

DATA COLLECTION BUDGET ALLOCATION FOR STOCHASTIC DATA ENVELOPMENT ANALYSIS

Loo Hay Lee
 Wai Peng Wong
 Wikrom Jaruphongs

Department of Industrial & Systems Engineering
 National University of Singapore
 10 Kent Ridge Crescent, Singapore 119260, SINGAPORE

ABSTRACT

Data Envelopment Analysis (DEA), a common tool used to measure efficiency, is unable to provide an accurate efficiency score when there are uncertainties in the data. This paper aims to address the problem, that if the users need to collect data to measure the efficiency and if data collection is expensive, how can the users allocate the resources intelligently in data collection in order to obtain a better estimate for the efficiency score.

1 INTRODUCTION

DEA (Data Envelopment Analysis) is a mathematical programming approach developed by Charnes et al. (1978) to measure efficiency for decision-making units (DMUs) with multiple inputs and multiple outputs. A main advantage of DEA is that it does not require any prior assumptions on the underlying functional relationships between the inputs and outputs (Cooper et al., 2006).

DEA is a data-driven approach which requires the estimation of the inputs/outputs data. An early criticism of DEA is that the data has to be deterministic. One of the earliest efforts in response to this was the development of chance-constrained programming (CCP) approach to accommodate stochastic variations in data (Charnes and Cooper, 1963). Extensive study on DEA using CCP has been carried out by Sengupta (1982, 1987, 1988, 1989, 1995) in which, one prominent feature is to incorporate the stochastic variables and then reformulate the stochastic model into a deterministic equivalent. Wong et al. (2008) recently suggested an alternative method to handle uncertainties in data, whereby they utilized Monte Carlo to estimate the distribution for the data and the efficiency. While all these earlier researches do offer important treatment of DEA in the presence of variations in the data, it does not provide a useful answer to the question: *Is the efficiency measured using this way accurate and how can we*

get an accurate efficiency when there are variations in the data? This paper aims at answering this question.

The remainder of the paper is organized in the following manner. Section 2 provides the problem definition. We describe the solution approach in Section 3. In Section 4, we will provide the numerical results. Finally we conclude the paper in Section 5.

2 PROBLEM DEFINITION

Let S be the set of inputs and R the set of outputs, where S and R are disjoint sets ($S \cap R = \emptyset$). Also denote K is the set of combined inputs/outputs; i.e., $K = S \cup R$. J is the set of DMUs. Let $\mathbf{X}_D = (x_{kj})_{k \in K, j \in J}$, where x_{kj} represents k -th input/output for DMU j . If $k \in S$, then x_{kj} is an input; otherwise if $k \in R$, then x_{kj} is an output. Given that \mathbf{X}_D is the matrix for the initial data of the inputs/outputs, the efficiency score for DMU j_0 denoted by $\theta(\mathbf{X}_D)$ can be computed using (1).

DEA model (adapted from Cooper et al., 2006)

$$\begin{aligned} \theta(\mathbf{X}_D) &= \min \quad \theta \\ \text{s.t.} \quad &\sum_{j \in J} \lambda_j x_{sj} \leq \theta x_{s_{j_0}} \quad s \in S \\ &\sum_{j \in J} \lambda_j x_{rj} \geq x_{r_{j_0}} \quad r \in R \\ &\lambda_j \geq 0, \quad j \in J \quad (1) \end{aligned}$$

The λ_j 's are the weights (decision variables) of the inputs/outputs that optimize the efficiency score of DMU j_0 . Model (1) attempts to proportionally contract DMU j_0 's inputs as much as possible while not decreasing its current level of outputs. Hence, the efficiency score will be equal to one if a DMU is efficient and less than one if a DMU is inefficient.

In the deterministic model i.e. when the true values of the input/output are known, the efficiency score can be eas-

ily calculated from (1) as a discrete a value. However, in actual application, i.e. when the environment is stochastic, the true values maybe unknown. Without loss of generality, if the inputs/outputs values that the users interested in are the true mean of the data, in order to use the same DEA model to estimate the efficiency, the users have to collect some data and then use the sample mean to replace the true mean. For instances, for cycle time, users have to collect a few data and use the sample mean to represent its true mean. The limitation of this method is that the sample mean is not the true mean unless the users have collected infinite amount of data; if there are only a few data collected, the sample means may be very different from the true mean, hence the efficiency score will not be accurate.

An alternative way to tackle this issue is based on the data collected, derive the belief (distribution) for the true mean for the stochastic inputs/outputs variables using Bayesian framework. Then, using the distribution of these inputs/outputs, we can estimate the distribution of the efficiency scores. This is called the Monte Carlo DEA approach. Interested readers may refer to Wong et al.(2008) for detailed implementation of the approach. In reality, when data collection is expensive, any attempt to collect data often raises one question. *How should we allocate the budget - how many data should we collect for each input/output?* If we naively allocate the data collection effort fairly, the efficiency estimated might not be accurate. Hence, it is important to know how to allocate our budget for data collection in order to get a better estimate for the efficiency. This question is similar to the question asked in the optimal computing budget allocation (OCBA) problem (see Chen et al. 2000), in the sense that OCBA is interested to assign different computing budgets to different designs so as to maximize the probability of correct selection, while in this problem we need to assign data collection budget to different inputs and outputs so as to maximize the accuracy of the efficiency prediction.

To formulate the data collection budget allocation problem (DCBA), we first need to define the objective of the problem. In this paper we use mean square error (MSE) which is defined as follows as the measure of accuracy for the efficiency.

$$MSE = E(\tilde{\theta} - \theta_D)^2 \quad (2)$$

where $\tilde{\theta}$ is the belief towards where the true efficiency lies and θ_D is the efficiency calculated from the initial data collected. Note that we use $\theta(\mathbf{X}_D)$ from (1) to represent θ_D .

The problem statement which we aim to solve in this paper is if we can only collect limited additional data, how do we distribute our efforts in collecting data so that we can get a better prediction of the efficiency. Assume we have collected some initial data for all the inputs/outputs and let \mathbf{X} be the matrix of their sample averages, we will determine the data collection scheme (or allocation design) so as to minimize the MSE of efficiency score subject to a limited

budget for additional samples, denoted by N . The allocation design is given by $\mathbf{n} = [n_k]_{k \in K}$, where n_k represents the number of additional data collected for input/output k . Hence, the mathematical model for our research problem is as shown below.

Problem DCBA

$$\begin{aligned} \min F(\mathbf{n}) &= E[(\tilde{\theta}(\mathbf{X}') - \theta(\mathbf{X}))^2] \\ \text{s.t. } \sum_{k \in K} n_k &= N \end{aligned} \quad (3)$$

The objective function $F(\mathbf{n})$ is defined as the MSE of the efficiency score for allocation design \mathbf{n} where \mathbf{X}' is the belief of the inputs/outputs after additional data are collected following the allocation design \mathbf{n} . Note that $\theta(\mathbf{X})$ is the efficiency score computed using Model 1. $\tilde{\theta}(\mathbf{X}')$ represents the belief for the true efficiency. The above model cannot be solved directly because the distribution for $\tilde{\theta}(\mathbf{X}')$ is unknown. In order to estimate $F(\mathbf{n})$, we need to derive \mathbf{X}' using Bayesian framework. Let $\mathbf{X}' = [x'_k]_{k \in K}$ and x'_k is a random variable. Under the Bayesian model, the belief for the unknown true value/mean of the input/output k denoted by μ_k is treated as a random variable and has a prior distribution which describes the knowledge or the subjective belief about μ_k before any sampling. The posterior distribution is updated after observed the samples $\{\hat{x}_k(t), t = 1, \dots, n_{ok}\}$. Note that n_{ok} = total number of samples, $\hat{x}_k(t)$ = the t -th sample of the performance measure. Similar to Chen et al. (2000), we assume that the μ_k has a conjugate normal prior distribution and consider non-informative prior distribution which implies that no prior knowledge is available about the performance of any design before conducting simulation. In that case, DeGroot (1970) shows that the posterior distribution of μ_k is

$$x'_k \sim N\left(\bar{x}_k, \frac{\sigma_k^2}{n_{ok}}\right) \text{ where } \bar{x}_k = \frac{1}{n_{ok}} \sum_{t=1}^{n_{ok}} \hat{x}_k(t) \text{ is the sample mean of the observations and } \sigma_k^2 \text{ is the true variance which can be approximated by the sample variance } \sigma_k^2 = \frac{1}{n_{ok} - 1} \sum_{t=1}^{n_{ok}} (\hat{x}_k(t) - \bar{x}_k)^2.$$

Based on the approximations made in Chen et al. (1997), when sample size increases, \bar{x}_k and σ_k^2 do not change and if additional n_k samples are collected, the predicted posterior distribution

$$\text{for } \mu_k \text{ can be approximated by } x'_k \sim N\left(\bar{x}_k, \frac{\sigma_k^2}{n_{ok} + n_k}\right).$$

After quantifying \mathbf{X}' , we can then estimate the distribution of the efficiency score through Monte Carlo method. Hence, we are able to estimate F given a value of \mathbf{n} as shown in (4) below.

$$F(n) \approx \frac{1}{M} \sum_{i=1}^M (\theta(\hat{\mathbf{X}}_{[i]}) - \theta(\mathbf{X}))^2 \quad (4)$$

where $\hat{\mathbf{X}}_{[i]}$ is the realization of the inputs/outputs \mathbf{X}' in the replication i of the Monte Carlo run for allocation design \mathbf{n} and M is the number of random data set.

Remarks: As a start, we focus on one DMU, i.e. θ is only for the DMU that we are interested in. The model can be generalized to all DMUs as well. Next, we present methods to solve it.

3 SOLUTION APPROACH

As there is no close form to compute $F(\mathbf{n})$, we need to estimate the $F(\mathbf{n})$ using Monte Carlo sampling. We can treat this problem as a simulation optimization problem. In this paper, we will use Genetic Algorithm (GA) to search the different allocation rules, \mathbf{n} , while OCBA-m to allocate the Monte Carlo sample in estimating $F(\mathbf{n})$.

GA works with a finite population, which evolves from one generation to the next, governed by the principles of natural selection and survival of the fittest among the individuals. Each generation consists of a population of chromosomes representing the possible solutions. Based on a random generated initial population, at every generation, GA evaluates the chromosomes and ranks them according to their fitness. The fitter chromosomes are selected to generate new offsprings by recombination and mutation operators. This evaluation-selection-reproduction cycle is repeated until a satisfactory solution is found.

Without loss of generality, we let the set of inputs/outputs $K = \{1, \dots, D\}$. We then use an array of integers to represent a solution or a chromosome as illustrated in Figure 1. We add the conditions for the chromosome in such a way that the sum of the integers equal to the budget,

$$\text{i.e. } \sum_{k \in K} n_k = N .$$

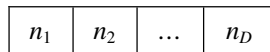


Figure 1: A chromosome representation

As we aim to balance towards more ‘exploitation’ and less ‘exploration’, we use ‘tournament selection’ whereby we randomly select some number of parents from the generation to form a tournament (sub-population) and then select the individual with the best fitness in this sub-population, effectively winning the tournament. We repeat the process several times until the required number of winners is chosen. The reproduction of two offsprings is obtained by a two-position crossover on two parents where the two positions are generated randomly. Note that if the offsprings are not feasible we will repair them. Each offspring is then assigned a small probability of mutation to create more diversification to the solutions.

OCBA-m is used to determine the number of samples to be assigned to each solution to estimate its fitness. The

details of OCBA-m algorithm can be found in Chen et al. (2008).

4 NUMERICAL RESULTS

The results obtained from the GA method will be compared against the ‘uniform’ allocation method where the data collection budget is equally allocate. Experiments are performed using the data sets and the supply chain model from Wong et al. (2008). The inputs/outputs which are considered to be stochastic are determined based on sensitivity analysis, i.e. starting from the most influential till the least influential on the efficiency. Note that D represents the total number of stochastic inputs/outputs.

Table 1: Comparison of N and savings

D	MSE	GA	Uniform	Savings
5	0.12694	30	120	4.00
5	0.07080	60	425	7.08
5	0.06503	90	695	7.72
10	5.98E-03	30	1570	52.33
10	3.30E-03	60	3560	59.33
10	2.31E-03	90	4930	54.78

The results showed that the savings can be very significant, which is as high as 50 times. Moreover, the saving increases when the size of the problem increases.

5 CONCLUSIONS

In this paper, we provided an approach on how to distribute the data collection budget so as to get a better estimate for the efficiency score in the circumstances when there are uncertainties or variations in the data. The results show that saving from allocating the budget intelligently can be very significant.

Some of the future research includes how to find better way in solving the DCBA problem.

REFERENCES

Charnes A., Cooper, W.W., 1963. Deterministic equivalents for optimizing and satisficing under chance constraints. *Management Science* 11, 18-39.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the inefficiency of Decision Making Units. *European Journal of Operational Research* 2(6), 429-444.

Chen, H. C., Chen, C. H. , Dai, L., Yücesan, E., 1997. New Development of Optimal Computing Budget Allocation For Discrete Event Simulation, *Proceedings of the 1997 Winter Simulation Conference*, pp. 334-341.

- Chen, C. H., Lin, J., Yücesan, E., Chick, S. E., 2000. Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization. *Journal of Discrete Event Dynamic Systems: Theory and Applications* 10, 251-270.
- Chen, C.H., He, D., Fu, M., Lee, L.H., 2008. Efficient Simulation Budget Allocation for Selecting an Optimal Subset. *INFORMS Journal on Computing* 20(4), 579-595.
- Cooper, W.W., Seiford, L.M., Tone, K., 2006. *Introduction to Data Envelopment Analysis and its Uses: with DEA-Solver Software and Reference*, Springer, New York.
- DeGroot, M.H., 1970. *Optimal Statistical Decisions*, McGraw-Hill, Inc., New York.
- Sengupta, J.K., 1982. Efficiency measurement in stochastic input-output systems. *International Journal of System Science* 13, 273-87.
- Sengupta, J.K., 1987. Data envelopment analysis for efficiency measurement in the stochastic case. *Computer Operations Research* 14, 117-129.
- Sengupta, J.K., 1988. Robust efficiency measures in a stochastic efficiency. *International Journal of Systems Science* 19, 779-791.
- Sengupta, J.K., 1989. Data envelopment analysis with maximum correlation. *International Journal of System Science* 20, 2085-2093.
- Sengupta, J.K., 1995. *Dynamics of Data Envelopment Analysis*. Kluwer Academic Publishers, Netherlands.
- Wong, W.P., Jaruphongsa, W., Lee, L.H., 2008. Supply Chain Measurement System – A Monte Carlo DEA based approach. *International Journal of Industrial and Systems Engineering* 3(2), 162-188.

WIKROM JARUPHONGSA is currently with Mon Transport Co. Ltd. in Thailand. He was an Assistant Professor in the Department of Industrial and Systems Engineering at the National University of Singapore during 2002-2008. He obtained his Ph.D. in Industrial Engineering from Texas A&M University. His research interest is in the areas of logistics, supply chain management, and applied optimization. (email: wikromj@gmail.com).

AUTHOR BIOGRAPHIES

LOO HAY LEE is an associate professor in the department of Industrial and Systems Engineering at National University of Singapore. His research interest lies in the area of Simulation-based optimization, Maritime logistics and supply chain system. He serves in the editorial board for the *Asia Pacific Journal of Operational Research* and is a member in the advisory board for *OR Spectrum*. (email: iseleeelh@nus.edu.sg).

WAI PENG WONG is a PHD candidate at the Department of Industrial and Systems Engineering, National University of Singapore. Her research interests include supply chain management, simulation and optimization. (email: g0500743@nus.edu.sg).