# APPLICATION OF PRINCIPAL COMPONENT ANALYSIS FOR PARSIMONIOUS SUMMARIZATION OF DEA INPUTS AND/OR OUTPUTS 

Tohru Ueda<br>Seikei University

Yoko Hoshiai<br>NTT Laboratories

(Received August 4, 1995; Final June 12, 1997)


#### Abstract

In Data Envelopment Analysis (DEA), when there are more inputs and outputs, there are more efficient Decision Making Units (DMUs). For example, if the specific inputs or outputs advantageous for a particular DMU are used, the DMU will become efficient. Usually the variables used as inputs or outputs are correlated. Therefore, the inputs and outputs should be selected appropriately by experts who know their characteristics very well. People who are less familiar with those characteristics require tools to assist in the selection. We propose using principal component analysis as a means of weighting inputs and/or outputs and summarizing parsimoniously them rather than selecting them. A basic model and its modification are proposed.

In principal component analysis, many weights for the variables that define principal components (PCs) have negative values. This may cause a negative integrated input that is a denominator of the objective function in fractional programming. The denominator should be positive. In the basic model, a condition that the denominator must be positive is added. When the number of PCs is less than the number of original variables, a part of original information is neglected. In the modified model, a part of the neglected information is also used.


## 1. Introduction

Data Envelopment Analysis (DEA) which was ingeniously developed by Charnes, Cooper and Rhodes [3] evaluates the relative efficiency of decision making units (DMU) that have many inputs and outputs. When there are more inputs and outputs, there are more efficient ones. For example, if the specific inputs or outputs advantageous for a particular DMU are used, the DMU will become efficient. Usually the variables used as inputs or outputs are correlated. Nunamaker [6] says that addition of a highly correlated variable may substantially alter the DEA efficiency evaluations. Therefore, the inputs and outputs should be selected appropriately by experts who know their characteristics very well. People who are less familiar with those characteristics require tools to assist in the selection. We propose using principal component analysis as a means of weighting inputs and/or outputs and summarizing parsimoniously them rather than selecting them. However, principal component analysis has two problems which have to be overcome. For these problems, a basic model and a modificaiton of it are proposed.

The first problem is as follows. In principal component analysis, many weights for the variables that define principal components ( PCs ) take negative values. This may cause a negative integrated input that is a denominator of the objective function in fractional programming. If both the numerator and the denominator have negative values, the fraction has a positive value, but it is difficult to compare the value with positive fraction values which are derived from positive numerators and positive denominators.

In order to overcome this problem, fractions whose denominators and numerators are both negative must be transformed into appropriate forms. We conserve that the smaller inputs are, the better the efficiency is and also the more outputs are, the better the efficiency is.

Even if denominators become positive by adding the same positive number, ordinal relations among denominators are conserved. When denominators are positive, ordinal relation among fractions can be always decided. From these, denominators must be positive. In the basic model, a constraint which satisfies this condition (Non-Negative Constraint: NNC) is added.

If all inputs are positive, NNC is redundant. However if there are negative inputs, NNC becomes effective. This means that in usual data the basic model is equivalent to models which do not have NNC and it can treat even negative inputs and outputs.

As the second problem, when the number of principal component is less than the number of original variables, a part of original informaiton is neglected as a variation factor. Mardia [5] says that principal component analysis looks for a few linear combinations which can be used to summarize the data, losing in the process as little information as possible. In the modified model, the information neglected in principal component analysis is recovered as much as possible.

## 2. Parsimonious Summarization of Inputs and/or Outputs (Basic Model)

### 2.1 Basic model formulation

In this paper, DEA is discussed basically as fractional programming in the following way.

$$
\begin{align*}
\max & D_{J}=\sum_{r=1}^{s} u_{r} O U T_{r J} / \sum_{i=1}^{m} v_{i} I N P_{i J}  \tag{2.1}\\
\text { subject to } & \sum_{r=1}^{s} u_{r} O U T_{r j} / \sum_{i=1}^{m} v_{i} I N P_{i j} \leq 1(j=1,2, \cdots, n), \\
& u_{r}, v_{i} \geq \varepsilon(r=1,2, \cdots, s ; i=1,2, \cdots, m)
\end{align*}
$$

where $J$ is the objective DMU,
$I N P_{i j}$ is the $i$-th input of $\mathrm{DMU} j$, and
$O U T_{r j}$ is the $r$-th output of $\operatorname{DMU}(j=1,2, \cdots, n)$.
First we discuss the input variables. The same discussion applies for the outputs. If $I N P_{i j}=I N P_{k j}(i \neq k$ and $\forall j), v_{i}$ and $v_{k}$ cannot be determined uniquely in the same way as in regression analysis. Between highly correlated inputs, $v_{i}$ may be unstable. Nunamaker [6] says that methods for handling the variable selection are very important. If we use the principal components (PCs) as inputs in equation (2.1), they have no correlations. (See [4] etc. for the principal component analysis.)

Let variables which have been usually used as input variables in DEA be $a_{i j}$, and let their $k$-th PCs be $\boldsymbol{x}_{k}$.

$$
\begin{align*}
x_{k j} & =\sum_{i=1}^{m} w_{k i} a_{i j} ; \sum_{i=1}^{m} w_{k i}^{2}=1(k=1,2, \cdots, p ; j=1,2, \cdots, n)  \tag{2.2}\\
\boldsymbol{x}_{k}{ }^{t} & =\left(x_{k 1}, x_{k 2}, \cdots, x_{k n}\right) . \tag{2.3}
\end{align*}
$$

When the original inputs are $a_{i j}{ }^{(0)}$, the inputs standardized as $a_{i j}$ are usually defined as

$$
\begin{equation*}
a_{i j}=\left(a_{i j}^{(o)}-\bar{a}_{i}^{(o)}\right) / s_{i}^{(o)}, \tag{2.4}
\end{equation*}
$$

where $\bar{a}_{i}^{(o)}$ and $s_{i}{ }^{(o)}$ are the mean and standard deviation of $a_{i j}{ }^{(o)}$. In this standardization, about half of the $a_{i j}$ have negative values, but the inputs used in equation (2.1) are desirable to be positive. It is known that $x_{k j}$ in equations (2.2) and (2.3) do not change even if a constant $c_{i}$ is added to the original inputs $a_{i j}{ }^{(0)} / s_{i}{ }^{(o)}$. Letting $c_{i}$ equal $\bar{a}_{i}^{(o)} / s_{i}^{(o)}$, the following standardization is proposed instead of equation (2.4).

$$
\begin{equation*}
a_{i j}=a_{i j}{ }^{(o)} / s_{i}{ }^{(o)} . \tag{2.5}
\end{equation*}
$$

This is justified because of the coincidence of the origin point.
Let $\boldsymbol{w}_{k}{ }^{t}=\left(w_{k 1}, w_{k 2}, \cdots, w_{k m}\right)$ be the weights used for the $k$-th principal components $\boldsymbol{x}_{k}$. The inner products of different weight vectors equal zero. If all elements of $\boldsymbol{w}_{1}$ are positive, then half the elements of $\boldsymbol{w}_{2}$ may be negative. This may result in negative denominators in equation (2.1). The numerators in equation (2.1) are permitted to be negative, but the denominators are not. Thus, a constant $C_{J}$ is added to the denominators in equation (2.1) and the following constraints are added to equation (2.1),

$$
\begin{equation*}
\sum_{k=1}^{p} v_{k} x_{k j}+C_{J} \geq \varepsilon \quad(j=1,2, \cdots, n) \tag{2.6}
\end{equation*}
$$

where $C_{J}$ is a constant depending on the objective DMU $J$.
Note that we cannot add the constants $d_{J}$ to the numerators in equation (2.1), because $D_{J}$ becomes close to one if all $u_{r}$ and $v_{i}$ are very small and $C_{J}=d_{J}$. A new formulation as fractional programming for the target $\mathrm{DMU} J$ is

$$
\begin{align*}
\max & D_{J}=\sum_{r=1}^{q} u_{r} y_{r J} /\left(\sum_{k=1}^{p} v_{k} x_{k J}+C_{J}\right)  \tag{2.7}\\
\text { subject to } & \sum_{r=1}^{q} u_{r} y_{r j} /\left(\sum_{k=1}^{p} v_{k} x_{k j}+C_{J}\right) \leq 1(j=1,2, \cdots, n), \\
& \sum_{k=1}^{p} v_{k} x_{k j}+C_{J} \geq \varepsilon(j=1,2, \cdots, n), \\
& u_{r}, v_{k} \geq \varepsilon(r=1,2, \cdots, q ; k=1,2, \cdots, p),
\end{align*}
$$

where $\quad x_{k j}$ is the $k$-th PC of DMU $j$ for standardized inputs $a_{i j}$,
$y_{r j}$ is the $r$-th PC of DMU $j$ for standardized outputs $b_{h j}$,
$p(<m)$ is the number of PCs for inputs, and
$q(<s)$ is the number of PCs for outputs.
When the second condition in problem (2.7) is excluded, the following equation (2.8) is a linear programming formulation of problem (2.7) and is called the output-oriented BCC model in Charnes et al. [2].

$$
\begin{align*}
\min & D_{J}=\sum_{k=1}^{p} v_{k} x_{k J}+C_{J},  \tag{2.8}\\
\text { subject to } & \sum_{r=1}^{q} u_{r} y_{r J}=1, \\
& -\sum_{k=1}^{p} v_{k} x_{k j}-C_{J}+\sum_{r=1}^{q} u_{r} y_{r j} \leq 0(j=1,2, \cdots, n), \\
& u_{r}, v_{i} \geq \varepsilon(r=1,2, \cdots, q ; i=1,2, \cdots, p) .
\end{align*}
$$

However, this equation cannot treat negative outputs, for example, for the case of ( $q=1$ ) because of the first condition in this equation. Therefore, we propose the following model equivalent to problem (2.7) because of the second condition in the problem (2.7):

$$
\begin{align*}
\max & D_{J}=\sum_{r=1}^{q} u_{r} y_{r},  \tag{2.9}\\
\text { subject to } & \sum_{k=1}^{p} v_{k} x_{k J}+C_{J}=1, \\
& -\sum_{k=1}^{p} v_{k} x_{k j}-C_{J}+\sum_{r=1}^{q} u_{r} y_{r j} \leq 0(j=1,2, \cdots, n), \\
& \sum_{k=1}^{p} v_{k} x_{k j}+C_{J} \geq \varepsilon(j=1,2, \cdots, n), \\
& u_{r}, v_{i} \geq \varepsilon(r=1,2, \cdots, q ; i=1,2, \cdots, p) .
\end{align*}
$$

Usually this model not only accords with the output-oriented BCC model (2.8) through linear transformation of $v_{k}$ and $u_{r}$, and inverse of $D_{J}$, but also has solutions even in the case where the model (2.8) does not have any solutions.

We consider improvement of inefficient DMU. For the case in which outputs cannot be improved, the inputs of inefficient $\operatorname{DMU} J$ must be decreased. The left side of the second equation of equation (2.9) is expressed in terms of $a_{i J}$ as

$$
\begin{equation*}
\sum_{k=1}^{p} v_{k} x_{k J}+C_{J}=\sum_{i=1}^{m} \tilde{v}_{i} a_{i J}+C_{J} \tag{2.10}
\end{equation*}
$$

where $\tilde{v}_{i}=\sum_{k=1}^{p} v_{k} w_{k i}$.
Let $I_{1}$ be a set of improvable inputs of DMU $J, I_{2}$ be a set of nonimprovable inputs of $\operatorname{DMU} J$, and $\alpha_{i J}$ be the objective values of inputs $a_{i J}$ of DMU $J$ for satisfying $\left\{D_{J}=1\right\}$. The $\alpha_{i J}$ must satisfy

$$
\begin{equation*}
\sum_{i \in I_{1}} \tilde{v}_{i} \alpha_{i J}+C_{J}=\sum_{r=1}^{q} u_{r} y_{r J}-\sum_{i \in I_{2}} \tilde{v}_{i} a_{i J} . \tag{2.11}
\end{equation*}
$$

### 2.2 Treatment of negative weights

The $k$-th $\mathrm{PC} \boldsymbol{x}_{k}{ }^{t}=\left(x_{k 1}, x_{k 2}, \cdots, x_{k n}\right)$ for inputs are given by equation (2.2). When there are many negative weights $w_{k i}$, the minimum weight $(<0)$ has the same effect as the maximum weight ( $>0$ ) in the principal component analysis. Therefore, evaluating the efficiency with absolute values of negative and positive weights is considered. Define $P_{k}$, $N_{k}, x_{k j}{ }^{(P)}$ and $x_{k j}{ }^{(N)}$ to be

$$
\begin{align*}
& P_{k}=\left\{h \mid w_{k h} \geq 0\right\}, \quad N_{k}=\left\{h \mid w_{k h}<0\right\},  \tag{2.12}\\
& x_{k j}(P)=\sum_{h \in P_{k}} w_{k h} a_{h j}, \quad x_{k j}^{(N)}=\sum_{h \in N_{k}}\left|w_{k h}\right| a_{h j} .
\end{align*}
$$

A method of using both $x_{k j}{ }^{(P)}$ and $x_{k j}{ }^{(N)}$ may be considered, but it has the following shortcomings.
(a) The number of elements in $N_{k}$ may be quite different from the number of elements in $P_{k}\left(N_{k} \gg P_{k}\right.$ or $\left.N_{k} \ll P_{k}\right)$.
(b) The effect of reducing the number of original inputs, $m$, to the number of PCs, $p$, is weakened.
(c) The discussion about PCs, for example, the contribution ratio, variance and so on, must be reconstructed.
(d) Suppose that all elements of $\boldsymbol{w}_{1}$ are positive ( $N_{1}$ is empty). If neither $P_{2}$ nor $N_{2}$ is empty, the second input, $x_{2 j}$, is divided into two inputs, $x_{2 j}{ }^{(P)}$ and $x_{2 j}{ }^{(N)}$. This results in emphasizing $\boldsymbol{x}_{2}$ over $\boldsymbol{x}_{1}$.
These points indicate that using $x_{k j}{ }^{(P)}$ and $x_{k j}{ }^{(N)}$ is not desirable. Considering that in the principal component analysis the first PC must be emphasized, the sign of the second and the following PCs for the inputs should be decided so that they have a positive correlation with the first PC for the outputs. The sign of the first PC for the inputs should be decided such that there are more positive $x_{1 i}$. The sign of PCs for the outputs should be decided in the same way.

### 2.3 Example

Here we present an application to a problem in the Nippon Telegraph and Telephone Corporation. A message area (MA) is an area in which users can talk by telephone for 3 minutes for 10 yen. The efficiency of 66 message areas was evaluated. The forty items shown in Table 1 were used as inputs and the following six items as outputs.

$$
\begin{aligned}
\text { Revenue }: & \text { Long distance, } b_{1 j} \\
& \text { Local, } b_{2 j}
\end{aligned}
$$


When DEA was applied directly to all inputs and outputs, the efficiencies of all DMUs became one, because the number of inputs are too many for the number of DMUs, considering that in Tone [10] the following condition is requested:

$$
n \geq \max \{m \times s, 3(m+s)\},
$$

where it is not always true that all efficiencies become one for such size of problems as this example. Then, the principal components $\boldsymbol{x}_{i}(i=1,2, \cdots, p)$ of the inputs and the principal components $\boldsymbol{y}_{k}(k=1,2, \cdots, q)$ of the outputs were obtained. Weights $\boldsymbol{w}_{1}, \boldsymbol{w}_{2}$ and $\boldsymbol{w}_{3}$ in $\boldsymbol{x}_{1}, \boldsymbol{x}_{2}$ and $\boldsymbol{x}_{3}$ are shown in Table 1. For outputs, the contribution ratio $C R_{1}$ of the first PC is 0.998 , and only the first PC was used as the DEA outputs, where

$$
\begin{equation*}
C R_{k}=\lambda_{k} / \sum_{j=1}^{K} \lambda_{j} \tag{2.13}
\end{equation*}
$$

$K$ is a number of original variables ( $K=m$ for inputs) and $\lambda_{j}$ is the $j$-th largest eigenvalue of a variance-covarience matrix of $K$ variables.
For inputs the contribution ratios of the first and second principal components are 0.794 and 0.064 . Therefore, two PCs were used as DEA inputs.

In principal component analysis, $\left(-\boldsymbol{w}_{k}\right)$ is not differentiated from $\boldsymbol{w}_{k}$. In DEA, $\left(-\boldsymbol{w}_{k}\right)$ gives a different evaluation from $\boldsymbol{w}_{k}$. At first, the sign of weight vectors was decided such that there are more positive $x_{k i}$ for each $k$. Table 2 shows the DEA efficiency $D_{J}$. Figure 1 shows the first PC for outputs divided by the first PC for inputs versus DEA efficiency.

Table 1. Weights.

| Population in 15 -year age bands |  | $\boldsymbol{w}_{1}$ | $\boldsymbol{w}_{2}$ | $\boldsymbol{w}_{3}$ |
| :---: | :---: | :---: | :---: | :---: |
|  | $0 \sim 14$ | 0.177 | -0.041 | -0.016 |
|  | 15~29 | 0.177 | -0.041 | -0.016 |
|  | 30~44 | 0.177 | -0.041 | -0.016 |
|  | 45~59 | 0.177 | -0.037 | -0.016 |
|  | 60~ | 0.176 | -0.036 | -0.010 |
| No. of families |  | 0.177 | -0.041 | -0.016 |
| Sector |  | 0.031 | 0.438 | -0.360 |
| Agriculture, forestry, fishing | No. of offices | 0.053 | 0.423 | -0.246 |
|  | No. of employees | 0.081 | 0.454 | -0.109 |
| Mining \& quarrying | No. of offices | 0.000 | 0.049 | -0.078 |
|  | No. of employees | 0.037 | 0.235 | 0.428 |
| Construction | No. of offices | 0.176 | -0.007 | -0.016 |
|  | No. of employees | 0.176 | -0.035 | -0.016 |
| Manufacturing | No. of offices | 0.175 | -0.011 | -0.010 |
|  | No. of employees | 0.174 | -0.029 | -0.010 |
| Energy | No. of offices | 0.156 | 0.119 | -0.078 |
|  | No. of employees | 0.176 | -0.034 | 0.018 |
| Transport \& communications | No. of offices | 0.175 | 0.011 | -0.010 |
|  | No. of employees | 0.176 | -0.034 | 0.029 |
| Wholesale trade | No. of offices | 0.177 | -0.035 | -0.016 |
|  | No. of employees | 0.176 | -0.068 | -0.010 |
| Retail trade | No. of offices | 0.132 | 0.233 | 0.386 |
|  | No. of employees | 0.177 | -0.019 | -0.016 |
| Restaurant | No. of offices | 0.175 | -0.028 | -0.010 |
|  | No. of employees | 0.170 | -0.032 | -0.041 |
| Banking, finance \& insurance | No. of offices | 0.168 | 0.097 | 0.198 |
|  | No. of employees | 0.177 | -0.037 | -0.016 |
| Realtor | No. of offices | 0.165 | -0.032 | 0.002 |
|  | No. of employees | 0.173 | -0.075 | -0.011 |
| Service | No. of offices | 0.176 | 0.006 | -0.010 |
|  | No. of employees | 0.174 | -0.047 | -0.006 |
| Public offices | No. of offices | 0.027 | 0.361 | 0.563 |
|  | No. of employees | 0.164 | -0.027 | -0.035 |
| Industrial products |  | 0.173 | -0.055 | 0.036 |
| Retail sales |  | 0.177 | -0.022 | -0.016 |
| Income per capita |  | 0.045 | 0.338 | -0.294 |
| Assets | Long distance | 0.175 | 0.005 | -0.010 |
|  | Local | 0.176 | -0.052 | -0.016 |
| Expenditure | Long distance | 0.175 | -0.030 | -0.013 |
|  | Local | 0.173 | -0.060 | -0.038 |

$\begin{array}{lllll}\text { Eigen values } & 0.794 & 0.064 & 0.044\end{array}$

Table 2. DEA efficiency.

| $J$ | $D_{J}$ | $J$ | $D_{J}$ | $J$ | $D_{J}$ | $J$ | $D_{J}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 1.000 | 21 | 0.644 | 41 | 0.773 | 61 | 0.615 |
| 2 | 0.812 | 22 | 0.498 | 42 | 0.508 | 62 | 0.562 |
| 3 | 0.682 | 23 | 0.620 | 43 | 0.353 | 63 | 0.632 |
| 4 | 0.689 | 24 | 0.672 | 44 | 0.485 | 64 | 0.585 |
| 5 | 0.775 | 25 | 0.592 | 45 | 0.679 | 65 | 0.759 |
| 6 | 0.569 | 26 | 0.543 | 46 | 1.000 | 66 | 0.690 |
| 7 | 0.721 | 27 | 0.705 | 47 | 0.750 |  |  |
| 8 | 0.298 | 28 | 0.460 | 48 | 0.687 |  |  |
| 9 | 1.000 | 29 | 0.496 | 49 | 0.588 |  |  |
| 10 | 0.691 | 30 | 0.454 | 50 | 0.777 |  |  |
| 11 | 0.684 | 31 | 0.414 | 51 | 0.710 |  |  |
| 12 | 0.592 | 32 | 0.483 | 52 | 0.738 |  |  |
| 13 | 0.692 | 33 | 0.615 | 53 | 0.685 |  |  |
| 14 | 0.529 | 34 | 0.521 | 54 | 0.535 |  |  |
| 15 | 0.613 | 35 | 0.591 | 55 | 0.484 |  |  |
| 16 | 0.588 | 36 | 0.551 | 56 | 0.530 |  |  |
| 17 | 0.450 | 37 | 0.658 | 57 | 0.862 |  |  |
| 18 | 0.876 | 38 | 0.596 | 58 | 0.570 |  |  |
| 19 | 0.242 | 39 | 0.634 | 59 | 0.448 |  |  |
| 20 | 0.535 | 40 | 0.711 | 60 | 0.672 |  |  |

Table 3. Weights and free

| $J$ | $v_{1}$ | $v_{2}$ | $C_{J}$ | $u_{1}$ |
| ---: | :---: | :---: | ---: | ---: |
| 1 | $\varepsilon$ | $\varepsilon$ | 1 | 1 |
| 2 | 40.74 | $\varepsilon$ | -0.11 | 40.62 |
| 3 | 12.51 | $\varepsilon$ | -0.03 | 12.48 |
| 4 | 85.47 | $\varepsilon$ | -0.23 | 85.24 |
| 5 | 4.389 | $\varepsilon$ | -0.012 | 4.377 |
| 6 | 110.1 | $\varepsilon$ | -0.3 | 109.8 |
| 7 | 84.83 | 1.18 | -0.34 | 84.46 |
| 8 | 19.86 | $\varepsilon$ | -0.05 | 19.80 |
| 9 | 294.9 | 4.09 | -1.19 | 293.6 |
| 10 | 76.67 | 1.06 | -0.31 | 76.34 |

The distance from the diagonal line represents the effect of the second PC for the inputs. Table 3 shows weights $v_{1}, v_{2}$ and $u_{1}$ and free variables $C_{J}$ in equation (2.9) for DMU1 to DMU10.

All second PCs excluding MA (DMU)1 for the inputs are positive, but the correlation coefficients $R(i, 1)$ between $i$-th PC for the inputs and the first PC for the outputs are

$$
R(1,1)=0.993, \quad R(2,1)=-0.089
$$

From the sign of $R(2,1)$, it may be considered that $\left(-\boldsymbol{w}_{k}\right)$ should be used instead of the $\boldsymbol{w}_{k}$ used for Figure 1, but all second PCs excluding MA1, for the inputs become negative. Considering that $R(2,1)$ is very small, only the first PC should be used. Here, DMU1 is a special DMU, having extremely larger (four to five times) inputs and outputs than the second largest DMU and having the opposite sign of the second PC to other DMUs as above mentioned. This DMU1 is too large to compare with other DMUs. Therefore, excluding DMU1, analysis was also proceeded. In that case, the contribution ratio of the first PC for output was 0.985 . The contribution ratio of the first, second, and third PCs for inputs were $0.728,0.057$, and 0.053 , so these three components were used as DEA inputs. Table 4 shows the weights and free variables and Figure 2 shows the first PC for outputs divided by the first PC for inputs versus DEA efficiency when DMU1 is excluded. About half of $C_{J}$ values in Table 4 are positive and $C_{J}$ does not have a bias toward negative, though all $C_{J}$ except for ( $J=1$ ) in Table 3 are negative.

As a result we propose that the number of PCs should be decided according to the values of the contribution ratios $C R_{k}$ and the correlation coefficients $R(i, 1)$ and $R(1, j)$, after exclusion of DMUs with extraordinary inputs or outputs.


Figure 1. First PCs ratio versus DEA efficiency (including DMU1).
Table 4. Weights and free variables (MA1 excluded).

| $J$ | $v_{1}$ | $v_{2}$ | $v_{3}$ | $C_{J}$ | $u_{1}$ |
| ---: | :---: | ---: | ---: | ---: | ---: |
| 2 | 9.856 | 0.206 | $\varepsilon$ | 0.004 | 9.080 |
| 3 | 3.558 | 0.814 | $\varepsilon$ | -0.006 | 3.177 |
| 4 | 20.64 | $\varepsilon$ | $\varepsilon$ | 0.022 | 19.09 |
| 5 | 1.179 | 0.554 | $\varepsilon$ | 0.011 | 1.031 |
| 6 | 25.30 | $\varepsilon$ | 0.47 | -0.096 | 22.71 |
| 7 | 19.45 | $\varepsilon$ | 0.53 | -0.069 | 18.10 |
| 8 | 5.884 | $\varepsilon$ | $\varepsilon$ | 0.006 | 5.441 |
| 9 | 60.04 | $\varepsilon$ | 8.26 | -1.20 | 53.67 |
| 10 | 17.85 | $\varepsilon$ | 0.49 | -0.06 | 16.61 |

## 3. Modified Model

When the number, $p$, of PCs is less than the number, $m$, of original variables and the cumulative contribution ratio of $p \mathrm{PCs}$ is $r,(1-r)$ of total information is usually considered as a variation factor, that is, a noise or a disturbance. In this section, the information (for example, $N_{1} H_{1}$ and $N_{2} H_{2}$ in Figure 3) is used positively and presented by one additional dimension, that is, total information is presented by $(p+1)$ dimensions.

### 3.1 Utilization of the mean and variance of a variation factor

When $p$ PCs are $\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \cdots, \boldsymbol{x}_{p}$, the $(p+1)$-st variable $\boldsymbol{x}_{p+1}$ with mean $\mu_{p+1}$ and variance $V_{p+1}$ are added, where

$$
\begin{equation*}
V_{p+1}=\sum_{k=1}^{m} \lambda_{k}-\sum_{k=1}^{p} \lambda_{k} . \tag{3.1}
\end{equation*}
$$



Figure 2. First PCs ratio versus DEA efficiency (excluding DMU1).


Figure 3. Perpendiculars from $N$ to $H$.
The $k$-th PC, $\left\{\boldsymbol{x}_{k}{ }^{t}=\left(x_{k 1}, x_{k 2}, \cdots, x_{k n}\right)\right\}$ and its mean $\mu_{k}$ are given by

$$
\begin{equation*}
x_{k j}=\sum_{i=1}^{m} w_{k i} a_{i j}, \quad \mu_{k}=\sum_{j=1}^{n} x_{k j} / n=\sum_{i=1}^{m} w_{k i} \bar{a}_{i}^{(o)} / s_{i}^{(o)} \quad(k=1,2, \cdots, p) \tag{3.2}
\end{equation*}
$$

where $a_{i j}=a_{i j}{ }^{(o)} / s_{i}{ }^{(o)}, \bar{a}_{i}^{(o)}=\sum_{j=1}^{n} a_{i j}{ }^{(o)} / n$.
Considering that the sum of means of $m$ standardized variables $\boldsymbol{a}_{i}{ }^{t}=\left(a_{i 1}, \cdots, a_{i n}\right)[i=$ $1,2, \cdots, m$ ] is

$$
\begin{equation*}
\sum_{i=1}^{m} \bar{a}_{i}^{(o)} / s_{i}{ }^{(o)} \tag{3.3}
\end{equation*}
$$

and $\left(\sum_{k=1}^{p} \mu_{k}\right)$ out of it is presented by $p \mathrm{PCs}$, let $\mu_{p+1}$ be

$$
\begin{equation*}
\mu_{p+1}=\sum_{i=1}^{m}\left(1-\sum_{k=1}^{p} w_{k i}\right) \bar{a}_{i}^{(o)} / s_{i}^{(o)} . \tag{3.4}
\end{equation*}
$$

Because $\mu_{p+1}$ and $V_{p+1}$ do not depend on the DMU, $\boldsymbol{x}_{p+1}$ is expressed as a scalar variable $x_{p+1}$. The same procedure as for the inputs is applied to the outputs, using the ( $q+1$ )-st variable, $y_{q+1}$. The $x_{p+1}$ and $y_{q+1}$ are supposed to be random variables. The following measure may come to mind in place of the first equation in equation (2.7).

$$
\begin{equation*}
D_{J 1}=\left(\sum_{r=1}^{q} u_{r} y_{r J}+u_{q+1} y_{q+1}\right) /\left(\sum_{i=1}^{p} v_{i} x_{i J}+v_{p+1} x_{p+1}+C_{J}\right) . \tag{3.5}
\end{equation*}
$$

For the same reason that in Sec.2.1 no constant can be introduced in the numerators, the three parameters $u_{q+1}, v_{p+1}$ and $C_{J}$ cannot be used simultaneously. Therefore equation (3.5) cannot be used.

### 3.2 Consideration of the discrepancy between PC space and the original space

Let the coordinates of a point, $N$, in vector space $V^{m}$ be

$$
\begin{equation*}
\boldsymbol{a}^{(N)}=\left(a_{1 N}, a_{2 N}, \cdots, a_{m N}\right)^{t} \tag{3.6}
\end{equation*}
$$

Let the coordinates in $V^{m}$ of the foot, $H$, of a perpendicular from $N$ to the vector space $V^{p}$ whose elements are $p \mathrm{PCs}$ be

$$
\begin{equation*}
\boldsymbol{a}_{N}{ }^{(H)}=\left(a_{1 N}{ }^{(H)}, a_{2 N}{ }^{(H)}, \cdots, a_{m N}{ }^{(H)}\right)^{t} \tag{3.7}
\end{equation*}
$$

(see Fig. 3). Because the number of PCs is limited to $p$, the vector of the values of $m \mathrm{PCs}$ at the point $H$ is $\boldsymbol{x}^{(H)}=\left(x_{1 H}, x_{2 H}, \cdots, x_{p H}, 0, \cdots, 0\right)^{t}$.

Let

$$
\boldsymbol{W}=\left(\boldsymbol{w}_{1}, \boldsymbol{w}_{2}, \cdots, \boldsymbol{w}_{m}\right)^{t}, \quad \boldsymbol{X}=\left(\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \cdots, \boldsymbol{x}_{m}\right)^{t} \quad \text { and } \quad \boldsymbol{A}=\left(\boldsymbol{a}_{1}, \boldsymbol{a}_{2}, \cdots, \boldsymbol{a}_{m}\right)^{t},
$$

where

$$
\begin{aligned}
& \boldsymbol{w}_{k}=\left(w_{k 1}, w_{k 2}, \cdots, w_{k m}\right)^{t} \quad \text { for } \quad k=1,2, \cdots, m, \\
& \text { especially, } \boldsymbol{w}_{h}=\mathbf{0} \quad \text { for } \quad h>p, \\
& \boldsymbol{x}_{k}=\left(x_{k 1}, x_{k 2}, \cdots, x_{k n}\right)^{t} \quad \text { for } \quad k=1,2, \cdots, m, \quad \text { and } \\
& \boldsymbol{a}_{k}=\left(a_{k 1}, a_{k 2}, \cdots, a_{k n}\right)^{t} \quad \text { for } \quad k=1,2, \cdots, m .
\end{aligned}
$$

Then,

$$
\begin{align*}
\boldsymbol{X} & =\boldsymbol{W} \boldsymbol{A}  \tag{3.8}\\
\boldsymbol{A} & =\boldsymbol{W}^{t} \boldsymbol{X}
\end{align*}
$$

Therefore,

$$
\begin{equation*}
\boldsymbol{a}_{N}{ }^{(H)}=\boldsymbol{W}^{t} \boldsymbol{x}^{(H)} . \tag{3.10}
\end{equation*}
$$

Considering that fewer $\left(a_{i A}-a_{i A}{ }^{(H)}\right)$ is desirable, we hit on the idea that the denominator of the first equation in equaiton (2.7) may be changed to

$$
\sum_{i=1}^{p} v_{i} x_{i J}+v_{p+1} \sum_{i=1}^{m}\left(a_{i J}-a_{i J}^{(H)}\right)+C_{J} ; \quad v_{p+1}>0
$$

This compensates the reduction in information of PCs. Here, from the viewpoint of parameter parsimony, individual parameters, $v_{p+i}$, should not be taken for each ( $a_{i J}-a_{i J}{ }^{(H)}$ ). Applying this idea to the outputs as well as the inputs, the following modified model is proposed instead of equation (2.7). In this model, $(p+2)$ variables, $\left[v_{1}, v_{2}, \cdots, v_{p+1}, C_{J}\right]$, for inputs and $(q+1)$ variables, $\left[u_{1}, u_{2}, \cdots, u_{q+1}\right]$, for outputs must be decided.
[Proposed Model]
max

$$
D_{J 2}=\left\{\sum_{r=1}^{q} u_{r} y_{r J}+u_{q+1} \sum_{i=1}^{s}\left(b_{i J}-b_{i J}^{(H)}\right)\right\} /\left\{\sum_{i=1}^{p} v_{i} x_{i J}+v_{p+1} \sum_{i=1}^{m}\left(a_{i J}-a_{i J}^{(H)}\right)+C_{J}\right\},
$$

subject to

$$
\begin{aligned}
& \left\{\sum_{r=1}^{q} u_{r} y_{r j}+u_{q+1} \sum_{i=1}^{s}\left(b_{i j}-b_{i j}{ }^{(H)}\right)\right\} /\left\{\sum_{i=1}^{p} v_{i} x_{i j}+v_{p+1} \sum_{i=1}^{m}\left(a_{i j}-a_{i j}{ }^{(H)}\right)+C_{J}\right\} \leq 1, \\
& \sum_{i=1}^{p} v_{i} x_{i j}+v_{p+1} \sum_{i=1}^{m}\left(a_{i j}-a_{i j}{ }^{(H)}\right)+C_{J} \geq \varepsilon \quad(j=1,2 \cdots, n), \\
& v_{i} \geq \varepsilon(i=1,2, \cdots, p+1), \quad u_{r} \geq \varepsilon(r=1,2, \cdots, q+1),
\end{aligned}
$$

where $b_{i j}$ and $b_{i j}{ }^{(H)}$ are defined in the same way as $a_{i j}$ and $a_{i j}{ }^{(H)}$.
Figure 4 shows the relation between the first PC of outputs divided by the first PC of inputs and the efficiency, $D_{J 2}$, where DMU1 was excluded and $\{p=1, q=1\}$. Figure 4 has a slightly larger variation above the lowest line than Figure 2.


Figure 4. First PCs ratio versus DEA efficiency, $D_{J 2}$ (excluding DMU1).

### 3.3 Improvement of inputs in the modified model

This section discusses the improvement of inputs in the modified model of Sec.3.2 for the case in which outputs cannot be improved on the lines of equation (2.11). Suppose that every input, $a_{i J}$, of an inefficient DMU $J$ is decreased at a constant rate, $K(<1)$, to $\left\{\hat{a}_{i J}=K a_{i J}\right\}$. The $k$-th PC of the DMU $J$ for inputs $\hat{a}_{i J}$ is

$$
\begin{equation*}
\hat{x}_{k J}=K x_{k J} \tag{3.12}
\end{equation*}
$$

that is, $\hat{x}_{k J}$ is also decreased at rate $K$. Moreover,

$$
\begin{equation*}
\hat{a}_{i J}^{(H)}=K a_{i J}{ }^{(H)} . \tag{3.13}
\end{equation*}
$$

Therefore, if

$$
\begin{equation*}
K=\left\{\sum_{r=1}^{q} u_{r} y_{r J}+u_{q+1} \sum_{i=1}^{s}\left(b_{i J}-b_{i J}^{(H)}\right)-C_{j}\right\} /\left\{\sum_{i=1}^{p} v_{i} x_{i J}+v_{p+1} \sum_{i=1}^{m}\left(a_{i J}-a_{i J}^{(H)}\right)\right\} \tag{3.14}
\end{equation*}
$$

then $D_{J 2}=1$.

## 4. Conclusion

We proposed that the sign of PCs for the inputs (outputs) should be decided according to the correlation coefficients between those PCs and the first PC for outputs (inputs), and that the number of PCs should be decided from the values of the contribution ratios, $C R_{k}$, and the correlation coefficients, $R(i, 1)$ and $R(1, j)$. We presented a basic model and a modification of it that takes factors unexplained by PCs into account.

We overcame the disadvantage of principal component analysis and made possible its use as a parsimonious summarization tool for DEA inputs and/or outputs. Of course, we do not use principal component analysis when the inputs or outputs are not so many and the correlations among the inputs or outputs are weak.

The number, $p$, of principal components is usually decided by the commulative contribution ratio, $C C R_{p}$, or the $p$-th eigenvalue, $\lambda_{p}$. The more the value of $p$ is, the more difficult it is to explain the meaning of each principal component. Therefore, we propose limiting $p$ to the small values and recovering information with a modified model shown in Sec.3.2. From references [4], [5], [7] and [8], we recommend $C C R_{p} \geq 0.7$ or $\lambda_{p} \geq 1$ for correlation matrix.

If the modified model is used for $(p-1) \mathrm{PCs}$, we can derive a model that has the same number of variables and does not neglect completely information which is presented by residual $(m-p+1)$ PCs. This model becomes a compromise between the basic model in Sec. 2 and models which use all variables. For the modified model there may be other ideas and we need further study.

If variables are classified into some groups whose members have a high correlation each other and the principal component analysis is applied to each group, intuitive interpretation of results becomes easy, but a number of inputs or outputs may not decrease very much.

In multivariate analysis, canonical correlation analysis is well-known as a means of analyzing two sets of variables. In this paper, they are a set of input variables and a set of output variables. Let the $i$-th canonical variables for inputs and outputs be $f_{i}$ and $g_{i}$, respectively. Canonical correlation analysis has shortcomings as a method of summarizing parsimoniously variables and evaluating efficiency in DEA. For example, there is no correlation between $f_{2}$ and $g_{1}$. We cannot explain any meanings of the linear combination of $f_{1}$ and $f_{2}$. The fractional programming which has $f_{2}$ in denominator and $g_{1}$ in numerator should not be approved. When as a measure of efficiency, we only use a ratio, $g_{1} / f_{1}$, of the first canonical variables, canonical correlation analysis may have some meanings.

In equation (2.9) a non-Archimedean infinitesimal $\varepsilon$ was introduced. We can derive an $\varepsilon$-free DEA in the same way as a 2 -phase process in Tone [9]. In the similar way we can also derive an $\varepsilon$-free DEA for equation (3.11).

We discussed DEA as a fractional programming problem and added constraints that denominators must be positive. Discussion in negative weights and modified models are also applicable to other formulations of DEA. Especially, for the purpose that we do not
mind signs of inputs, it may be appropriate to use additive DEA models (see Charnes et al. [1]).

## References

[1] Charnes, A., Cooper, W.W., Golany, B., Seiford, L. and Stutz, J.: Foundations of Data Envelopment Analysis for Pareto-Koopmans efficient empirical production functions, J. of Econometrics, Vol. 30 (1985), 91-107.
[2] Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M.: Data envelopment analysis -Theory, methodology and applications-. Kluwer Academic Publishers, 1994.
[3] Charnes, A., Cooper, W.W. and Rhodes, E.: Measuring the efficiency of decision making units, European Journal of Operational Research, Vol. 2 (1978), 429-444.
[4] Chatfield, C. and Collins, A.J.: Introduction to multivariate analysis. Chapman \& Hall Ltd., 1984.
[5] Mardia, K.V., Kent, J.T. and Bibby, J.M.: Multivariate Analysis. Academic Press, 1979.
[6] Nunamaker, T.R.: Using data envelopment analysis to measure the efficiency of nonprofit organization, A critical evaluation, Managerian and Decision Economics, Vol.6, No. 1 (1985), 50-58.
[7] Okuno, T., Kume, H., Haga, T. and Yoshizawa, T.: Multivariate Analysis (revised). Nikkagiren, 1981 (in Japanese).
[8] Tanaka, Y. and Wakimoto, K.: Methods of Multivariate Statistical Analysis. Gendaisuugakusya, 1983 (in Japanese).
[9] Tone, K.: An $\varepsilon$-free DEA and a new measure of efficiency, J. of Operations Research Society of Japan, Vol.36, No. 3 (1993), 167-174.
[10] Tone, K.: Data Envelopment Analysis. Nikkagiren, 1993 (in Japanese).

Tohru Ueda<br>Department of Industrial Engineering<br>Faculty of Engineering<br>Seikei University<br>Kichijoji-Kitamachi, Musashino<br>Tokyo 180, Japan

