Outliers In Data Envelopment Analysis

Shaik Khaleel Ahamed *khaleelska@gmail.com*
Research Scholar, C.S.E.Dept
S.V.U.College of Engineering
S.V. University
Tirupati, A.P, 517501, India

Prof.MM Naidu *mmnaidu@yahoo.com*
Professor, C.S.E. Dept
S.V.U.College of Engineering
S.V. University
Tirupati, A.P, 517501, India

Prof.C.Subba Rami Reddy *csruma@yahoo.com*
Professor, Statistics.Dept
S.V. University
Tirupati, A.P, 517501, India

Abstract

Data Envelopment Analysis is a linear programming technique that assigns efficiency scores to firms engaged in producing similar outputs employing similar inputs. Extremely efficient firms are potential Outliers. The method developed detects Outliers, implementing Stochastic Threshold Value, with computational ease. It is useful in data filtering in BIG DATA problems.

Keywords: Constant Return to Scale, Data Envelopment Analysis, Super Efficiency, Threshold Value.

1. INTRODUCTION

An ‘outlier’ is an observation that is radically dissimilar with majority of observations. It falls outside a cloud of normal observations. The presence of an outlier may be due to reporting errors. Such observations shall be corrected or removed for a valid empirical analysis and consequent conclusions. If an outlier arrives from the same probability distribution as others, they do occur with small probability. Such observations shall be carefully examined since they carry special information that cannot be retrieved from the normal observations. Outliers do not possess any item in a neighborhood of a specified radius. Detection of outliers is constituted by two sub problems.

(i) Define inconsistency in a data set and
(ii) To provide an efficient method to identify the inconsistent observations (outliers).

2. DATA ENVELOPMENT ANALYSIS

Data Envelopment Analysis is a linear programming technique that measures efficiency of decision making units. In efficiency evaluation production plans are projected onto the envelopment frontier determined by the most efficient observations that are potential outliers. Outliers elevate the frontier leading to the under estimation of efficiency scores of inefficient decision making units. Charnes, Cooper and Rhodes (1978) proposed a technology set that is based on the axioms of inclusion, free disposability and minimum extrapolation, whose boundary serves as envelopment frontier that admits constant returns to scale. The efficiency scores of interior production units are under estimated in the presence of outliers in the CCR (1978) model. Banker, Charnes and Cooper (BCC, 1984) extended the CCR model, whose production possibility set is based on the axioms of inclusion, convexity, free disposability and minimum extrapolation. The extremely efficient decision making units are potential Outliers.

3. DEA – Outliers

a) Timmer (1971) was the first one to recognize high sensitivity of DEA scores when outliers are present, in linear programming problems. By suitably finding the threshold value, a specified percent of firms were removed from the reference set to arrive at output elasticities with respect to inputs, in the frame work of Cobb-Douglass production function (1928), with acceptable magnitudes. The deleted input and output plans are viewed as Outliers**. The percentage of firms removed from the data is subjective.**

b) In DEA all efficient decision making units are flagged as potential outliers. The efficiency score of efficient firms is 100%. Andersen and Petersen (1993) suitably tailored the DEA constraints to assess super efficiency scores of efficient firms. Such production unit with larger efficiency score (input approach) is ranked better. The input super efficiency score is larger than or equal to unity, for such production plans. In their approach firm's input and output vector, whose efficiency is under evaluation, is removed from the reference set, the assessed DMU being efficient. Consequently, the input vector falls below the input efficient frontier and the deletion pushes the frontier upwards, toward inefficient units all producing a given level or more of an output. Deletion of an efficient production plan from the reference set leads to the contraction of input sets. Such input efficient decision making unit whose deletion from the reference set resulted in maximum contraction of input set is the most influential observation, possibly an outlier (refer to the figure). The property of frontier displacement refers to efficient decision making units. If the input and output combination of efficient firm is removed from the reference set, for the same firm its production plan is projected on to the constrained frontier. If input orientation is pursued this score emerges to be one or more than one. Suppose the input efficiency score is 1.5, then this score is interpreted as, that this firm will continue to be efficient in the presence of input expansion up to a factor 1.5. This approach can be extended in a straight forward manner to output and graph orientation. **The super efficiency measurement above gives a single measurement of irregular polyhedron. The threshold value to identify outliers is due to subjective choice.**

c) Wilson (1995) identified outliers following leave-one-out approach, and the search was in relation to efficient frontier, under exclusively input perspective and output perspective. **Wilson's method requires more computational labour while his threshold value is subjective.**

d) Simar (2003) suggests that a production plan shall be treated as an outlier if it is sufficiently influential under both orientations (input and output). **His threshold values to identify outliers are subjective.**

e) Tran et.al (2008) proposed a new method for detecting outliers in Data Envelopment Analysis. They consider the CCR-DEA formulation and the observed plans which determined the CCR frontier as potential outliers. Their approach depended on the intensity parameters of efficient firms arrived at construction of the DEA hull. With reference to CCR-DEA hull the intensity parameters are non-negative. If a firm is inefficient, its intensity parameter is assigned with a zero value by every firm, including itself. An efficient firm evaluated relatively efficient by itself may participate in the construction of DEA frontier for the evaluation of inefficient decision making units, there by possess positive intensity parameters. An efficient firm that appears the most with positive intensity parameter values while inefficient firms are evaluated may be viewed as an influential observation. For identification of outlier not only the count of positive intensity parameter values is important as metric but their sum can also be used as another metric. Stosic and Sampario de souza (2003) proposed a method which is based on a combination of a boot strap and resampling schemes for automatic detection of outliers, which takes into consideration the concept of leverage. The leverage metric measures the effect produced on the efficiency scores of all others DMUs, when a particular firm is removed from the data set. Outliers are expected to display leverage much above the mean leverage and hence should be selected with lower probability than the other DMUs when resampling is performed.

f) Sampario de Souza et.al (2005) defined the leverage of j^{th} DMU as,

$$l_j = \sqrt{\frac{\sum_{k=1,k+j}^{n} \left(\theta^*_{kj} - \theta_k\right)^2}{n-1}}$$

where θ^*_{kj} is the efficiency score of k^{th} DMU based on the data set from which j^{th} DMU's production plan is removed, and θ_k is efficiency score of k^{th} DMU. **Based on unaltered data set, one can compute mean leverage, in boot strap samples choice of threshold value being subjective.**

g) Johnson et.al (2008) believed outliers are found not only among extremely efficient but also inefficient observations. The leverage of an input and output observation to displace the frontier is chosen as a metric to identify an outlier both in efficiency and inefficiency perspectives. The leverage estimate is provided by super efficiency and super inefficiency score. **For this purpose the efficient and inefficient frontiers are used, which bind the production possibility set from above and below, the choice of threshold value is subjective.**

h) Chen and Johnson (2010) formulated an alternative to the above approach. They consider Hull that satisfies the axioms of inclusion and convexity. The axiom of free disposability is withdrawn, on which the convex Hull is built. The methodology developed to identify outliers is similar to the super efficiency evaluation proposed by Andersen and Petersen (1993). The leverage of a DMU to contract the production possibility set while its input vector and output vector are removed from the reference technology determines if the DMU under evaluation is outlier or not. Removal of free disposability axiom, removes the weak efficient subset of the DEA production possibility set from the reference technology, overall boundary shift attributed to an efficient decision making unit serves as a metric to classify it as an outlier or not**. The threshold value is subjective and the method involves greater computational labour.**

4. NEW METHOD- ITS MERITS OVER OTHER METHODS

The proposed study is an attempt to identify outliers in a scenario that there are n production units combining m similar inputs to produce s similar outputs. The production units may be profitable or non-profitable organizations. The input and output vectors of the production units spin a production possibility set under the axioms of inclusion, free disposability, closure under ray expansion and contraction and minimum extrapolation. The production units can be decomposed into four disjoint sets constituted by, (i) extremely efficient, (ii) efficient, (iii) weakely efficient and (iv) inefficient. The surface of the pp set is spun by the extremely efficient ones. All the extremely efficient firms constitute the reference technology of production process. If the input and output vectors of an extremely efficient firm is deleted from the reference technology then the production possibility set experiences contraction. The new pp set is a subset of the original pp set. An inefficient firm's input and output vectors deletion leaves the pp set intact. The potential outliers are the extremely efficient firms. An important direction in the attempt to identify outliers is suggested by Andersen and Petersen (1993) through their super efficiency measurement problem. Their approach reveals such extremely efficient firm with the largest (smallest) super efficiency score under input (output) orientation is certainly an outlier. In this method for identification of outliers, a threshold value needs to be specified which is subjective. Further, super efficiency score provides one measurement of an irregular polyhedran that accounts for contracted region. When an extremely efficient firm's input and output vectors are deleted from the reference technology, for some inefficient firms, their efficiency scores will increase and for the remaining inefficient firms, their efficiency scores would be intact. The increments of efficiency-scores of inefficient firms provide additional measurements of contracted region embedded in an irregular polyhedron. These additional measurements combined with the difference between the super efficiency score and unity provides a means to obtain statistically based threshold value that facilitates outliers identification. The various methods of outlier identification outlined in the review suffer from subjective threshold value and heavy computational labour. **The merits of the new method are that the threshold value is statistically determined, requires least computational labour. This method is of immence use in data filtering in problems that constitute inputs and outputs with a monotonic relationship between inputs and outputs, particularly useful in BIG DATA problems.**

4.1 Data Envelopment Analysis-Constant Return To Scale-Outliers

Charnes, Cooper and Rhodes (1978) proposed a fractional programming problem to measure technical efficiency of decision making units. Applying Charnes and Cooper transformation, this problem can be transformed into a linear programming problem. Under input perspective the optimal solution not only assigns a technical efficiency score to each decision making unit, but provides such scores to its peer DMUs that are based upon the input and output weights of the decision making unit for which the CCR-DEA problem is solved.

Let $x_{ij,} i \in I; y_{rj,} r \in S$ be the inputs and outputs of the decision making unit $j \in J$. For j=0, the following CCR problem is solved:

$$\delta_0^1 = \max \sum_{r=1}^{s} v_r y_{r0}$$

$$\text{s.t} \sum_{i=1}^{m} u_i x_{i0} = 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

$$\sum_{r=1}^{s} v_r y_{rj} - \sum_{i=1}^{m} u_i x_{ij} \leq 0, \forall j \in J$$

$$v_r \geq 0, r \in S; u_i \geq 0, i \in I$$

For efficient decision making units δ_0^1 =1 and the corresponding slack is zero for $j = 0 \in J$.

The potential decision making units are the efficient ones. Solving the above problem for each decision making unit, efficient firms can be identified. These firms are potential super efficient. To assess super efficiency of extremely efficient decision making units. Andersen and Petersen (1993) formulated an input oriented envelopment problem.

$$\delta_0^2 = \min \lambda$$

$$\text{s.t} \sum_{\substack{j=1 \\ j \neq 0}}^{n} \lambda_j x_{ij} \leq \lambda x_{i0}, i \in I \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

$$\sum_{\substack{j=1 \\ j \neq 0}}^{n} \lambda_j y_{rj} \geq y_{r0}, r \in S$$

$$\lambda_j \geq 0, \forall j \in J - \{0\}$$

i) The super efficiency problem is solved for the extremely efficient decision making units.
ii) Super efficiency score measures the ability of an extremely efficient decision making unit to remain efficient in the event of further radial augmentation of inputs upto some degree.

iii) Under constant return to scale frame work the super efficiency problem is always feasible if input and output values are positive.

iv) Super efficiency score reveals the ability of the firm to contract the production possibility set.

v) The dual of the above envelopment problem is,

$$\delta_0^2 = \max \sum_{r=1}^{s} v_r y_{r0}$$

$$\text{s.t } \sum_{i=1}^{m} u_i x_{i0} = 1 \quad \text{……………………. (3)}$$

$$\sum_{r=1}^{s} v_r y_{rj} - \sum_{i=1}^{m} u_i x_{ij} \le 0,\ j \in\ j - \{0\}$$

$$v_r \ge 0, r \in S$$

$$u_i \ge 0, i \in I$$

The optimal solution of (1) is a feasible solution of (2). Therefore,

$$\delta_0^2 \ge \delta_0^1$$

For extremely efficient firm, $\delta_0^1 = 1 \Rightarrow \delta_0^2 \ge 1$.

Problem (1) and (3) can be equivalently expressed as,

$$\delta_0^1 = \max \frac{\sum_{r=1}^{s} v_r y_{ro}}{\sum_{i=1}^{m} u_i x_{io}}$$

$$\text{s.t } \frac{\sum_{r=1}^{s} v_r y_{rj}}{\sum_{i=1}^{m} u_i x_{ij}} \le 1, j \in J \quad \text{……………………….. (4)}$$

$$v_r \ge 0, r \in S; u_i \ge 0, i \in I$$

$$\delta_0^2 = \max \frac{\sum_{r=1}^{s} v_r y_{ro}}{\sum_{i=1}^{m} u_i x_{io}}$$

$$\text{s.t } \frac{\sum_{r=1}^{s} v_r y_{rj}}{\sum_{i=1}^{m} u_i x_{ij}} \le 1, j \in J - \{0\} \quad \text{……………………. (5)}$$

$$v_r \ge 0, r \in S; u_i \ge 0, i \in I$$

Applying Charnes and Cooper transformation problem (4) and (5) can be reduced to (1) and (3) respectively.

Every feasible solution of program (4) is a feasible solution of (5). If $\left(\overline{v,u}\right)$ and $\left(\overline{\overline{v}},\overline{\overline{u}}\right)$ are optimal solutions of (4) and (5) respectively, then we have,

$$\frac{\sum_{r=1}^{s}\overline{v}_r y_{rj}}{\sum_{i=1}^{m}\overline{u}_i x_{ij}} \le \frac{\sum_{r=1}^{s}\overline{\overline{v}}_r y_{rj}}{\sum_{i=1}^{m}\overline{\overline{u}}_i x_{ij}} \le 1,\ j\in J$$

$$\Rightarrow \frac{OD^{'}}{OD} \le \frac{OD^{"}}{OD}$$

$$\frac{OE^{'}}{OE} \le \frac{OE^{"}}{OE}$$

$$\frac{OF^{'}}{OF} \le \frac{OF^{"}}{OF}$$

For j=0, $\frac{\sum_{r=1}^{s}\overline{v}_r y_{ro}}{\sum_{i=1}^{m}\overline{u}_i x_{io}} \le \frac{\sum_{r=1}^{s}\overline{\overline{v}}_r y_{ro}}{\sum_{i=1}^{m}\overline{\overline{u}}_i x_{io}}$

since this firm is efficient, $\frac{\sum_{r=1}^{s}\overline{v}_r y_{ro}}{\sum_{i=1}^{m}\overline{u}_i x_{io}} = 1$

$$\frac{\sum_{r=1}^{s}\overline{\overline{v}}_r y_{ro}}{\sum_{i=1}^{m}\overline{\overline{u}}_i x_{io}} \ge 1$$

$$\frac{OB^{'}}{OB} \ge 1$$

$$\frac{OB^{'}}{OB} - d_B = 1$$

$$d_B = \frac{OB^{'}}{OB} - 1$$

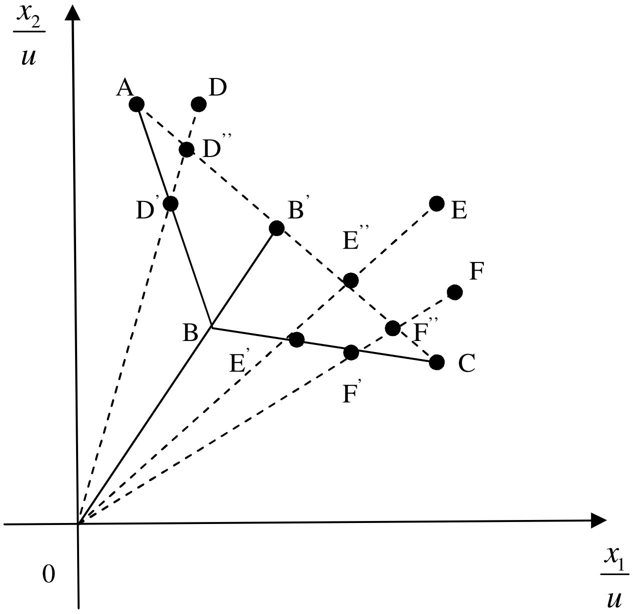

FIGURE 1: Unit Output Isoquant.

In the figure above first and second input requirements to produce unit output are measured along horizontal and vertical axes respectively. The input isoquant is determined by the extremely efficient firms A,B and C. the firms D,E and F are inefficient for which the firm B is an efficient peer, solving problem(1) for firm B, its standard efficiency score and cross efficiency scores for the remaining decision making units can be obtained. The cross efficiency scores are as follows: $\frac{OD'}{OD}, \frac{OE'}{OE}, \frac{OF'}{OF}$. Such efficiency scores of a firm evaluated with other firm's efficiency scores are called cross efficiency scores.

Solving the super efficiency problem (3), super efficiency scores for firm B and cross efficiency scores for other firms can be obtained. The cross efficiency scores of other firms are,

$$\frac{OD''}{OD}, \frac{OE''}{OE}, \frac{OF''}{OF}$$

$$\frac{OD''}{OD} \geq \frac{OD'}{OD}$$

$$\frac{OE''}{OE} \geq \frac{OE'}{OE}$$

$$\frac{OF''}{OF} \geq \frac{OF'}{OF}$$

The area of the triangle ABC measures the contraction of the production possibility set. The super efficiency score of B, provides one measurement of contracted production possibility set,

$d_B = \frac{OB'}{OB} - 1$ that lies between zero and one.

d_B gives a measurement of production possibility set contraction.

Define

$$d_D = \frac{OD^{''}}{OD} - \frac{OD^{'}}{OD}$$

$$d_E = \frac{OE^{''}}{OE} - \frac{OE^{'}}{OE}$$

$$d_F = \frac{OF^{''}}{OF} - \frac{OF^{'}}{OF}$$

$d_{D,}d_E$ and d_F are also measurements of contraction of the production possibility set. We take average of all these measurements to arrive at a more meaning full measure of contraction.

$\overline{d_B} = \dfrac{d_B + d_D + d_E + d_F}{\eta_B}$, where $\eta_B = 4$

The above arithmetic mean gives rise to a Student t-test, in which $\overline{d}$ is tested against zero, if sample size is small

$t_B = \dfrac{\overline{d_B}}{s/\sqrt{\eta_B}}$ follows Student's t-distribution with $\eta_B - 1$ degrees of freedom.

If $\overline{d_B} \geq t_\alpha \dfrac{s}{\sqrt{\eta_B}}$, then firm B is an outlier, where α is the level of significance.

If there are other decision making units that are inefficient and for which firm B is not an efficient peer, for such firms problems (1) and (3) assign the same efficiency scores, so that their deviations vanish.

(i) For outlier determination a threshold value is needed, whose choice often subjective. This method provides a threshold value $t_\alpha \dfrac{s}{\sqrt{\eta_B}}$ that is statistically determined which depends upon the level of significance.
(ii) Further, this method need not choose every extremely efficient decision making unit as an outlier.
(iii) It is a common practice to identify large super efficient firms as outliers, 'how large' is a subjective matter.
(iv) For the identification of an outlier this method uses not only the super efficiency scores, but also the potential improvements of efficiency of inefficient decision making units.

5. FUTURE RESEARCH DIRECTION

Economic data often are subjected to returns to scale. Returns to scale may be constant, increasing or decreasing. The present study assumes constant returns to scale. The super efficiency problems are always feasible, if input and output values are positive and returns to scale are constant. However, if return to scale are either increasing or decreasing it is likely that for some extremely efficient firms their super efficiency problems are infeasible. A natural extension of the present study is identification of outliers, suitably fine tuning the super efficiency problems to be free from infeasibility, in the presence of non-constant returns to scale.

6. REFERENCES

[1] Andersen, P. and N. C. Petersen. (1993). "A Procedure for Ranking Efficient Units in Data Envelopment Analysis." *Management Science*, 39:1261-1264.

[2] Charnes, W.W. Cooper, Z.M. Huang and D.B. Sun, Polyhedral cone-ratio DEA models with an illustrative application to large commercial bank, Journal of Economics 46 (1990) 73-91.

[3] Banker, Charnes and Cooper (1984).”Estimating Most Productive Scale Size Using Data Envelopment Analysis.” *European Journal Of Operations Research* 35-44

[4] Charnes, A., Cooper W.W., and Rhodes, E., (1978), “Measuring the Efficiency of Decision-Making Units”, European Journal of Operations Research, 2, 429-444.

[5] Chen and Johnson (2010) ; “A Unified model for detecting Outliers in DEA, Computers and Operations Research, Vol. 37. 417-425.

[6] Daraio, C. and L. Simar (2003), Introducing environmental variables in nonparametric frontier models: a probabilistic approach, Discussion paper 0313, Institute de Statistique, Universities Catholique de Louvain, Belgium.

[7] Johnson, A.L., Chen W.C., McGinnis, L.F., (2008).,” Internet-based benchmarking for warehouse operations”. Working Paper, 2008.

[8] J.R. Doyle and R.H. Green, Efficiency and cross-efficiency in DEA: derivations, meanings and uses, Journal of Operational Research Society 45 (1994) 567-578.

[9] J.H. Dula and B.L. Hickman, Effects of excluding the column being scored from the DEA envelopment LP technology matrix, Journal of Operational Research Society 48 (1997) 1001-1012.

[10] J. Zhu, Robustness of the efficient DMUs in data envelopment analysis, European Journal of Operational Research 90 (1996) 451-460.

[11] J. Zhu, Super-efficiency and DEA sensitivity analysis, European Journal of Operational Research 129 (2001) 443-455.

[12] M. Halme and P. Korhonen, Restrciting weights in value efficiency analysis, European Journal of Operational Research 126 (2000) 175-188.

[13] P.C. Pendharkar, “A Data Envelopment Analysis-Based Approach for Data Preprocessing,” IEEE Transactions on Knowledge & Data Engineering, Vol. 17, No. 10, 2005, pp. 1379-1388.

[14] R.G. Dyson and E. Thanassoulis, Reducing weight flexibility in data envelopment analysis, Journal of Operational Research Society 39 (1988) 563-576.

[15] R. Green, J.R. Doyle and W.D. Cook, preference voting and project ranking using DEA and cross-evaluation, European Journal of the Operational Research 90 (1996) 461-472.

[16] Stosic, B. and Sampaio de Sousa, M.C. (2003) “Jackstrapping Dea Scores For Robust Efficiency Measurement.” Series Texto para Discussão N° 291, Universidade de Brasília.

[17] S. Talluri and J. Sarkis, Extensions in efficiency measurement of alternate machine component grouping solutions via data envelopment analysis, IEEE Transactions on Engineering Management 44 (1997) 27-31.

[18] Tran,N.M., Sheverly,G., and Preckel,P., (2008) “ A New Method for detecting Outliers in DEA”, Applied Economic Letters, 1-4.

[19] T. R. Anderson, A. Uslu, and K. B. Hollingsworth, "Revisiting extensions in efficiency measurement of alternate machine component grouping solutions via data envelopment analysis," Working paper 1998.

[20] Timmer, C. Peter,(1971)," Using a probabilistic frontier production function to measure technical efficiency", Journal of Political Economy 79, 776-794.

[21] Wilson, P. W. (1995) "Protecting Influential Observations in Data Envelopment Analysis." Journal of Productivity Analysis, 4:27–45.