日本オペレーションズ・リサーチ学会
2022年 春季研究発表会

# Streamlined DEA computation in the big data context

申請中　Osaka University　*ZHUANG Qianwei
01604524　Osaka University　MORITA Hiroshi

## 1. Background

Data envelopment analysis (DEA) was developed to assess the performance of a set of homogeneous decision-making units (DMUs), with multiple inputs and multiple outputs, by linear programming methods.

In traditional DEA for a set of $n$ DMUs, DEA model is solved $n$ times, one for each DMU. As the cardinality $n$ increases, the running-time sharply increases. Several studies provided their framework to reduce the computation time. The framework PH (Pre-score Hull) by Khezrimotlagh, et al. (2019) is currently the most powerful one.

## 2. Problem statement

The goal of this study is to develop a new method to decrease the computation time required to solve large-scale DEA problems. Our major contribution is we built a hull to find all the efficient DMUs in a faster way than PH method.

In this study, we focus on the envelopment form of radial input-oriented variable return-to-scale (VRS) DEA model. The envelopment form contains $n+1$ decision variables, $m+s+1$ constraints and $n$ non-negative restrictions, where $n$ is the number of DMUs and $m$ and $s$ are the numbers of inputs and outputs, respectively. The model is as follows, where $DMU_l$ is under evaluation:

$$\min \varphi_l - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right)$$

$$s.t.$$

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = \varphi_l x_{il}, \qquad i = 1,2,\dots,m,$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{rl}, \qquad r = 1,2,\dots,s,$$

$$\sum_{j=1}^{n} \lambda_j = 1,$$

$$\lambda_j, s_i^-, s_r^+ \geq 0, \qquad \forall i, r, j.$$

**Definition1.** $DMU_l$ is called an efficient DMU if $\varphi_l = 1$.Otherwise if $\varphi_l < 1$ then it is inefficient. The set of all the efficient DMU is denoted by $\mathfrak{F}$.

**Definition2.** The piece-wise plane (or hyperplane) formulated by efficient DMUs is called production frontier.

## 3. Method

### 3.1 Theoretical basis

For a DMU with $m$ inputs and $s$ outputs, there are $2^m - 1$ input and $2^s - 1$ output combinations respectively. And $2^{m+s} - 1$ input-output combinations. Let $IC = \{C(m,i) = \binom{i}{m} | i = 1,\dots,m\}$, $OC = \{C(s,o) = \binom{o}{s} | o = 1,\dots,s\}$ represent the set of input and output combinations respectively. And $IC(p), OC(q)$ represent the $p_{th}$ and $q_{th}$ input and output combinations, where $1 \leq p \leq 2^m - 1$, $1 \leq q \leq 2^s - 1$. $F(\cdot)$ represents the correspondence from a DMU set to the corresponding subscript index set.

**Lemma1.** Given a group of $n$ DMUs, for $DMU_l$, if there is a pair of $p, q$ satisfies $\frac{\sum_{b\in F(OC(q))} y_{bl}}{\sum_{a\in F(IC(p))} x_{al}} = max_j\{\frac{\sum_{b\in F(OC(q))} y_{bj}}{\sum_{a\in F(IC(p))} x_{aj}}\}$, $j = 1,\dots,n$, then $DMU_l$ must be VRS efficient.

**Lemma2.** $DMU_l$ must be VRS efficient if it is the unique DMU satisfies one of the following equations:

a) $\sum_{a\in F(IC(p))} x_{al} = min_j\{\sum_{a\in F(IC(p))} x_{aj}\}$, $j = 1,\dots,n$,

b) $\sum_{b\in F(OC(q))} y_{bl} = max_j\{\sum_{b\in F(OC(q))} y_{bj}\}$, $j = 1,\dots,n$,

c) $\sum_{a\in F(IC(p))} x_{al} - \sum_{b\in F(OC(q))} y_{bl} = min_j\{\sum_{a\in F(IC(p))} x_{aj} - \sum_{b\in F(OC(q))} y_{bj}\}$, $j = 1,\dots,n$.

Since we cannot use the above 2 lemmas directly because of scale influence. So it is necessary to do preprocess (e.g. standardization) of the data before applying the lemmas.

### 3.2 Framework

The key to reduce the computation time when large dataset is present is to identify all efficient DMUs first, thus we developed the following framework:

1. start,
2. $\mathfrak{D} \leftarrow$ Get a sample of DMUs,
3. $\mathfrak{EF} \leftarrow$ Select a group of efficient DMUs as a hull,

4. $\mathfrak{E} \leftarrow$ Find exterior DMUs in $\mathfrak{D} \setminus \mathfrak{E}\mathfrak{F}$ respect to the hull of $\mathfrak{E}\mathfrak{F}$,

5. If $\mathfrak{E} = \{\ \}$:

$\mathfrak{F} = \mathfrak{E}\mathfrak{F}$ and all DMUs are already evaluated (go to 6).

Otherwise,

5.1 $\mathfrak{F} \leftarrow$ find the rest of efficient DMUs in $\mathfrak{E}$,

5.2 Evaluate DMUs in $\mathfrak{D} \setminus \mathfrak{F}$ respect to $\mathfrak{F}$,

6. End.

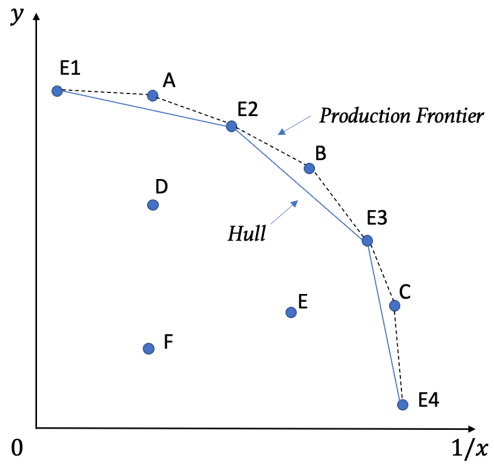


***Fig.1****. Sketch of production frontier and the hull*

We chose the 1 input and 1 output case in Fig.1 for an intuitive illustration. Each point represents a DMU. The piece-wise dash line represents the production frontier. Literally the piece-wise hull we built in step4 is formed by points E1, E2, E3, E4 on the production frontier. Points A, B, C on the above of the hull is called exteriors. Point D, E, F below the hull is called interiors.

### 3.3 Computer and software

To make comparison with PH method, in this study we just used one computer with an Intel® Core™ i3-9100 CPU @3.60GHz, 4.00GB memory and a 64-bit windows 11 operating system. We used python pulp package to solve the DEA model.

### 3.4 Datasets

In this study, we used generated datasets with dimensions ranging from 1 input 1 output to 6 inputs 6 outputs and cardinalities from 1000 to 20000 by uniform distribution. For all the datasets, Results on each data set confirm that our method performs much better in comparison with PH method.

## 4. Results

Here is part of the results:

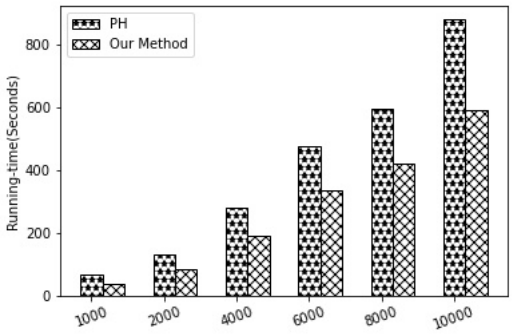


***Fig.2****. Comparison on different cardinalities (with 5 inputs and 5 outputs)*

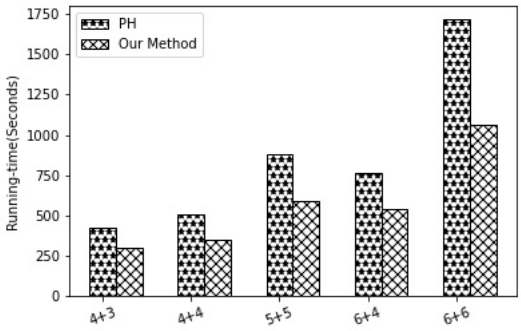


***Fig.3****. Comparison on different dimensions (with cardinality 10000)*

Fig.2 and Fig.3 shows whether with dimensions or cardinality increasement our method is faster than PH method. The major reason is that our established hull is a group of efficient DMUs, while PH's hull is only majorly formed by potentially-efficient DMUs, which resulted in larger amount of exterior set thus longer running time.

## 5. Conclusion and future work

We have proposed a framework basing on the 2 lemmas to substantially decrease the running-time for the large-scale DEA problems. Our framework presented a better performance than the current most powerful PH method. This framework can still be modified, in the future we will apply statistical learning methods or other ways to improve its performance.

## References

[1] Khezrimotlagh, D., Zhu, J., Cook, W. D., & Toloo, M. (2019). Data envelopment analysis and big data. European Journal of Operational Research, 274(3), 1047-1054.