 4), the proposed method replaces all the 7 inputs with 4 modified components, thus reduces the overestimation to 17.8%. Note also that both the modified PC-DEA and the standard DEA work better when data are highly correlated because the constraints attributable to the variables are rather similar to each other. Nonetheless, even in the best scenario (Experiment 9, of which n = 100, with 4 highly correlated inputs), the modified PC-DEA is still better than the standard DEA by having a much slighter overestimation (0.06% compared to 4.24%). The modified PC-DEA performs well in all cases to overcome the problem of overestimation. Although it produces underestimations (0.24% – 2.11%) due to the loss of information, the effect is deemed slender compared to the improvement in the discriminatory power.

Table 7: Results of Monte Carlo Simulations (100 trials) on the Percentages of Overestimation and Underestimation

Experiment	% Overestimation		% Underestimation	
	Std DEA	Modified PC-DEA	Std DEA	Modified PC-DEA
1	11.50	1.30	0	1.65
2	22.05	11.05	0	0.20
3	18.75	2.80	0	1.15
4	42.00	17.80	0	0.30
5	7.20	0.30	0	1.70
6	14.58	6.98	0	0.16
7	13.26	1.18	0	1.20
8	33.16	12.32	0	0.28
9	4.24	0.06	0	2.11
10	10.05	4.32	0	0.28
11	9.08	0.33	0	1.80
12	26.31	8.91	0	0.24

Conclusion

Literature shows that PCA-DEA outperforms other methods when all the variables under consideration are relevant. Furthermore, it is a convenient approach to reduce the dimensionality because it involves the least run time and estimation results are satisfactory. Principal components are the uncorrelated weighted linear combinations of original variables that capture the maximum variance. As the linear combinations are formed with a mixture of positive and negative weights, principal components could not meet the free disposability assumption in a DEA model. Consequently, the problem of unboundedness might arise in the linear program of the DEA model.

To overcome this problem, this study proposed that the eigenvectors be modified whereby each of the modified axes is constructed based on a set of variables that correlate in the same direction to the respective principal component. The modification involves the exclusion of contrast variables that capture a smaller portion of SSL, thus, there would not be significant information loss due to the modification. This was illustrated in redundancy analysis using Monte Carlo experiments. Compared to other possible alternatives to obtain non-negative weights for the principal components, the modified components due to proposed method captured the largest redundancy – in fact, they retained almost as much the explained variation as in the extracted principal components.

This study showed that the modified PC-DEA performs well to overcome the problem of overestimation, particularly when data are highly correlated. Because the modification can be obtained easily by adding programming codes to existing PCA-DEA its run time is not different from that of PCA-DEA. Better data reconstruction avoids the problem of unboundedness in a linear program, thus, the modified PC-DEA is a practical alternative to reduce dimensionality in a DEA model. In circumstances when there are many relevant variables, but not many comparable observations, researchers may consider applying the proposed method to aid meaningful benchmarking processes.

Acknowledgement

The authors would like to thank L. P. Teo for constructive advice.

References

Adler, N, & Golany, B. (2001). Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe. *European Journal of Operational Research*, *132*(*2*), 18-31.

Adler, N, & Golany, B. (2002). Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society*, *53*, 985-991.

Adler, N., & Yazhemsky, E. (2010). Improving discrimination in data envelopment analysis: PCA-DEA or variable reduction. *European Journal of Operational Research*, *202*, 273-284.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, *39*(*10*), 1261-1264.

Angulo-Meza, L., & Lins, M. R. E. (2002). Review of methods for increasing discrimination in data envelopment analysis. *Annals of Operations Research*, *116*, 225-242.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L. M., & Stutz, J. (1985). Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics 30*, 91-107.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429-444.

Daraio, C., & Simar, L. (2007). Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications. New York, NY: Springer.

Doyle, J. R., & Green, R. H. (1994). Efficiency and cross-efficiency in DEA: derivations, meanings and uses. *Journal of the Operational Research Society*, *45*(*5*), 567‑578.

Dunteman, G. H. (1989). *Principal components analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series no. 07-069. Newbury Park: Sage.

Dyson, R., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, *132*, 245-259.

Fare, R. (1998). *Fundamentals of production theory*. Berlin, Germany: Springer-Verlag.

Green, R. H., Doyle, J. R., & Cook, W. D. (1996). Preference voting and project ranking using DEA and cross-evaluation. *European Journal of Operational Research*, *90*, 461-472.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: Massachusetts Institute of Technology.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321-377.

Jenkins, L., & Anderson, M. (2003). A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European Journal of Operational Research*, *147*, 51-61.

Jolliffe, I. T. (2002). *Principal component analysis*, *2nd Ed*. New York, NY: Springer-Verlag.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187-200.

Kao, L. J., Lu, C. J., & Chiu, C. C. (2011). Efficiency measurement using independent component analysis and data envelopment analysis. *European Journal of Operational Research*, *210*, 310-317.

Kline, P. (2002). *An easy guide to factor analysis*. London, UK: Routledge.

Kneip, A., Park, B. U., & Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, *14*, 183-793.

Kneip, A., Simar, L., & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in non-parametric frontier models. *Econometric Theory*, *24*, 1663-1697.

Lee, T. W. (1998). *Independent component analysis: Theory and application*. Boston, MA: Kluwer Academic Publisher.

Nataraja, N. R., & Johnson, A. L. (2011). Guidelines for using variable selection techniques in data envelopment analysis. *European Journal of Operational Research*, *215*, 662-669.

Pastor, J. T., Ruiz, J. L., & Sirvent, I. (2002). A statistical test for nested radial DEA models. *Operations Research*, *50*(*4*), 728-735.

Pedraja-Chaparro, F., Salinas-Jimenez, J., & Smith, P. (1999). On the quality of the data envelopment analysis model. *Journal of the Operational Research Society*, *50*, 636-644.

Podinovski, V. V., & Thanassoulis, E. (2007). Improving discrimination in data envelopment analysis: some practical suggestions. *Journal of Productivity Analysis*, *28*, 117-126.

Ruggiero, J. (2005). Impact assessment of input omission on DEA. *International Journal of Information Technology & Decision Making*, *4*(*3*), 359-368.

Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: critique and extensions. In *Measuring efficiency: An assessment of data envelopment analysis*, R. H. Silkman, Ed. *32*, 73-105. San Francisco, CA: Jossey-Bass.

Shephard, R. W. (1970). *Theory of cost and production functions*. Princeton, NJ: Princeton University Press.

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

Simar, L., & Wilson, P. W. (2000a). A general methodology fir bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, *27*, 779-802.

Simar, L., & Wilson, P. W. (2000b). Statistical inference in nonparametric frontier models: the state of the art. *Journal of Productivity Analysis*, *13*, 49-78.

Simar, L., & Wilson, P. W. (2001). Testing restriction in nonparametric efficiency models. *Communication in Statistics*, *30*, 159-184.

Sirvent, L., Ruiz, J. L., Borras, F., & Pastor, J. T. (2005). A Monte Carlo evaluation of several tests for the selection of variables in DEA models. *International Journal of Information Technology & Decision Making*, *4*(*3*), 325-343.

Smith, P. (1997). Model misspecification in data envelopment analysis. *Annals of Operations Research*, *73*, 233-252.

Stewart, D. K., & Love, W. A. (1968). A general canonical correlation index. *Psychological Bulletin*, *70*, 160-163.

Thanassoulis, E. (2001). *Introduction to the theory and application of data envelopment analysis: A foundation text with integrated software*. USA: Kluwer Academic Publishers.

Ueda, T., & Hoshiai, Y. (1997). Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. *Journal of the Operational Research Society of Japan*, *40*, 466-478.

van den Wollenberg, A. L. (1977). Redundancy analysis: Alternative for canonical analysis. *Psychometrika*, *42*, 207-219.