2021-1200 IJOI
https://www.ijoi-online.org/

International Journal of Organizational Innovation

DATA REDUCTION BY ENTROPY MEASURE OF FACTOR IMPORTANCE IN DATA ENVELOPMENT ANALYSIS

Chao-Chin Chao*
Department of Food and Beverage Management, Far East University
No.49, Zhonghua Rd., Xinshi Dist., Tainan City 74448, Taiwan, R.O.C.
Tel: +886-7-6222131 ext 51184
*Corresponding E-mail: Chawpp@mail2000.com.tw

Abstract

Data Envelopment Analysis (DEA) is a mathematical programming approach for measuring relative efficiencies within a group of decision making units based on multiple inputs and outputs. However, the number of inputs and outputs affects the discrimination level in efficiency evaluation of DMUs. This study introduces the concept of entropy to measure the importance of factors and uses the entropy values to identify the omitted factor(s). This method enables the decision maker to determine the less influential factor for omission between two highly correlated factors. It does not only improve the discrimination in the DEA evaluation but also retain information for further ranking.

Keywords: Data envelopment analysis, Data reduction, Entropy, Correlation

Introduction

The Data Envelopment Analysis (DEA) method, first proposed by Charnes et al. (1978), is widely known as an evaluation technique for efficiency rating within a group of decision making units (DMUs) based on multiple inputs and outputs. The efficiency of a DMU within the DEA frame is defined as the ratio of multiple weighted outputs to multiple weighted inputs. Under the DEA restriction that no DMU has more than 100% efficiency, the weights are chosen to show that a specific DMU is as

efficient as possible. DEA is a widely utilized technique to deal with efficiency evaluation and is often found in the management literature (for example, Chen and Bao 2013; Chen, 2011; Chang and Chen, 2008).

In efficiency measures, the number of inputs and outputs will affect the discrimination level and computational time in DEA evaluation. Based on the DEA frame, the more DMUs in the DEA model, the more constrained the weights, and the higher discrimination level the DEA result has. On the other hand, the more factors in the DEA model, the less discerning the analysis is. Because omitting factor(s) can have many advantages in DEA evaluation, there has been much work down in DEA to reduce the number of factors. To achieve a reasonable level of discrimination, a number of guidelines have been proposed in the literature suggesting limiting the number of variables relative to the number of DMUs. Two guidelines commonly applied are that the total number of inputs and outputs should be less than one third of the number of DMUs in the DEA model (Friedman and Sinuany-Stern, 1998) or that the number of DMUs should be at least two times the product of the number of inputs and number of outputs (Dyson et al., 2001).

The principal component analysis is a popular approach to reduce the number of factors, and it has been widely applied to deal with this task (e.g. Adler and Golany, 2001; Ueda and Hoshiai, 1997; Zhu, 1998). The principal component is a linear combination of factors so that the data used in the DEA model is not the original data of inputs and outputs. Another study from Jenkins and Anderson (2003) discussed the choice of variables to omit. They applied the regression and correlation analysis to reduce the number of factors that are highly correlated. They normalized the original data of all the factors to have a mean of zero and a variance of one, suggested trying all the combinations to find which factors best represent all the data, and used partial covariance of inputs or outputs as a measure of information contained in these retained variables. Recently, Wagner and Shimshak (2007) proposed a procedure of stepwise selection of variables, in a way similar to the stepwise regression method, that involves sequentially maximizing (or minimizing) the average change in the efficiencies as variables are added or dropped from the DEA model. However, although trying all the combinations to find the best represent factors is robust, it needs complex computations.

The earlier approach to reduce the number of factors arose from observing that often many of them were highly correlated, and one or more of the highly correlated variables could simply be omitted (Charnes, et al., 1988; Jenkins and Anderson, 2003; Kao et al., 1993; Saen et al., 2005). However, which factor(s) should be omitted and which should be retained is rarely obvious. Rather than only looking at the correlation coefficients of inputs or outputs and arbitrarily deciding which factor(s) to be omitted, this study focuses on determining the less influential factor by utilizing the entropy measure of each factor importance.

This paper aims to advance the work on factor omission method to DEA modeling by utilizing the entropy measure of factors importance and is organized as follows. In Section 2, we present the method of entropy measure of factors importance. Section 3 demonstrates the proposed approach by using is a numerical example. Our conclusions are offered in Section 4.

Entropy Measure Of Factor Importance

The efficiency evaluation of DEA is a kind of information processing activity in which efficiency-relevant information is evaluated via inputs/outputs. In this sense, the factors serve as information sources. The more information is emitted by the *i*th factor (i.e., the *i*th information source), the more relevant the factor is in a given efficiency-evaluation situation. Therefore, the importance of a factor can be related to the amount of information that can be transmitted to the efficiency evaluation. This amount of information can be measured by an adapted *entropy measure*.

Suppose there are *n* DMUs to be evaluated by using *s* outputs and *m* inputs, and we denote y_{ik} as the *i*th output and x_{rk} as the *r*th input of DMU_k, where all y_{ik} and x_{rk} are greater than zero. The value set of outputs can be written as

$y^k = (y_{1k}, y_{2k}, \ldots, y_{sk})$ and that of inputs is $x^k = (x_{1k}, x_{2k}, \ldots, x_{mk})$, $k = 1,2,...,n$. To measure the entropy of factors importance, this study scales the data based on the ideal value of each factor. Specifically, if for factor *i* a larger value means better performance (commonly the outputs), we scale the data by the formula $d_{ik} = \frac{y_{ik}}{y_i^*}$, $i = 1,2,...,s$, where y_i^* is denoted as

the ideal value of factor *i* (i.e. $y_i^* = \max_k \{y_{ik}\}$). The d_{ik} is the scaled value on factor *i* of DMU$_k$. Conversely, if for factor *r* a smaller value means better performance (commonly the inputs), we scale the data by the formula $d_{rk} = \frac{x_r^*}{x_{rk}}$, $r = 1,2,...,m$, where x_r^* is denoted as the ideal value of factor *r* (i.e. $x_r^* = \min_k \{x_{rk}\}$). Through the scaling procedure, the value of each d_{ik} should be in the range of zero to one, i.e. $0 < d_{ik} \le 1$, $i = 1,2,...,(s+m)$ and $k = 1,2,...,n$, and all inputs have the same effect orientation as outputs. If $d_{ik} = 1$, it means DMU$_k$ achieves the best performance on factor i among all DMUs. The less divergent the values d_{ik} s are, the less important the *i*th factor becomes. Notably, the purpose of this data scaling technique is to transform all the data in the range of zero to one for comparison.

We use Equation (1) given by Zeleny (1982) for the entropy measure of factor *i*.

$$e(i) = -\frac{1}{\ln n}\sum_{k=1}^{n}[\frac{d_{ik}}{D_i}\ln(\frac{d_{ik}}{D_i})], \quad (1)$$

where $D_i = \sum_{k=1}^{n} d_{ik}$, $i = 1,2,...,(s+m)$, and ln denotes the natural logarithm. Based on Equation (1), we have $0 \le e(i) \le 1$. The larger $e(i)$ is, the less information is transmitted by factor *i* (Zeleny, 1982). If all d_{ik} become identical for a given factor *i*, then $\frac{d_{ik}}{D_i} = \frac{1}{n}$ and $e(i)$ assumes its maximum value and is equal to one. Actually, if $e(i) = 1$, the factor *i* would not transmit any useful information at all for discrimination. If there are two outputs (or inputs), namely factors *i* and *j*, and those are highly correlated, it means that the information of one factor can be substituted adequately by the other. Moreover, if $e(i)$ is larger than $e(j)$, then factor *i* is a less influential factor and it can be omitted to reduce the number of factors.

The entropies of factors provide the information of which highly correlated factor(s) should be omitted. The procedure to identify omission factors is as follows:

Step 1. Calculate the correlation coefficients and the entropies of importance for all factors.

Step 2. Omit the output (or input) factor with a larger value of entropy between the factors with the highest

absolute value of correlation coefficient within outputs (or inputs).

Step 3. Repeat Step 2 until the number of the retained outputs/inputs satisfies some guidelines.

Before determining which factors to omit, the correlation coefficients for all factors are necessary. The omission procedure starts with the two outputs (or inputs) with the highest absolute value of correlation coefficient and we omit the less influential factor, the one with larger entropy, and then repeat this in descending order of absolute correlation coefficient to conclude the omission procedure.

Average intrinsic information of a factor can be measured through the entropy measure. Note that the more distinct and divergent the values, the larger amount of efficiency information is contained in and transmitted by the factor. Any reevaluation of factor values or any addition or removal of DMUs will change the entropies of factors. The importance of each factor thus changes dynamically.

Numerical Example

The purpose of this section is to demonstrate how the proposed approach determines the less influential factors. An example, originally published in the study of Hokkanen and Salminen (1997), is illustrated, and Sarkis (2000) applied a number of different DEA models to the same data with categorization of all criteria to be minimized as DEA inputs and those to be maximized as DEA outputs. There are twenty-two DMUs assessed with five inputs (X1, X2, X3, X4 and X5) and three outputs (Y1, Y2 and Y3) based on Sarkis' categorization. The total number of inputs and outputs does not satisfy any guidelines. We attempt to omit at least one factor in this example, and the original data is presented in Table 1.

Table 2 shows the correlation coefficients of factors highlighting the strong correlation, and Table 3 is the entropies of all factors derived from Equation (1). The strongest correlation coefficient in Table 2 is -0.96 corresponding to inputs X2 and X4. Because X2 and X4 are highly correlated and the entropy of X2 (0.999) is less than that of X4 (1.000), X4 is a less influential factor. We omitted X4 and then the number of efficient DMUs is fifteen (see Table 4), one less than that of the full factor DEA model.

Table 1. The original data of the numerical example

	Inputs					Outputs		
DMU	X1	X2	X3	X4	X5	Y1	Y2	Y3
1	656	552678100	609	1190	670		14	13900
2	786	539113200	575	1190	682		18	23600
3	912	480565400	670	1222	594		24	39767
4	589	559780715	411	1191	443		10	13900
5	706	532286214	325	1191	404		14	23600
6	834	470613514	500	1226	384	6.5	18	40667
7	580	560987877	398	1191	430		10	13900
8	682	532224858	314	1191	393		14	23600
9	838	466586058	501	1229	373	6.5	22	41747
10	579	561555877	373	1191	405		9	13900
11	688	532302258	292	1191	370		13	23600
12	838	465356158	499	1230	361	6.5	17	42767
13	595	560500215	500	1191	538		12	13900
14	709	532974014	402	1191	489		17	23600
15	849	474137314	648	1226	538	6.5	20	40667
16	604	560500215	500	1191	538		12	13900
17	736	532974014	402	1191	489		17	23600
18	871	474137314	648	1226	538	6.5	20	40667
19	579	568674539	495	1193	538		7	13900
20	695	536936873	424	1195	535		18	23600
21	827	457184239	651	1237	513		16	45167
22	982	457206173	651	1239	513		16	45167

Table 2. Correlation coefficients of factors

Factor	X1	X2	X3	X4	X5	Y1	Y2	Y3
X1	1.00							
X2	-0.93	1.00						
X3	0.63	-0.57	1.00					
X4	0.86	-0.96	0.67	1.00				

X5	0.08	0.15	0.63	-0.11	1.00			
Y1						1.00		
Y2						-0.43	1.00	
Y3						-0.49	0.76	1.00

Table 3. The entropies of all factors

Factor	X1	X2	X3	X4	X5	Y1	Y2	Y3
Entropy	0.996	0.999	0.990	1.000	0.995	0.993	0.987	0.969

After omitting X4, the retained factors include four inputs and three outputs satisfying the guideline that the number of inputs and outputs is less than one third of the number of DMUs. To reduce the number of factors and to attain a higher discrimination level in the DEA evaluation, the omission candidates are X1 and X2, since their absolute correlation coefficient is the largest one after omitting X4. The entropies shown in Table 3 indicate that X2 is a less influential factor compared to X1, so that X2 is omitted and the number of efficient DMUs is twelve, with three less efficient DMUs than when omitting only X4. Although omitting (X1, X4) and (X2, X4) render the same number of efficient DMUs, omitting (X2, X4) is a better choice, since its total efficiency score (21.369) is larger than that (21.106) of omitting (X1, X4) (see the last row of Table 5). It means that the retained inputs (X1, X3 and X5) contain more information than that of retained inputs (X2, X3 and X5). If the decision maker wants to omit another factor, the best candidates are Y2 and Y3, since they have strong correlation, with a coefficient of 0.76. Based on the entropies, omitting Y2 is better than omitting Y3. Under the situation of omitting X2, X4 and Y2, the number of efficient DMUs is reduced to five and the total efficiency score attains 20.036. The efficiency scores of DMUs by omitting multiple factors are presented in Tables 5.

The goal of the proposed method is to omit the less influential factors but to retain more information for further ranking. If a DMU becomes inefficient after omitting the less influential factor(s), but it is efficient in the full factor DEA model, it means that this DMU does not have better performance with the retained factors. In other words, this type of DMU is not

truly efficient. Based on the proposed approach to omit the less influential factors, the discrimination level will be increased and the ranking of efficient DMUs will be more efficient in computation.

Table 4. DEA efficiencies of DMUs after omitting one factor

	Efficiency					
DMU	Full factors	Without X1	Without X2	Without X4	Without Y2	Without Y3
1	0.837	0.727	0.837	0.837	0.603	0.837
2	0.871	0.803	0.871	0.871	0.584	0.871
3	1.000	1.000	1.000	1.000	0.891	1.000
4	1.000	1.000	1.000	1.000	1.000	1.000
5	0.991	0.991	0.991	0.987	0.972	0.991
6	0.984	0.984	0.984	0.981	0.981	0.954
7	1.000	1.000	1.000	1.000	1.000	1.000
8	1.000	1.000	1.000	1.000	0.987	1.000
9	1.000	1.000	1.000	1.000	0.990	1.000
10	1.000	1.000	1.000	1.000	1.000	1.000
11	1.000	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000	1.000	0.961
13	1.000	1.000	1.000	1.000	1.000	1.000
14	1.000	1.000	1.000	1.000	0.914	1.000
15	0.978	0.978	0.978	0.960	0.928	0.963
16	1.000	1.000	1.000	1.000	1.000	1.000
17	1.000	1.000	1.000	0.987	0.914	1.000
18	0.978	0.978	0.978	0.959	0.928	0.963
19	1.000	0.998	1.000	1.000	1.000	1.000
20	1.000	0.963	1.000	1.000	0.821	1.000
21	1.000	1.000	1.000	1.000	1.000	1.000
22	1.000	1.000	0.999	1.000	0.999	1.000
Number of efficient DMUs	16	13	15	15	8	15

Table 5. DEA efficiencies of DMUs after omitting multiple factors

DMU	Without (X1 and X2)	Without (X1 and X4)	Without (X2 and X4)	Without (X1, X2 and Y3)	Without (X2, X4 and Y3)	Without (X2, X4 and Y2)	Without (X1, X4 and Y2)
1	0.727	0.623	0.837	0.727	0.837	0.594	0.623
2	0.803	0.714	0.871	0.803	0.871	0.581	0.713
3	1.000	1.000	1.000	1.000	1.000	0.813	1.000
4	1.000	1.000	0.985	1.000	0.985	0.983	1.000
5	0.991	0.986	0.968	0.990	0.968	0.953	0.986
6	0.984	0.981	0.971	0.923	0.917	0.972	0.954
7	1.000	1.000	1.000	1.000	1.000	0.998	1.000
8	1.000	1.000	1.000	1.000	1.000	0.987	1.000
9	1.000	1.000	1.000	1.000	1.000	0.984	1.000
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	0.961	0.961	1.000	0.961
13	1.000	1.000	1.000	1.000	1.000	0.973	1.000
14	1.000	0.970	1.000	1.000	1.000	0.892	0.970
15	0.978	0.959	0.952	0.957	0.909	0.895	0.959
16	1.000	1.000	0.990	1.000	0.990	0.959	1.000
17	1.000	0.970	0.984	1.000	0.984	0.869	0.970
18	0.978	0.959	0.928	0.957	0.887	0.879	0.959
19	0.998	0.984	1.000	0.998	1.000	1.000	0.984
20	0.963	0.960	1.000	0.963	1.000	0.821	0.960
21	1.000	1.000	1.000	0.898	0.806	1.000	1.000
22	0.999	1.000	0.883	0.897	0.706	0.883	1.000
Number of efficient DMUs	13	12	12	11	10	5	11
Total score	21.421	21.106	21.369	21.074	20.821	20.036	21.039

Conclusion

The greater the number of inputs and outputs in a DEA model, the higher the dimensionality of the LP solution space, the less the discrimination level, and the more computational time will be required. In order to achieve a higher discrimination level of the DMUs, a common approach to reduce the number of factors in DEA model is to omit factors highly correlated with those retained. Unfortunately, the research literature rarely specifies the exact logic of which factor(s) to omit and which to retain. This study introduces entropy measure of factors importance to identify the less influential factor(s) suitable for omission. The scaling technique based on the ideal value is analogous to the DEA frame that calculates efficiency scores based on the frontier.

The proposed approach can also be applied to omit factors with a correlation coefficient that is beyond a specific threshold (for example 0.7) to reduce computations for further ranking, even though the total number of inputs and outputs has satisfied some guidelines in a DEA model. Because of omitting factor(s) may loss some information, there is no one methodology can be prescribed as the complete solution to the question of factor selection in DEA modeling. This paper provides another viewpoint of data reduction for factor(s) omission by utilizing the entropy measure of factor importance. Based on the entropy measures associated with the correlation coefficients of factors, decision makers can make an appropriate omission decision.

References

Adler, N., and 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, 260-273.

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

Charnes,.A., Cooper, W. W., and Sueyoshi, T., 1988. A goal programming/constrained regression review of the bell system breakup. *Management Science* 34, 1-26.

Chen, T.H., and Bao, C.P., 2013. Applying cost-reliability analysis to improve system reliability, *Jour-*

nal of Industrial and Production Engineering, 30(7), 467-472.

Chen, T.H., 2011. Performance measurement in a small Taiwanese hotel chain, *Cornell Hospitality Quarterly* 52(3), 354-362.

Chang, S.Y., and Chen, T.H., 2008. Performance ranking of Asian lead frame firms: A slack-based method in data envelopment analysis, *International Journal of Production Research* 46, 3875-3885.

Dyson, R G, and Allen, R., 2001, Camanho A S, Podinovski V V, Sarrico C S. Pitfall and protocols in DEA. *European Journal of Operational Research* 132, 245-259.

Friedman, L., and Sinuany-Stern, Z., 1998. Combining ranking scales and selecting variables in DEA context: The case of industrial branches. *Computers and Operations Research* 25(9), 781-791.

Hokkanen, J., and Salminen, P., 1997. Choosing a solid waste management system using multi-criteria decision analysis. *European Journal of Operational Research* 98, 19-36.

Jenkins, L, and 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.

Kao, C, Chang, P., and Hwang, S. N., 1993. Data envelopment analysis in measuring the efficiency of forest management. *Journal of Environmental Management* 38, 73-83.

Saen, R. F., Memariani, A., and Lotfi, F. H., 2006. The effect of correlation coefficient among multiple input vectors on the efficiency mean in data envelopment analysis. *Applied Mathematics and Computation* 162, 503-521.

Sarkis, J., 2000. A comparative analysis of DEA as a discrete alternative multiple criteria decision tool. *European Journal of Operational Research* 123, 543-557.

Ueda, T,, and Hoshiai, Y.,1997. Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs, *Journal of Operations Research Society of Japan* 40 (4), 466-478.

Wagner, J. M., and Shimshak, D. G., 2007. Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives. *European Journal of Operational Research* 180, 55-67.

Zeleny, M., 1982. *Multiple criteria decision making*. McGraw-Hill: New York; 1982.

Zhu, J.,1998. Data envelopment analysis vs. principal components analysis: An illustrative study of economic performance of Chinese cities. *European Journal of Operational Research* 111, 50-61.