Relaxed normality assumption in stochastic DEA for efficient handling of Big Data

Panagiotis D. Zervopoulos

Department of Health Management
Open University of Cyprus
Cyprus

Ioannis Mitropoulos

Department of Business Administration
Technological Educational Institute of Western Greece
Greece

20th Conference of the International Federation of Operational Research Societies

Barcelona, 13 - 18 July, 2014

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: ARCHIMEDES III. Investing in knowledge society through the European Social Fund.

1. The scope

- Efficiency measurement when noisy data are present
- Efficiency measurement using big data sets
- Limitation of the computational burden

2. Introduction to the methodology

- Data Envelopment Analysis (DEA) is a widely used non-parametric, linear-programming method for measuring efficiency
- Stochastic DEA, under the assumption of normal distribution, deals effectively with noise in the data set
- Stochastic DEA is a non-linear programming extension of conventional DEA programs

3. The problem

- Stochastic DEA, under the assumption of normal distribution, requires:
  - significant computational burden
  - at least x50 the time that a stochastic DEA, under the assumption of uniform distribution, needs

4. Methodology

- CRS DEA

$$
\begin{aligned}
&\min \theta \\
&s.t. \ \sum_{j=1}^{n} \lambda_j x_{ij} \le \theta x_{io} \quad i = 1, \ldots, m \\
&\quad \sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{ro} \quad r = 1, \ldots, s \\
&\quad \lambda_j \ge 0 \qquad\qquad\qquad\qquad (1)
\end{aligned}
$$

- Assuming that:

$$\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right) \sim N\left(E\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right), \mathrm{var}\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right)\right) \quad \text{and}$$

$$\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right) \sim N\left(E\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right), \mathrm{var}\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right)\right)$$

4. Methodology

- The chance-constrained stochastic DEA expression of program (1) is written as follows:

$$
\begin{aligned}
&\min \theta \\
&s.t.\quad \mathrm{P}\left(\sum_{j=1}^{n} \lambda_j x_{ij} \le \theta x_{io}\right) \ge a \quad i = 1,\dots,m \\
&\qquad\ \ \mathrm{P}\left(\sum_{j=1}^{n} \lambda_j y_{rj} \ge y_{ro}\right) \ge a \quad r = 1,\dots,s \\
&\qquad\ \ \lambda_j \ge 0 \qquad\qquad\qquad\qquad (2)
\end{aligned}
$$

- Program (2) is transformed as follows:

$$
P\left(\frac{\sum_{j=1}^{n} \lambda_j x_{ij} - E\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right)}{\left(\mathrm{var}\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right)\right)^{1/2}} \le \frac{\theta x_{io} - E\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right)}{\left(\mathrm{var}\left(\sum_{j=1}^{n} \lambda_j x_{ij}\right)\right)^{1/2}} \right) \ge \alpha \quad \text{and}
$$

$$
1 - P\left(\frac{\sum_{j=1}^{n} \lambda_j y_{rj} - E\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right)}{\left(\mathrm{var}\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right)\right)^{1/2}} \le \frac{y_{ro} - E\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right)}{\left(\mathrm{var}\left(\sum_{j=1}^{n} \lambda_j y_{rj}\right)\right)^{1/2}} \right) \ge \alpha \qquad (3)
$$

4. Methodology

- Finally, program (2) is defined:

$$\min \theta$$

$$s.t.\quad \sum_{j=1}^{n}\lambda_j x_{ij} + \left(E\left(\sum_{j=1}^{n} x_{ij} \right) - \sum_{j=1}^{n} x_{ij} \right)\lambda_j + \Phi^{-1}(\alpha)\left(\sum_{j=1}^{n}\sum_{k=1}^{l}\lambda_j\lambda_k \operatorname{cov}\left(x_{ij}, x_{ik}\right) \right)^{1/2} \le \theta x_{io}$$

$$\sum_{j=1}^{n}\lambda_j y_{rj} + \left(E\left(\sum_{j=1}^{n} y_{rj} \right) - \sum_{j=1}^{n} y_{rj} \right)\lambda_j - \Phi^{-1}(\alpha)\left(\sum_{j=1}^{n}\sum_{k=1}^{l}\lambda_j\lambda_k \operatorname{cov}\left(y_{rj}, y_{rk}\right) \right)^{1/2} \ge y_{ro}$$

$$\lambda_j \ge 0 \qquad (4)$$

where $\Phi^{-1}(\alpha) = 1.645$ for $\alpha = 0.05$

4. Methodology

- Assuming that:

$$\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right) \sim U\left(\min\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right),\ \max\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right)\right) \quad \text{and}$$

$$\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right) \sim U\left(\min\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right),\ \max\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right)\right)$$

- Program (1) is written as follows:

$$\frac{\theta x_{io} - \min\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right)}{\max\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right) - \min\left(\sum_{j=1}^{n}\lambda_j x_{ij}\right)} \geq \alpha \qquad \text{and}$$

$$1 - \frac{y_{ro} - \min\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right)}{\max\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right) - \min\left(\sum_{j=1}^{n}\lambda_j y_{rj}\right)} \geq \alpha \qquad (5)$$

4. Methodology

- Finally, we obtain the stochastic DEA program, under uniform distribution:

$$
\begin{aligned}
&\min \theta \\
&s.t. \quad \sum_{j=1}^{n} \lambda_j x_{ij} + \sum_{j=1}^{n} \lambda_j \left(\alpha \left(x_{\max,i} - x_{\min,i} \right) + x_{\min,i} - x_{ij} \right) \le \theta x_{io} \\
&\qquad \sum_{j=1}^{n} \lambda_j y_{rj} + \sum_{j=1}^{n} \lambda_j \left(\left(1-\alpha\right) \left(y_{\max,r} - y_{\min,r} \right) + y_{\min,r} - y_{rj} \right) \ge y_{ro} \\
&\qquad \lambda_j \ge 0
\end{aligned}
\qquad (6)
$$

5. Numerical example: Empirical results

- Dataset: analgesics market in the UK

Table 1. Selected market data

PRODUCT ID	STORE	MARKET	RETAILER_ID	a_vol	a_prv	ISM	DISP	WAISM	WADISP	MULTI	WA MULT	ACV
81	14457	22	40002	30	0.19	0.056	0.03	0.687	0.391	0.101	0.301	1.541
85	14464	27	40002	20	0.2	0.34	0.05	0.731	0.421	0.308	0.195	2.124
79	14465	23	40002	24	0.222	0.37	0.09	0.459	0.207	0.621	0.341	1.554

B1G1FRi	B1G25LPi	B1G2HPi	B1G50LPi	B1 GDOUBLELPi	B1 GFRWTRi	B2 F1000Pi	B2 F100Pi	B2F450Pi	B2F750Pi	B3 F1000Pi	B3FPR2i	HALFPRICEi
0.515	0.390	0.322	0.287	0.540	0.487	0.286	0.148	0.213	0.400	0.186	0.292	0.332
0.678	0.610	0.632	0.552	0.697	0.234	0.625	0.757	0.571	0.687	0.511	0.626	0.512
1.021	0.920	0.462	0.779	0.513	0.398	0.518	0.373	0.530	0.903	0.629	1.102	0.759

PP150Pi	PP250Pi	PP450Pi	RB100Pi	RD100Pi	RD200Pi	RD300Pi	RD400Pi	RD500Pi	S100Pi	S150Pi	S15PCi	S200Pi
0.251	0.256	0.316	0.267	0.149	0.460	0.212	0.505	0.298	0.304	0.544	0.389	0.290
0.772	0.670	0.806	0.594	0.591	0.647	0.817	0.241	0.349	0.790	0.884	0.569	0.803
0.851	0.678	0.658	0.582	0.724	0.478	0.572	0.540	0.725	0.927	1.024	0.437	0.752

S20PCi	S25PCi	S3RDi	S50Pi	B1G1FRd	B1G25LPd	B1G2HPd	B1 G50LPd	B1 GDOUBLELPd	B1 GFRWTRd	B2 F1000Pd	B2 F100Pd	B2F450Pd
0.226	0.272	0.410	0.477	0.357	0.276	0.400	0.451	0.287	0.566	0.295	0.485	0.495
0.926	0.628	0.532	0.528	0.562	0.720	0.574	0.906	0.549	0.493	0.682	0.635	0.669
0.602	0.538	0.716	0.629	0.708	0.815	0.560	0.677	0.445	0.643	0.848	0.438	0.759

B2F750Pd	B3F1000Pd	B3FPR2d	HALFPRICEd	PP150Pd	PP250Pd	PP450Pd	RB100Pd	RD100Pd	RD200Pd	RD300Pd	RD400Pd	RD500Pd
0.358	0.337	0.481	0.079	0.355	0.360	0.405	0.217	0.138	0.002	0.432	0.512	0.283
0.286	0.577	0.325	0.628	0.715	0.777	0.713	0.673	0.716	0.592	0.594	0.753	0.763
0.497	0.733	0.632	0.658	0.459	0.563	0.671	0.711	0.671	0.604	0.899	0.671	0.584

S100Pd	S150Pd	S15PCd	S200Pd	S20PCd	S25PCd	S3RDd	S50Pd	sv_5_15d	sv_5_15i	sv_15_25d	sv_15_25i	sv_25_35d
0.707	0.254	0.153	0.328	0.366	0.452	0.368	0.235	0.122	0.262	0.145	0.459	0.334
0.529	0.330	0.256	0.357	0.450	0.667	0.484	0.587	0.655	0.392	0.472	0.488	0.480
0.586	0.689	0.839	0.636	0.831	0.734	0.616	0.577	0.727	0.685	0.564	0.581	0.781

sv_25_35i	sv_35_45d	sv_35_45i	sv_45_55d	sv_45_55i	sv_gt55d	sv_gt55i
0.072	0.405	0.212	0.553	0.358	0.384	0.232
0.534	0.718	0.524	0.704	0.651	0.629	0.746
0.778	0.653	0.824	0.648	0.683	0.885	0.619

5. Numerical example: Empirical results

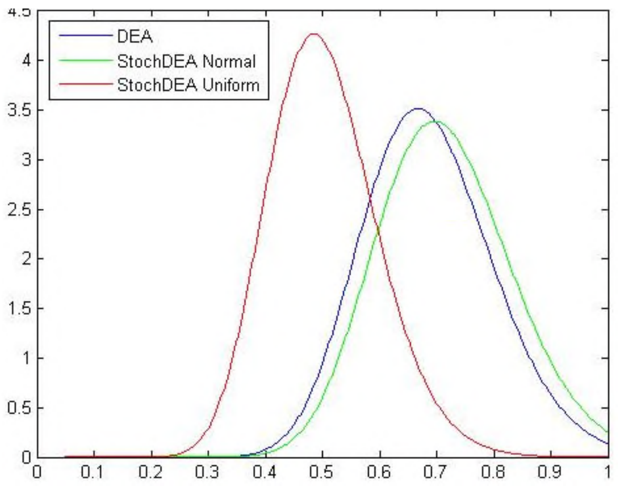

Figure 1. Distributions of efficiency scores (500 units)

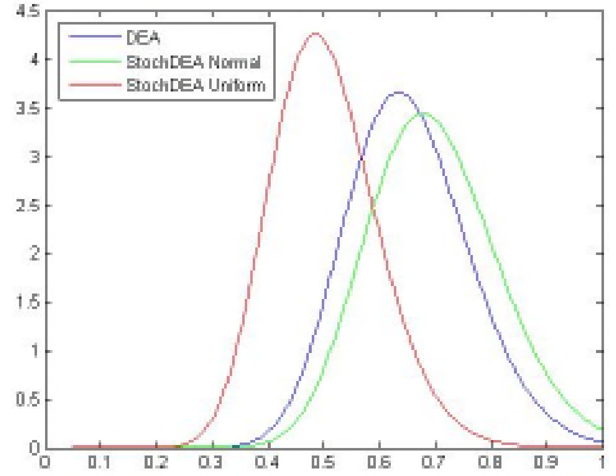

Figure 2. Distributions of efficiency scores (10,000 units)

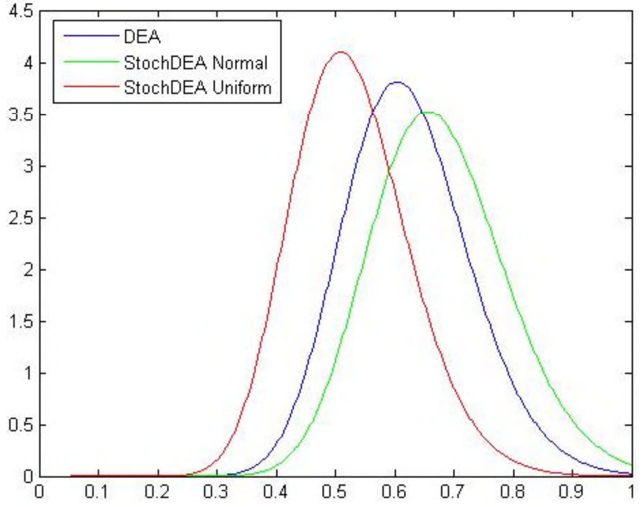

Figure 3. Distributions of efficiency scores (100,000 units)

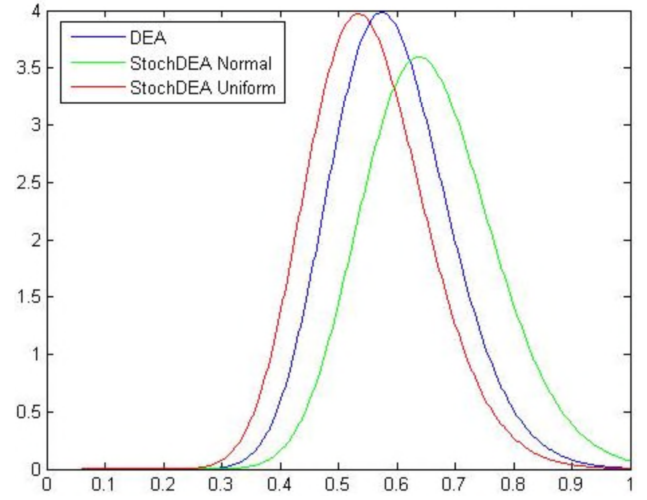

Figure 4. Distributions of efficiency scores (500,000 units)

5. Numerical example: Empirical results

Table 2. Comparative analysis using the t-test

Methods			Units	Mean	St. Deviation	95% Confidence Interval of the Difference		t	p-value (2-tailed)
						Lower	Upper		
	SDEA Normal	SDEA Uniform	400	0.21544	0.04461	0.21105	0.21983	96.48	0.00000
DEA	SDEA Normal		400	-0.02996	0.00713	-0.03066	-0.02925	-83.92	0.00000
DEA		SDEA Uniform	400	0.18548	0.04368	0.18118	0.18978	84.82	0.00000
	SDEA Normal	SDEA Uniform	10,000	0.19438	0.04148	0.19029	0.19846	93.60	0.00000
DEA	SDEA Normal		10,000	-0.04291	0.00880	-0.04378	-0.04205	-97.45	0.00000
DEA		SDEA Uniform	10,000	0.15147	0.03950	0.14758	0.15535	76.59	0.00000
	SDEA Normal	SDEA Uniform	100,000	0.14921	0.03675	0.14559	0.15282	81.10	0.00000
DEA	SDEA Normal		100,000	-0.05480	0.01144	-0.05592	-0.05367	-95.64	0.00000
DEA		SDEA Uniform	100,000	0.09441	0.03490	0.09097	0.09784	54.04	0.00000
	SDEA Normal	SDEA Uniform	500,000	0.10367	0.03245	0.10048	0.10687	63.82	0.00000
DEA	SDEA Normal		500,000	-0.06568	0.01440	-0.06710	-0.06426	-91.13	0.00000
DEA		SDEA Uniform	500,000	0.03799	0.03150	0.03489	0.04109	24.09	0.00000

5. Numerical example: Empirical results

- Estimated time for measuring efficiency scores

Table 3. Regression model summary

Methods	F	p-value	Adjusted R^2
SDEA Normal	52.69	0.0000	0.8960
SDEA Uniform	1390.99	0.0000	0.9957

Table 4. Regression model coefficients

Stochastic DEA (Normal distribution)				
	Coefficients	Confidence Interval		p-value
		Lower bound	Upper bound	
Units	2.9505	2.0929	3.8080	0.0000
Variables	74.3156	17.8651	130.7661	0.0150
Stochastic DEA (Uniform distribution)				
Units	0.0491	0.0433	0.0549	0.0000
Variables	1.6874	1.3054	2.0695	0.0000

Table 5. Estimated processing time

Method	Units	Variables	Processing Time (Hours)
SDEA Normal	1,000,000	4	819.49
SDEA Uniform	1,000,000	4	13.65

Thank you!