Vol. 34 (2014)
No. 3

数 学 杂 志
J. of Math. (PRC)

ROBUST DATA ENVELOPMENT ANALYSIS WITH UNCERTAIN DATA

YU Hong-xia[1], JIN Li[2]

(*1.School of Math. & Physics, University of Shanghai Electric Power, Shanghai 201300, China*)

(*2.School of Math., Physics and Information Science, Zhejiang Ocean University, Zhoushan 316004, China*)

Abstract: In this paper, we study the data envelopment analysis (DEA) model with uncertain output parameters. By using robust optimization method, a robust DEA model is presented, numerical experiment indicates that this model is reliable for efficiency estimating and ranking strategies. This robust DEA model can deal with unknown-but-bounded uncertainty, in which the distributions of the random data entries are permitted to be unknown. Compared with the existing model which considers uncertain data with symmetrical distribution only, the robust DEA model suggested in the paper has wider applications.

Keywords: data envelopment analysis; uncertainty; robust optimization

2010 MR Subject Classification: 90C90; 90C05

Document code: A **Article ID:** 0255-7797(2014)03-0423-09

1 Introduction

Data envelopment analysis (DEA) is a mathematical programming methodology for evaluating and measuring the relative efficiencies of a set of decision making units (DMUs) that use multiple inputs to produce multiple outputs. Due to its solid underlying mathematical basis and wide applications to real-world problems, much effort has been devoted to the DEA methods since the pioneering work of [1, 2] summarized the major research in DEA over the last 30 years, which provided a good research framework.

In the conventional DEA models, all the data are assumed to have the form of specific numerical values which are "known exactly". However, this situation may not always be true. In reality, the data of real-world problems more often than not are uncertain—not known exactly at the time the problem is being solved. In applications one cannot ignore the possibility that even a small uncertainty in the data can make the nominal optimal solution to the problem completely meaningless from a practical viewpoint. [3] showed that a small perturbation on data of linear programming problem could lead heavily infeasible solution.

* **Received date:** 2012-12-02 **Accepted date:** 2013-06-13

Foundation item: Supported by Major Program of National Natural Science Foundation of China (70732003).

Biography: Yu Hongxia(1978–), female, born in Liaoning, Ph.D., major in robust optimization and its applications.

As a linear programming based approach, DEA will never be able to escape from the impact of uncertainty, i.e., a small perturbation on output data of DMU could make a big change on the efficiencies, so the results of the ranking could be unreliable. Consequently, there exists a real need of a methodology capable of generating a robust solution, one that is immunized against the effect of data uncertainty.

In this paper, we consider the perturbation in output data and propose a new robust DEA model based on the adaptation of recently developed robust optimization approaches proposed by [4]; [5] proposed a robust DEA model with consideration of uncertainty on output parameters for the performance assessment of electricity distribution companies. However, the assumption of symmetric data uncertainty made in their paper could be too restrictive for many real-world applications; [6] considered the DEA with uncertain data, they proposed a second order cone model. In this paper, we relax the assumption of symmetric data uncertainty and construct the robust formulation for asymmetric data uncertainty, the model we present is a linear programming problem.

This paper is organized as follows. We first present the fundamentals of robust optimization in Section 2. In Section 3, we illustrate the classical DEA model and propose the robust DEA model when the output data are uncertain. In Section 4, we demonstrate some experimental results. Finally, in Section 5, we sum up our conclusions.

2 Robust Optimization

To present the robust optimization method, consider a linear programming problem

$$\begin{aligned} \max \quad & c^T x, \\ \text{s.t.} \quad & Ax \le b, \\ & l \le x \le u, \end{aligned} \tag{2.1}$$

where $c, l, u \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $A = (a_{ij})$ is a $m \times n$ matrix, and $x \in \mathbb{R}^n$ is the vector of decision variables. We assume, without loss of generality, that only the elements of the matrix A are subject to uncertainty. In fact, if c and b are also uncertain, we can rewrite the problem as

$$\begin{aligned} \max \quad & \tilde{c}^T \tilde{x}, \\ \text{s.t.} \quad & \tilde{A}\tilde{x} \le 0, \\ & \tilde{l} \le \tilde{x} \le \tilde{u} \end{aligned}$$

with $\tilde{x} = (z, x, y)$, $\tilde{c} = (1, 0, 0)$, $\tilde{A} = \begin{pmatrix} 0 & A & -b \\ 1 & -c^T & 0 \end{pmatrix}$, $\tilde{l} = (z_L, l, 1)$, and $\tilde{u} = (z_U, u, 1)$, where z_L and z_U are finite constants as long as all the components of c are bounded. In this new formulation, only the matrix $\tilde{A}$ contains uncertain data.

Consider a particular row i of the matrix A and let J_i be the set of coefficients in row i that are subject to uncertainty. Each entry $a_{ij}, j \in J_i$ is modeled as a symmetric and bounded random variable that takes values in $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$.

Given the data uncertainty structure for A, the traditional linear optimization methodology elect to solve the nominal formulation, where each random a_{ij} is replaced by its mean value:

$$\begin{aligned} \max \quad & c^T x, \\ \text{s.t.} \quad & \sum_j \bar{a}_{ij} x_j \le b_i, \quad \forall i, \\ & l \le x \le u. \end{aligned} \tag{2.2}$$

Small data uncertainty is just ignored as if the given ("nominal") data were exact, and the resulting nominal solution is what recommended for use, in hope that small data uncertainties will not affect significantly the feasibility and optimality properties of this solution. However, this hope is not necessarily justified. The analysis of linear optimization problems from the NETLIB collection reported in Ben-Tal & Nemirovski [3] revealed that for 13 of 94 NETLIB problems, random 0.01percent perturbations of the uncertain data can make the nominal optimal solution severely infeasible: with a non-negligible probability, it violates some of the constraints by 50 percent and more. Consequently, this leads one to consider a solution that is guaranteed to satisfy $Ax \le b$ for all realizations of the random a_{ij}'s, while maximizing the objective value. To obtain such a solution, Soyster [7] proposed the following robust formulation :

$$\begin{aligned} \max \quad & c^T x, \\ \text{s.t.} \quad & \sum_j \bar{a}_{ij} x_j + \sum_{j \in J_i} \hat{a}_{ij} y_j \le b_i, \quad \forall i, \\ & -y_j \le x_j \le y_j, \quad \forall j, \\ & l \le x \le u, \\ & y \ge 0. \end{aligned} \tag{2.3}$$

It can be proved that the optimal solution of (2.3) is a feasible solution of (2.1) for every possible realization of A. Although the Soyster's method provides the most robust solution, it is also the most conservative in practice in the sense that the robust solution has an objective function value much worse than the objective function value of the solution of the nominal linear optimization problem. To address this conservatism, [8] proposed the following robust problem:

$$\begin{aligned} \max \quad & c^T x, \\ \text{s.t.} \quad & \sum_j \bar{a}_{ij} x_j + \sum_{j \in J_i} \hat{a}_{ij} y_j + \Omega_i \sqrt{\sum_{j \in J_i} \hat{a}_{ij}^2 z_{ij}^2} \le b_i, \quad \forall i, \\ & -y_{ij} \le x_j - z_{ij} \le y_{ij}, \quad \forall i, j \in J_i, \\ & l \le x \le u, \\ & y \ge 0, \end{aligned} \tag{2.4}$$

where Ω_i, which is dependent upon the user's risk preference, is a user defined parameter and adjusts the trade-off between robustness and optimality. The authors have shown that the probability that the i constraint is violated at most $\exp(-\Omega_i^2)/2$. This robust model is less conservative than Model (2.3) as every feasible solution of the latter problem is a feasible solution to the former problem. However, Model (2.4) is a second order cone problem, which is more demanding computationally than the linear Model (2.3).

To overcome this difficulty, [4] presented a new robust formulation. One attractive aspect of this method is that the new formulation is a linear programming problem. The authors introduced a budget parameter $\Gamma_i \in [0, |J_i|]$ for each row $i = 1, \ldots, m$ of the matrix A, which is a user defined parameter that adjusts the robustness of the model and interpreted as the maximum number of uncertain parameters allowed to take their worst case value. The role of Γ_i is in the following way:

If $\Gamma_i = 0$, each $a_{ij}, j \in J_i$, is forced to take its mean value $\bar{a}_{ij}$. If $\Gamma_i = |J_i|$, each $a_{ij}, j \in J_i$, can take values from its range $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$. If $0 < \Gamma_i < |J_i|$, $\lfloor \Gamma_i \rfloor$ elements among $a_{ij}, \forall j \in J_i$, can take values from their respective ranges; furthermore, if Γ_i is not an integer, one other random element, says a_{it_i}, can take values from its reduced range $[\bar{a}_{it_i} - (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i}, \bar{a}_{it_i} + (\Gamma_i - \lfloor \Gamma_i \rfloor)\hat{a}_{it_i}]$; the remaining $|J_i| - \lceil \Gamma_i \rceil$ random elements are forced to take their respectively mean values. The role of the parameter Γ_i is to adjust the robustness against the level of conservatism of the solution. As surmised in [4], nature could be restricted in its behavior in that only a subset of the random elements actually deviate from their respective mean values, in order to adversely affect the solution.

Given Γ_i for all i, [4] sought a solution that maximizes the objective value under the restriction that it must remain feasible to (2.1) as long as up to Γ_i elements out of $a_{ij}, \forall j \in J_i$, are allowed to change. Let S_i be a subset of J_i, such that $|S_i| = \lfloor \Gamma_i \rfloor$, and let $t_i \in J_i \backslash S_i$. To obtain such a solution, [4] constructed the following robust formulation:

$$
\begin{aligned}
\max \quad & c^T x, \\
\text{s.t.} \quad & \sum_j \bar{a}_{ij} x_j + \max_{S_i \cup t_i \subseteq J_i} \left\{ \sum_{j \in S_i} \hat{a}_{ij} y_j + (\Gamma_i - \lfloor \Gamma_i \rfloor) \hat{a}_{it_i} y_{t_i} \right\} \leq b_i, \quad \forall i, \\
& -y_j \leq x_j \leq y_j, \quad \forall j, \\
& l \leq x \leq u, \\
& y \geq 0.
\end{aligned} \tag{2.5}
$$

[4] proved that the probability that the i constraint violation is bounded above by $\exp(-\Gamma_i^2)/2|J_i|$. Thus, as Γ_i increases, more protection is given and the solution is more robust. The authors showed that the above nonlinear robust formulation can be recast as an equivalent linear programming formulation:

$$\begin{aligned}
\max \quad & c^T x, \\
\text{s.t.} \quad & \sum_j \bar{a}_{ij} x_j + \Gamma_i z_i + \sum_{j \in J_i} p_{ij} \leq b_i, \quad \forall i, \\
& z_i + p_{ij} \geq \hat{a}_{ij} y_j, \quad \forall i, \forall j \in J_i, \\
& p_{ij} \geq 0, \quad \forall i, \forall j \in J_i, \\
& z_i \geq 0, \quad \forall i, \\
& -y_j \leq x_j \leq y_j, \quad \forall j, \\
& l \leq x \leq u, \\
& y \geq 0.
\end{aligned} \tag{2.6}$$

3 Robust DEA Model

Assume that we deal with a set of n DMUs converting m inputs into s outputs, with input-output vectors (x_j, y_j); $j = 1, \cdots, n$, in which $x_j = (x_{1j}, \cdots, x_{mj})^T$ and $y_j = (y_{1j}, \cdots, y_{sj})^T$. Define $X = [x_1, x_2, \ldots, x_n]$ and $Y = [y_1, y_2, \cdots, y_n]$ as $m \times n$ and $s \times n$ matrices of inputs and outputs, respectively. The original fractional DEA model which is an input-oriented CCR model is presented as :

$$\begin{aligned}
\max \quad & \frac{v^T y_{j_0}}{u^T x_{j_0}}, \\
\text{s.t.} \quad & \frac{v^T y_j}{u^T x_j} \leq 1, \; j = 1, \cdots, n, \\
& v, u \geq 0,
\end{aligned} \tag{3.1}$$

which evaluates the relative efficiencies of n DMUs by maximizing the ratio of weighted summation of outputs to weighted summation of inputs. u and v are weight vectors associated with inputs and outputs, respectively. In addition, x_{j_0} and y_{j_0} are the input and output for the DMU under evaluation. This model is a nonlinear programming problem, and it is equivalent to the following linear programming problem which is more computational convenient:

$$\begin{aligned}
\max \quad & \mu^T y_{j_0}, \\
\text{s.t.} \quad & \mu^T y_j - \nu^T x_j \leq 0, \; j = 1, \ldots, n, \\
& \nu^T x_{j_0} = 1, \\
& \mu, \nu \geq 0.
\end{aligned} \tag{3.2}$$

We assume that only the output data cannot be exactly obtained due to the existence of uncertainty, so as to avoid the appearance of uncertainty in the input related equality

constraint. They are only known to lie within the upper and lower bounds represented by the range $[y_{rj}^L, y_{rj}^U]$, where $y_{rj}^L > 0$. Assume the mean of the random variable y_{rj} is $\bar{y}_{rj}$. In order to avoid the appearance of uncertainty in objective function, we express the objective function as $\max z$, and add the constraint $z - \mu^T y_{j_0} \leq 0$ into the constraints. In order to generate an uncertainty-immune solution, we adopt the robust optimization technique proposed by [4]. Let J_j be the set of coefficients in column j of matrix Y that are subject to uncertainty. Choose $\Gamma_j \in [0, |J_j|]$. Let

$$\begin{aligned}\Re(\Gamma_j) &= \{y_j = (y_{1j}, \cdots, y_{sj}) | y_{ij} \in [\bar{y}_{ij} - \beta_{ij}(\bar{y}_{ij} - y_{ij}^L), \bar{y}_{ij} + \beta_{ij}(y_{ij}^U - \bar{y}_{ij})], \\ &\quad \forall i; 0 \leq \beta_{ij} \leq 1, \forall i; \sum_{i \in J_j} \beta_{ij} \leq \Gamma_j, \text{ at most one } \beta_{ij} \text{is fractional}\}.\end{aligned}$$

The robust DEA model with output uncertainty is expressed as following:

$$\begin{aligned}\max \quad & z, \\ \text{s.t.} \quad & \min_{y_{j_0} \in \Re(\Gamma_{j_0})} \mu^T y_{j_0} \geq z, \\ & \max_{y_j \in \Re(\Gamma_j)} \mu^T y_j - \nu^T x_j \leq 0,\ j = 1, \ldots, n, \\ & \nu^T x_{j_0} = 1, \\ & \mu, \nu \geq 0.\end{aligned} \tag{3.3}$$

Let S_j be a subset of J_j such that $|S_j| = \lfloor \Gamma_j \rfloor$, and let $t \in J_j \backslash S_j$. It can be deduced that

$$\min_{y_{j_0} \in \Re(\Gamma_{j_0})} \mu^T y_{j_0} = \mu^T \bar{y}_{j_0} - \max_{S_{j_0} \cup \{t\} \subseteq J_{j_0}} \{\sum_{i \in S_{j_0}} \mu_i y_{ij_0}^L + (\Gamma_{j_0} - \lfloor \Gamma_{j_0} \rfloor) \mu_t y_{tj_0}^L\},$$

and

$$\max_{y_j \in \Re(\Gamma_j)} \mu^T y_j = \mu^T \bar{y}_j + \max_{S_j \cup \{t\} \subseteq J_j} \{\sum_{i \in S_j} \mu_i y_{ij}^U + (\Gamma_j - \lfloor \Gamma_j \rfloor) \mu_t y_{tj}^U\}.$$

Let

$$\psi_j(y, \Gamma_j) = \max_{S_j \cup \{t\} \subseteq J_j} \{\sum_{i \in S_j} \mu_i y_{ij}^U + (\Gamma_j - \lfloor \Gamma_j \rfloor) \mu_t y_{tj}^U\},$$

the function $\psi_j(y, \Gamma_j)$ equals to the objective function of the following linear programming problem:

$$\begin{aligned}\max \quad & \sum_{i \in J_j} \mu_i y_{ij}^U \omega_{ij}, \\ \text{s.t.} \quad & \sum_{i \in J_j} \omega_{ij} \leq \Gamma_j, \\ & 0 \leq \omega_{ij} \leq 1,\ \forall i \in J_j.\end{aligned} \tag{3.4}$$

Its dual problem is

$$\begin{aligned} \min \quad & \Gamma_j z_j + \sum_{i \in J_j} p_{ij}, \\ \text{s.t.} \quad & z_j + p_{ij} \geq \mu_i y_{ij}^U, \forall i \in J_j, \\ & p_{ij} \geq 0, \forall i \in J_j, \\ & z_j \geq 0. \end{aligned} \tag{3.5}$$

By strong duality of linear programming, since problem (3.4) is feasible and bounded for all $\Gamma_j \in [0, |J_j|]$, then the dual problem (3.5) is also feasible and bounded and their objective values coincide. So we have that $\psi_j(y, \Gamma_j)$ is equal to the objective function value of problem (3.5). Substituting to problem (3.3) we obtain that the robust DEA model is equivalent to the following linear optimization problem:

$$\begin{aligned} \max \quad & z, \\ \text{s.t.} \quad & \mu^T \bar{y}_{j_0} - (\Gamma_{j_0} z_{j_0} + \sum_{i \in J_{j_0}} p_{ij_0}) \geq z, \\ & \mu^T \bar{y}_j + \Gamma_j z_j + \sum_{i \in J_j} p_{ij} - \nu^T x_j \leq 0, \\ & \nu^T x_{j_0} = 1, \\ & z_j + p_{ij} \geq \mu_i y_{ij}^U, \quad \forall i \in J_j, \forall j, \\ & z_{j_0} + p_{ij_0} \geq \mu_i y_{ij_0}^L, \quad \forall i \in J_{j_0}, \\ & p_{ij} \geq 0, \quad \forall i \in J_j, \forall j, \\ & z_j \geq 0, \quad \forall j, \\ & \mu \geq 0, \nu \geq 0. \end{aligned} \tag{3.6}$$

This linear programming model can be solved by the simplex method or interior point method.

4 Numerical Example

In this section, a simple numerical example computed by the Matlab software is considered to clearly illustrate the proposed robust DEA approach. Now we look at a DEA example which includes five DMUs, each DMU has three inputs and two outputs. Assume that the input data x_{ij} is deterministic and the output data y_{ij} is uncertain, which lies within the range $[\bar{y}_{ij} - d_{ij}^L, \bar{y}_{ij} + d_{ij}^U]$, where $\bar{y}_{ij}$ is the nominal output data. Table 1 shows the input data and the nominal output data.

Table 1 The input and nominal output data for each DMU

DMU	Input			Output	
	x_1	x_2	x_3	y_1	y_2
DMU1	9	5	8	4	18
DMU2	5	9	8	10	10
DMU3	8	10	11	5	20
DMU4	8	7	8	9	9
DMU5	8	9	10	2	25

For simplicity, we assume that d_{ij}^L and d_{ij}^U are 0.5 and 1 respectively, J_j is equal to 2, so Γ_j can vary between 0 and 2. Denote $\Gamma = (\Gamma_1, \Gamma_2, \cdots, \Gamma_5)$, we increase Γ from 0 to 1.75, the efficiency scores of the DMUs are summarized in Figure 1. It is evident that the efficiency scores of the DMUs are decreasing as Γ increasing, we can call this the price of robustness. As shown, compared with the others, DMU 2 is the most efficient decision making unit.

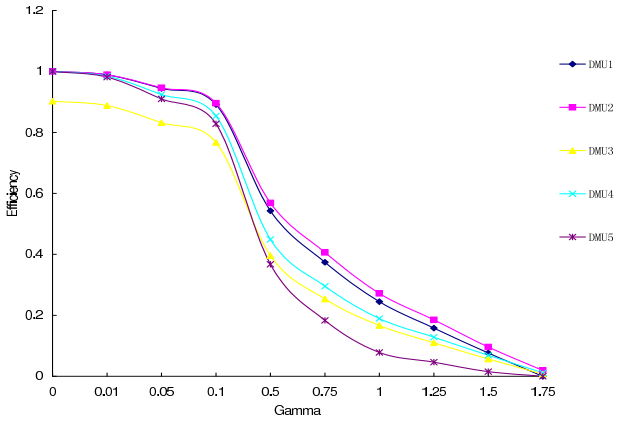

Fig. 1 Efficiency scores of the DMUs

5 Conclusion

In this paper, we analyze the impact of output data uncertainty on the result of DEA and propose a robust model based on the newly developed robust optimization approaches. The conventional robust DEA model assumed the symmetric data uncertainty: the ranges of the uncertain elements are symmetrically bounded around their means. We relax the assumption of symmetric data uncertainty to make our robust DEA model more suitable for applications. We construct the robust DEA model for asymmetric data uncertainty and obtain an equivalent linear programming formulation. We implement the proposed model in a numerical example and the results indicate that considering the output data uncertainties

when applying DEA approach is very important, and using robust DEA approach could be more reliable for efficiency evaluation and ranking in multiple criteria decision making problems.

References

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

[2] Cook W D, Seiford L M. Data envelopment analysis(DEA) thirty years on[J]. Eur. J. Oper. Res., 2009, 192(1): 1–17.

[3] Ben-Tal A, Nemirovski A. Robust solutions of linear programming problems contaminated with uncertain data[J]. Math. Program., 2000, 88(3): 411–421.

[4] Bertsimas D, Sim M. The price of robustness[J]. Oper. Res., 2004, 52(1): 35–53.

[5] Sadjadi S J, Omrani H. Data envelopment analysis with uncertain data: an application for Iranian electricity distribution companies [J]. Energy Policy, 2008, 36(11): 4247–4254.

[6] Wang K, Wei F J. Robust data envelopment analysis based MCDM with the consideration of uncertain data[J]. J. Syst. Eng. Electron., 2010, 21(6): 981–989.

[7] Soyster A L. Convex programming with set-inclusive constraints and applications to inexact linear programming[J]. Oper. Res., 1973, 21: 1154–1157.

[8] Ben-Tal A, Nemirovski A. Robust solutions of uncertain linear programs[J]. Oper. Res. Lett., 1999, 25(1): 1–13.

数据不确定的鲁棒数据包络分析方法

于洪霞[1] , 金 丽[2]

(1.上海电力学院数理学院, 上海 201300)

(2.浙江海洋学院数理与信息学院, 浙江 舟山 316004)

摘 要: 本文研究了输出参数不确定的数据包络分析(DEA)模型. 利用鲁棒优化方法, 建立了一个鲁棒DEA模型, 通过数值试验表明了该模型在有效性评价和排序方面的可靠性. 本文的模型可处理分布未知的不确定数据, 与已有的只考虑对称分布不确定参数的模型相比, 适用范围更广.

关键词: 数据包络分析; 不确定性; 鲁棒优化

MR(2010)**主题分类号**: 90C90; 90C05 **中图分类号**: O221.1