Journal of the Operational Research Society (2002) 53, 985–991 0160-5682/02 $15.00

www.palgrave-journals.com/jors

Including principal component weights to improve discrimination in data envelopment analysis

N Adler[1] and B Golany[2]*

[1]*Hebrew University of Jerusalem, Jerusalem, Israel; and* [2]*Israel Institute of Technology, Haifa, Israel*

This research further develops the combined use of principal component analysis (PCA) and data envelopment analysis (DEA). The aim is to reduce the curse of dimensionality that occurs in DEA when there is an excessive number of inputs and outputs in relation to the number of decision-making units. Three separate PCA–DEA formulations are developed in the paper utilising the results of PCA to develop objective, assurance region type constraints on the DEA weights. The first model applies PCA to grouped data representing similar themes, such as quality or environmental measures. The second model, if needed, applies PCA to all inputs and separately to all outputs, thus further strengthening the discrimination power of DEA. The third formulation searches for a single set of global weights with which to fully rank all observations. In summary, it is clear that the use of principal components can noticeably improve the strength of DEA models.
Journal of the Operational Research Society (2002) **53**, 985–991. doi:10.1057/palgrave.jors.2601400

Keywords: data envelopment analysis; principal component analysis; performance measurement; assurance regions; ranking

Introduction

The aim of this paper is to further develop the combined use of principal component analysis (PCA) and data envelopment analysis (DEA). The idea to combine these two methodologies was developed independently by Ueda and Hoshiai[1] and Adler and Golany.[2] The goal of the PCA–DEA model is to improve the discriminatory power within DEA, which often fails when there is an excessive number of inputs and outputs in relation to the number of decision-making units (DMUs). When this occurs, the Pareto frontier may be defined by a large number of DMUs, ie a large proportion of the units are considered efficient. However, whilst principal component analysis can reduce the dimensions of inputs and/or outputs, this refinement may still be insufficient to provide an adequate ranking of the DMUs. Therefore, we develop in this paper an extended PCA–DEA model with constraints aimed at further strengthening the model's discriminatory power and answering some of the problems that cropped up in the previous research.

The following section describes both DEA and PCA. Subsequently PCA–DEA constraint-modified models are developed and demonstrated through a well-known example drawn from the literature. Three separate formulations are presented. The first formulation introduces assurance regions using PCA weights as objectively based constraints. The second and third formulations apply PCA to all inputs and separately to all outputs. In the second model assurance region constraints are applied and discrimination is maximized with none to minimal loss of information. The third formulation computes one set of global weights enabling a complete ranking of all DMUs. Finally, the last section presents a summary and conclusions of this research.

**Correspondence*: *B Golany, Faculty of Industrial Engineering and Management, The Technion, Israel Institute of Technology, Haifa, Israel.*
E-mail: golany@ie.technion.ac.il

Data envelopment analysis and principal component analysis

DEA is a technique that measures the relative efficiency of DMUs with multiple inputs and outputs but with no obvious production function to aggregate the data in its entirety. Since it may be unclear how to quantitatively combine the various inputs consumed and outputs produced, the DEA methodology computes the relative efficiency of each DMU individually through the use of weighted averages. The additive model (see Charnes *et al*[3]), used here for its translation invariance property, reflects all inefficiencies identified in both inputs and outputs simultaneously. By comparing n units with q outputs denoted by Y and r inputs denoted by X, the efficiency measure for unit a is expressed as in model (1), where both the envelopment side and multiplier side formulations are presented.

$$\begin{array}{llll} \max\limits_{s,\sigma} & e^t s + e^t \sigma & \min\limits_{V,U} & VX^a - UY^a \\ \text{s.t.} & Y\lambda - s = Y^a & \text{s.t.} & VX - UY \geqslant 0 \\ & -X\lambda - \sigma = -X^a & & V \geqslant e \\ & \lambda,\ s,\ \sigma \geqslant 0 & & U \geqslant e \end{array} \tag{1}$$

In (1), λ represents a vector of DMU weights chosen by the linear program, $e\,(e^t)$ a (transposed) vector of ones, σ and s vectors of input and output slacks, respectively, and X^a and

Y^a the input and output column vectors for DMU_a respectively. Conventional wisdom refers to output slacks as s^+ and input slacks as s^-, which have been replaced in this formulation with s and σ, respectively, the purpose of which becomes apparent in the subsequent models.

PCA explains the variance structure of a matrix of data through linear combinations of variables, consequently reducing the data to a few principal components, which generally describe 80–90% of the variance in the data. If most of the population variance can be attributed to the first few components, then they can replace the original variables with minimum loss of information. As stated in Hair *et al*,[4] let the random vector $X = [X_1, X_2, \ldots, X_p]$ (ie the original inputs/outputs chosen to be aggregated) have the correlation matrix C with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ and normalised eigenvectors $l_1, l_2, \ldots, l_p$. Consider the linear combinations, where the superscript t represents the transpose operator:

$$\begin{aligned} &X_{\text{PC}_i} = l_i^t X = l_{1i}X_1 + l_{2i}X_2 + \cdots + l_{pi}X_p \\ &\text{Var}(X_{\text{PC}_i}) = l_i^t C l_i, \qquad i = 1, 2, \ldots, p \\ &\text{Correlation}(X_{\text{PC}_i}, X_{\text{PC}_k}) = l_i^t C l_k, \quad i = 1, 2, \ldots, p, \\ &\qquad\qquad k = 1, 2, \ldots, p \end{aligned}$$

The principal components (PCs) are the uncorrelated linear combinations $X_{\text{PC}_1}, X_{\text{PC}_2}, \ldots, X_{\text{PC}_p}$ ranked by their variances in descending order. The PCA used here is based on correlation rather than covariance, as the variables used in DEA are often quantified in different units of measure.

In general, inputs and outputs of a DEA need to be strictly positive; however, PCs can contain negative values. Since the additive model accepts both negative and zero values (see Pastor[5]), we will base our analyses on this formulation thus avoiding the necessity for further data manipulation.

The PCA–DEA constrained model formulation

This section first develops the PCA–DEA additive model formulation and then discusses the PCA–DEA constrained model. Finally, the last section provides a small numerical illustration drawn from the literature comparing departments at various English universities.

PCA–DEA model

In order to use PC scores instead of the original data, the DEA model needs to be transformed to take into account the use of linear aggregation of data. Indeed, the formulation presented ensures that were we to use all the PCs, we would attain the same solution as that achieved under the original DEA formulation. When using both original inputs and outputs and some aggregated variables, separate $X = \{X_o, X_{Lx}\}$ and $Y = \{Y_o, Y_{Ly}\}$, where $X_o(Y_o)$ represent inputs (outputs) whose original values will be used in the subsequent DEA. X_{Lx} (Y_{Ly}) represents inputs (outputs) whose values will be transformed through PCA. This separation of variables should be based on a logical understanding of the specific application under question. For example, in an application where outputs could be separated into direct explanations of resources, such as capital and labour, and indirect explanations such as quality or environmental measures, it would be natural to apply PCs to the indirect variables, which are often highly correlated and could be reduced with minimal to no loss of information. Another example can be drawn from an air transport application described in Adler and Golany,[2] in which the input variables could be logically split into three separate areas, namely sets of transportation, comfort and size variables. Each group could be handled individually as the correlation between variables within the group is naturally high.

Let $L_x = \{l_{ij}^x\}$ be the matrix of the PCA linear coefficients of input data and let $L_y = \{l_{ij}^y\}$ be the matrix of the PCA linear coefficients of output data. Now, $X_{\text{PC}} = L_x X_{Lx}$ and $Y_{\text{PC}} = L_y Y_{Ly}$ are weighted sums of the corresponding original data, X_{Lx} and Y_{Ly}. We can consequently replace model (1) with model (2).

$$\begin{aligned} \min_{s_{\text{PC}}^+, s_{\text{PC}}^-, s_o, \sigma_{\text{PC}}^+, \sigma_{\text{PC}}^-, \sigma_o} \quad & -e^t s_o - e^t L_y^{-1}(s_{\text{PC}}^+ - s_{\text{PC}}^-) - e^t \sigma_o \\ & - e^t L_x^{-1}(\sigma_{\text{PC}}^+ - \sigma_{\text{PC}}^-) \\ \text{s.t.} \quad & Y_o\lambda - s_o = Y_o^a \\ & Y_{\text{PC}}\lambda - (s_{\text{PC}}^- - s_{\text{PC}}^+) = Y_{\text{PC}}^a \\ & -X_o\lambda - \sigma_o = -X_o^a \\ & -X_{\text{PC}}\lambda - (\sigma_{\text{PC}}^- - \sigma_{\text{PC}}^+) = -X_{\text{PC}}^a \\ & s_o,\ s_{\text{PC}}^-,\ s_{\text{PC}}^+,\ \sigma_o,\ \sigma_{\text{PC}}^-,\ \sigma_{\text{PC}}^+,\ \lambda \geqslant 0 \end{aligned} \qquad (2)$$

The terms L_x^{-1} (L_y^{-1}) in (2) represent the inverse matrix of input (output) weights attained through the PCA. Since the adapted data matrices, X_{PC} and Y_{PC}, can contain negative values, the new slacks do not need to be non-negative, hence we represent them by the difference between two, auxiliary, non-negative variables, namely s_{PC}^-, s_{PC}^+ and σ_{PC}^-, σ_{PC}^+.

It is worthwhile noting that the original additive model of Charnes *et al*[3] was extended by Cooper *et al*[6] to a weighted additive model whereby weights were introduced into the objective function in order to counter bias that might occur due to differences in magnitude of the values of the original variables. Along the same line, the new slacks in the objective function of model (2) are weighted in order to counter the transformation and ensure an equivalent problem/solution to that of the original linear program. In addition, two more constraints must be included in the formulation to maintain equivalence with model (1):

$$\begin{aligned} L_y^{-1}(s_{\text{PC}}^+ - s_{\text{PC}}^-) \geqslant 0 \\ L_x^{-1}(\sigma_{\text{PC}}^+ - \sigma_{\text{PC}}^-) \geqslant 0 \end{aligned}$$

These constraints relate to the new slacks, which are relevant only to the PC data. The PCA–DEA formulation is exactly equivalent to the original additive model when the PCs explain 100% of the correlation in the original input and output matrices. Clearly, if we use less than full information, we will lose some of the explanatory powers of the data but we will improve the discriminatory power of the model. Furthermore, since DEA is a frontier-based methodology, it is particularly affected by data errors in measurement. To some extent, this may be avoided by removing some PCs.

The only other methodology for reducing dimensionality known to the authors is to drop specific variables. By doing so, all information drawn from those variables are automatically lost. By removing certain PCs, we will not lose an entire variable, unless the PC weight is placed entirely on that variable in the PC dropped, with a zero weight in all other PC combinations.

Writing the dual of model (2) for DMU_a, we attain the following formulation, in which the effect of the PCA can be seen directly through the weights L_x (L_y).

$$\begin{aligned} \text{Min} \quad & V_o X_o^a + V_{PC} X_{PC}^a - U_o Y_o^a - U_{PC} Y_{PC}^a \\ \text{s.t.} \quad & V_o X_o + V_{PC} X_{PC} - U_o Y_o - U_{PC} Y_{PC} \geqslant 0 \\ & V_o \geqslant e \\ & U_o \geqslant e \\ & V_{PC} L_x \geqslant e \\ & U_{PC} L_y \geqslant e \\ & V_{PC}, U_{PC} \text{ free} \end{aligned} \tag{3}$$

Clearly, not only will U_o and V_o be computed, but also U_{PC} and V_{PC}, which can then be translated backwards to compute the 'original' weights. If all PCs are utilised, the weights should be the same as the original model, however given the potential in DEA for degeneracy this may not always be the case. See the end of the results section for an example of this phenomenon. Even if all PCs are not accounted for, the weights can still be evaluated backwards on the original data, ensuring a complete DEA result that is interpretable for inefficient units in the standard manner.

PCA–DEA constrained model

The use of PCA, whilst useful in reducing the problems of dimensions, may still be insufficient in providing an adequate ranking of the DMUs. Therefore, the introduction of constraints on the PC weights within DEA (V_{PC} and U_{PC}) may further aid the discrimination capabilities of the analysis. One of the major concerns with the assurance region and cone-ratio models is the additional preference information that needs to be drawn from the decision makers in order to accomplish this task. This is particularly problematic given the less-than-opaque understanding of the DEA weight variables. Consequently, the objective use of weight constraints based on the output of the PCA may be considered helpful. This section will be separated into three parts. The first section discusses the addition of assurance region constraints to the PCA–DEA model, utilising PCA weight information. The second model applies PCA to all inputs and outputs separately and then utilises constraints. The last section compares and contrasts the DMUs in the spirit of Cook and Kress[7], utilising a discrimination factor to produce a single set of weights.

PCA–DEA partially constrained model. Several sets of weight constraints could be considered for this analysis, such as that of assurance regions, first developed in Thompson *et al.*[8] Golany[9] proposed a version of assurance regions with ordinal constraints on the weights. Other researchers followed with various extensions and the topic is reviewed in Cooper *et al.*[10] Assurance region constraints introduce bounds on the DEA weights, for example to include all variables through a strictly positive value or to incorporate 'decision makers' knowledge. In the light of the fact that PCA prioritises the PCs in descending order of importance, additional constraints could simply require the weight of PC_1 to be at least that of PC_2, the weight of PC_2 to be at least that of PC_3 and so on, as specified in Equation (4).

$$\begin{aligned} V_{PC_i} - V_{PC_i+1} \geqslant 0 \\ U_{PC_i} - U_{PC_i+1} \geqslant 0 \end{aligned} \quad \text{for } i = 1, \ldots, m-1, \tag{4}$$

where m PCs are analysed.

Equation (4) applies only to those weights associated with PC-modified data and only the PCs chosen to be included in the analysis.

Complete PCA–DEA constrained model. An alternative to the previous section would be to apply PCA to all inputs and outputs separately, subsequently including all PCs in the adapted model ensuring equivalence to the standard DEA model with no loss of information. The set of constraints described in Equation (4) could then be applied to all PC variables, requiring only objective information and improving the discriminatory power of the original model. This is shown clearly in Equation (5), where we present the standard additive model and the complete PCA–DEA constrained models.

$$\begin{aligned} \min_{V,U} \quad & VX^a - UY^a \\ \text{s.t.} \quad & VX - UY \geqslant 0 \\ & V \geqslant e \\ & U \geqslant e \end{aligned} \qquad \begin{aligned} \min_{VL_y, UL_x} \quad & V_{PC} X_{PC}^a - U_{PC} Y_{PC}^a \\ \text{s.t.} \quad & V_{PC} X_{PC} - U_{PC} Y_{PC} \geqslant 0 \\ & V_{PC} L_x \geqslant e \\ & U_{PC} L_y \geqslant e \\ & V_{PC_i} - V_{PC_{i+1}} \geqslant 0 \\ & U_{PC_i} - U_{PC_{i+1}} \geqslant 0 \end{aligned} \tag{5}$$

where $V = V_{PC} L_x$ and $U = U_{PC} L_y$. Clearly PCs adding little to no information to the analysis could be dropped and the model reduced accordingly with minimal loss of information. Nasierowski and Arcelus[11] suggested a different approach in their paper, which used statistical techniques to support the development of measures aimed at assessing national productivity. They utilise a four-step procedure, defined in Hair *et al*,[4] which applies PCA as an external instrument to identify important variables, thus reducing the number of variables utilised in a subsequent analysis based on objective, statistical information.

Maximum discrimination in the PCA–DEA formulation. A third approach, applicable when PCA is applied to all inputs and separately to all outputs, follows the modelling approach of Cook and Kress.[7] In their paper they aim to maximise the differentiation capabilities of DEA through the use of a discrimination factor, ε. The result will be a single set of weights with the aim of ranking the DMUs as fully as possible.

$$\begin{aligned} \text{Max} \quad & \varepsilon \\ \text{s.t.} \quad & V_{PC}X_{PC} - U_{PC}Y_{PC} \geqslant 0 \\ & V_{PC}L_x \geqslant e \\ & U_{PC}L_y \geqslant e \\ & V_{PC_i} - V_{PC_{i+1}} \geqslant \varepsilon \\ & U_{PC_i} - U_{PC_{i+1}} \geqslant \varepsilon \\ & V_{PC_1} = 1 \end{aligned} \quad (6)$$

where ε represents the discriminating variable. Under model (6) we maximise the discriminatory power of PCA–DEA through objective constraints on the PC-modified data. The normalisation condition, $V_{PC_1} = 1$ is required in order to avoid unbounded solutions. The right-hand side of this equation can be increased up to a very large upper bound without affecting the ordinal ranking of the DMUs. We further note that other normalisation schemes can be implemented, including $\sum V_{PC} = 1$ or $\sum V_{PC} + \sum U_{PC} = 1$, but the behaviour of the ranking as a result of these conditions must still be investigated.

Example of the PCA–DEA constrained model

In order to illustrate the potential of these models, we will present the results of a simple dataset existing in the literature. The numerical illustration is drawn from Wong and Beasley[12] in which seven university departments were compared over six variables. The original data is presented in Table 1 and the results of the PCs analysis in Table 2.

Table 2 PCA linear coefficients of input and output data based on correlation

Inputs L_x			*Outputs L_y*		
PC_1	*PC_2*	*PC_3*	*PC_1*	*PC_2*	*PC_3*
0.5687	0.6316	0.5269	0.5759	0.7919	−0.2033
0.6184	0.0941	−0.7802	0.5788	−0.2193	0.7854
0.5424	−0.7696	0.3370	0.5774	−0.5700	−0.5846
Variance explained					
86.7866	13.1647	0.0487	94.5568	2.9892	2.4541

In this example, it could be argued that a single PC input and a single PC output explain the vast majority of the variance in the original data matrices. It should be noted here that the results of a PCA are unique up to a sign and must therefore be chosen carefully to ensure feasibility of the subsequent PCA–DEA model (Dillon and Goldstein[13]).

Table 3 presents the results of various DEA models, including the standard additive model, three PC-constrained, two PC-constrained, one PC and the subjectively constrained results of Wong and Beasley[12] for comparison.

In the basic additive and three PC-unconstrained models, given the excessive number of variables compared to DMUs, only one university department (DMU_4) is deemed inefficient. Using all three PCs with constraints, a slightly greater degree of differentiation is attained, such that two DMUs are now considered inefficient. It is clear that the three-PC model without constraints would result in the exact same ranking as the basic additive model (they are algebraically equivalent problems), consequently the addition of constraints has improved the differentiation whilst taking into consideration all variables in the model. Furthermore, the standard input and output weights can be evaluated easily by multiplying the PC weights with V_{PC} and U_{PC} drawn from the PCA–DEA formulation ($V = V_{PC}L_x$ and $U = U_{PC}L_y$).

Table 1 Raw data from university departments example

	Inputs			*Outputs*		
DMU	*No. of academic staff*	*Academic staff salaries*	*Support staff salaries*	*No. of undergraduate students*	*No. of postgraduate students*	*No. of research papers*
1	12	400	20	60	35	17
2	19	750	70	139	41	40
3	42	1500	70	225	68	75
4	15	600	100	90	12	17
5	45	2000	250	253	145	130
6	19	730	50	132	45	45
7	41	2350	600	305	159	97

Table 3 Efficiency scores of various DEA formulations

	Additive model (non-normalised scores), (efficient DMU = 0, inefficient DMU = positive value)				*CCR (efficient DMU = 1, inefficient DMU = (0,1))*
DMU	*Original additive and three-PC unconstrained*	*Three PCs on both sides + constraints*	*Two PCs on both sides + constraints*	*One PC on each side*	*Subjectively constrained*
1	0.0	0.0	0.0	33.3	1
2	0.0	0.0	0.0	51.1	0.995
3	0.0	0.0	298.2	328.8	0.862
4	202.6	202.6	235.7	316.5	0.691
5	0.0	0.0	0.0	430.6	1
6	0.0	0.0	0.0	0	1
7	0.0	671.4	885.8	1041.6	1

For greater differentiation, two PCs (on both the input and output side) could be used, explaining over 97% of the variance of each set of variables. This results in three inefficient DMUs. Finally, if we compute the model with one PC on each side, only DMU_6 remains efficient whilst still explaining over 86% of the variance in the original input data and over 94% on the output side. Indeed the choice of DMU_6 as being the only efficient DMU matches results presented in Li and Reeves,[14] in which a multiple objective linear program model is developed using a minimax objective function.

When applying model (6) to the example, we attain the ordinal ranking presented in Table 4. The table presents the ranked DMUs according to a single set of input and output weights, as specified in Table 5.

As can be seen in Table 4, the ranking is similar to the results of the other models, with DMU_6 ranked highest and DMU_4 and DMU_7 ranked lowest. From Table 5, we can see that all original variables are considered in the analysis with a weight of at least one, as required in model (6). We can also view the PCA weights in columns one and two, whereby the weight on PC_1 is larger than that of PC_2 and so on.

Finally, we turn to the issue of multiple optimal solutions that was discussed previously. The numerical illustration was solved using the standard additive model and the second PCA–DEA formulation, with complete, unconstrained PCs. From Table 6 one can see that the same efficiency scores can be computed with different sets of weights. The first row shows the efficiency score achieved from both formulations, as they are algebraically identical. The next six rows specify the weights on the original inputs and outputs evaluated in the PCA–DEA formulation ($V = V_{PC}L_X$ and $U = U_{PC}L_Y$). The last six rows show the weights computed per DMU in the standard additive model. Whilst the efficiency scores are the same, the choice of weights is not, showing the potential for multiple optimal solutions.

Table 4 Results of maximum discrimination model

DMU	*Rank according to the ratio value*	*Ratio value UY/VX*
DMU_6	1	1
DMU_2	2	0.9243
DMU_3	3	0.8671
DMU_1	4	0.8419
DMU_5	5	0.6961
DMU_4	6	0.5775
DMU_7	7	0.4859

Summary and conclusions

This paper has shown how PCA may be utilised within the DEA context to noticeably improve the discriminatory power of the model. Frequently, an extensive number of decision making units are considered relatively efficient within DEA, due to an excessive number of inputs and outputs relative to the number of units. Assurance regions and cone-ratio constraints are often used in order to restrict the variability of the weights, thus improving the discrimi-

Table 5 Single set of weights used to fully rank DMUs

V_{PC} values		*U_{PC} values*		*'Original' V values ($V = V_{PC}L_x$)*		*'Original' U values ($U = U_{PC}L_y$)*	
V_{PC_1}	10	U_{PC_1}	10.4436	V_1	10.1569	U_1	9.0374
V_{PC_2}	1.8931	U_{PC_2}	2.3367	V_2	1.5138	U_2	1
V_{PC_3}	−6.2139	U_{PC_3}	−5.7702	V_3	6.0614	U_3	8.0716

Table 6 PCA–DEA solution and additive DEA solution

	DMU_1	DMU_2	DMU_3	DMU_4	DMU_5	DMU_6	DMU_7
Efficiency score (efficient DMU = 0, otherwise positive)	0.000	0.000	0.000	202.591	0.000	0.000	0.000
Using complete PC variables without constraints:							
V_1	1.000	1.000	1.000	1.000	1.000	1.000	70.444
V_2	1.000	1.609	1.000	1.000	1.071	1.000	1.000
V_3	1.000	1.000	18.106	1.000	1.000	1.000	1.000
U_1	3.905	8.741	11.851	5.371	1.000	4.547	1.000
U_2	5.032	1.000	1.000	1.000	7.158	2.209	33.023
U_3	1.268	1.000	1.000	1.000	8.817	2.209	2.912
Using the standard additive model:							
V_1	12.334	61.835	1.107	1.000	28.176	19.494	104.603
V_2	1.394	1.296	1.004	1.000	1.319	1.403	1.239
V_3	3.008	1.553	21.928	1.000	1.738	2.631	1.314
U_1	3.729	13.899	7.860	5.371	2.777	5.768	8.908
U_2	12.629	4.006	2.619	1.000	8.197	5.784	29.094
U_3	5.874	3.977	15.206	1.000	18.834	11.207	6.656

natory power of the standard DEA formulations. However, use of these techniques requires additional, preferential information, which is often difficult to attain. This research has suggested an objective solution to the problem.

The first model presented uses PCA on sets of inputs or outputs and reduces the number of variables in the model to those PCs that prove significant, with for example an eigenvalue greater than 1.8. See Dillon and Goldstein[13] for other stopping-rule suggestions. Some information will be lost, however this is minimised through the use of PCA and does not require the decision maker to remove complete variables from the analysis or specify which variables are more important. Consequently, the adapted formulation avoids losing all the information included within a variable that is then dropped, rather losing a principal component explaining by definition very little of the correlation between the original grouped data. Additional strength can be added to the model by including constraints over the PC variables, whereby the weight on the first PC (that explains the greatest variation in the original dataset) must be at least equal to that of the second PC and so on. These assurance region type constraints are based on objective information alone. Furthermore, multiplying the new PCA–DEA weights by the PC weights computes ‘original’ DEA weights, thus providing complete DEA information.

The second model suggested is to apply PCA to all inputs and separately to all outputs. The PCA–DEA model can then be constrained in the same manner as discussed previously, this time over all PCs, thus improving the discriminatory power of the model with no loss of information. If required, PCs can be dropped and the remaining PCs constrained to further improve discriminatory power. The third formulation searches for a common set of weights to fully rank the units through the use of a discrimination factor and a set of objective, PC-based constraints.

The three models are demonstrated using a well-known example drawn from Wong and Beasley[12] in which subjective constraints resulted in four efficient units of the seven considered. The formulations presented here lead to the identification of DMU_6 as the single most efficient department amongst the small dataset.

Acknowledgements—The authors would like to thank two anonymous referees for very helpful comments on an earlier version of this paper. N.A. would also like to thank the Recanati Foundation for partial support of this research.

References

1 Ueda T and Hoshiai Y (1997). Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. *J Opl Res Soc Jpn* **40**: 466–478.
2 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. *Eur J Opl Res* **132**: 18–31.
3 Charnes A, Cooper WW, Golany B, Seiford L and Stutz J (1985). Foundations of data envelopment analysis for Pareto–Koopmans efficient empirical production functions. *J Econometrics* **30**: 91–107.
4 Hair JF, Anderson RE, Tatham RL and Black WC (1995). *Multivariate Data Analysis*. Prentice-Hall: Englewood Cliffs, NJ.
5 Pastor J (1996). Translation invariance in data envelopment analysis: a generalization. *Ann Opns Res* **66**: 93–102.
6 Cooper WW, Park KS and Pastor JT (1999). RAM: a range adjusted measure of inefficiency for use with additive models and relations to other models and measures in DEA. *J Prod Anal* **11**: 5–42.
7 Cook WD and Kress M (1990). A data envelopment model for aggregating preference rankings. *Mngt Sci* **36**: 1302–1310.
8 Thompson RG, Singleton FD, Thrall RM and Smith BA (1986). Comparative site evaluations for locating a high-energy lab in Texas. *Interfaces* **16**: 35–49.

9 Golany B (1988). A note on including ordinal relations among multipliers in data envelopment analysis. *Mngt Sci* **34**: 1029–1033.

10 Cooper WW, Seiford LM and Tone K (2000). *Data Envelopment Analysis: A Comprehensive Text with Models*, *Applications*, *References and DEA-Solver Software*. Kluwer Academic: Boston, MA.

11 Nasierowski W and Arcelus FJ (1999). Interrelationships among the elements of national innovation systems: a statistical evaluation. *Eur J Opl Res* **119**: 235–253.

12 Wong YHB and Beasley JE (1990). Restricting weight flexibility in data envelopment analysis. *J Opl Res Soc* **41**: 829–35.

13 Dillon WR and Goldstein M (1984). *Multivariate Analysis*: *Methods and Applications*. J Wiley: New York.

14 Li X-B and Reeves GR (1999). A multiple criteria approach to data envelopment analysis. *Eur J Opl Res* **115**: 507–517.

Received August 2001;
accepted March 2002 after one revision