確率的DEA法

森田　浩

1. はじめに

データに基づいた手法ではそのデータのもつ不確実性を考慮に入れる必要がある．統計的手法はこの不確実性をばらつきとして積極的に捉えて評価している手法であるといえよう．何回か観測して平均値を取るのも，推定量の分散を小さくしたり，ばらつきの大きさを同定したりするためである．回帰分析や分散分析でも同様である．そこでは，データにはばらつきがあるものという前提で理論化されているのである．

一方，DEA (data envelopment analysis) ではデータのもつ不確実性は陽には取り入れられていない．統計的手法では得られたデータから全体像を把握するための評価を推定しようとしているが，データの値が実際に得られた状況におけるDMU (decision making unit) そのものの評価を下そうとしているのがDEAである．この違いがデータのばらつきを積極的に捉えているか否かにつながっているのである．

不確実性を陽に考えないにもかかわらずDEAが広く使われているのは，数理計画手法によってさまざまな情報が提供されることも一因であろう．効率的でないものに対する改善目標や非効率要因の特定が可能であり，モデルの拡張性でも優れている．偶発的な「誤差」をどう捉えるかに注力している統計学とシステムのもつ構造的な「誤差」を捉えようとしているDEAという見方もできる．

しかしながら，DEAで不確実性を考える必要がないということはありえず，DEAの確定的なモデルが多く開発されてきている今日では，次の段階としてデータ固有の偶発的「誤差」を陽に取り入れたDEAモデルの開発も必要となっている．Bauer[2]は「データには統計的ノイズがあるためにそこから得られる効率的フロンティアもそのノイズに汚染されたものになる」と指摘しており，データの持つエラーに関する明確な説明がなされない限り非統計的な効率性尺度に懐疑的な統計学者は多いだろうと言っている．Seifordは1996年に報告したDEAの現状をまとめた論文[8]で「今後10年では確率的DEAの研究が最も注目を集めるであろう」と予見している．その10年の半分を過ぎた今，不確実性の下でのDEA分析手法にいくつかの進展が見られている．浅学のためその多くを把握することはできていないが，いくつかを以下で紹介したい．

もりた　ひろし
大阪大学大学院　工学研究科
〒565-0871 吹田市山田丘2-1

2. ファジィDEA法と確率的DEA法

確率的DEA法の話題に入る前にファジィDEA法について少し触れておきたい．

データのもつ不確実性を記述する数学的な方法には，確率的な変動として表す確率変数として扱う方法とあいまいなものとして表すファジィ数として扱う方法があろう．当初，不確実データを扱うDEAとしては前者の確率的データによる分析があったが，多方面で実用化されるにつれてファジィ数による評価法も多く見られるようになってきた（例えば，乾口ら[6]や上田[10]）．確率的DEA法では確率的計画法による定式化がなされるのに対し，ファジィ数を扱うモデルではファジィ計画法によって定式化される．求解のアルゴリズムの点からみると，確率的計画法よりファジィ計画法の方が扱いやすいこともあるが，特に日本ではファジィ理論の研究者たちが新たなデータ解析手法としてDEAに注目し，そのファジィDEA法への拡張が進められている．ファジィDEA法では得られる効率値もファジィ数として与えられる．効率性もあいまいなものになるという考え方である．一方，確率的DEA法では分布問題として扱う場合を除いて，一般に，ある規範を設けて確率的な条件を確定的な問題へと帰着させているため，得られる効率値は確定した数値として与えられる．もちろん効率値は確定値であっても確率的変動の大きさなどは加味されて算出されて

おり，通常の確定的なDEA法とは異なる．

もっとも不確実性のあるのはモデルではなくデータであるから，ファジィ数として表すのか確率変数として表すのかはデータによって決められるものである．統計的な予測理論に基づいた推定値などを使うときには確率的な枠組みで議論すべきであろう．そのような理論によらない予測値を使ったり，漠然としていて確定的な値がわからなかったりする場合にはファジィ数として議論すべきである．

3. DEA

m 入力 s 出力をもつDMUが n 個あり，DMU_j は入力 x_j と出力 y_j をとるものとする．CCRモデルでは生産可能集合（production possibility set）P には次の仮定がおかれている．

(A 1)　すべての $j=1,2,\cdots,n$ に対して $(x_j, y_j)\in P$

(A 2)　正の k に対して $(x, y)\in P$ ならば $(kx, ky)\in P$

(A 3)　$\bar{x}\geq x, \bar{y}\leq y$ に対して $(x, y)\in P$ ならば $(\bar{x}, \bar{y})\in P$

この生産可能集合 P の有効フロンティア上にあるDMUが効率的と評価され，DMU_o の効率値を求めるための計画問題は次の問題(1)の包絡形とその双対形である問題(2)の乗数形となる．

$$\text{包絡形}\quad\left|\begin{array}{ll}\text{minimize} & \theta\\ \text{subject to} & \theta x_o - X\lambda\geq 0\\ & y_o - Y\lambda\leq 0\\ & \lambda\geq 0\end{array}\right. \tag{1}$$

$$\text{乗数形}\quad\left|\begin{array}{ll}\text{maximize} & u'y_o\\ \text{subject to} & v'x_o=1\\ & -v'X+u'Y\leq 0\\ & u\geq 0, v\geq 0\end{array}\right. \tag{2}$$

問題(1)では，$(\theta x_o, y_o)\in P$ となる最小の θ を求めていることになる．その値 θ^* は効率値（efficiency score）と呼ばれ，$\theta^*=1$ ならば P の境界線上にあり，$\theta^*<1$ ならば P の内部にあることがわかる．この境界線上にあるDMUをD効率的（D-efficient）といい，さらに入出力にスラックがないものを効率的（efficient）という．

一方，問題(2)では評価対象のDMUにとって最も有利なウェイト (u^*, v^*) が決められる．効率値が1となるDMUではこのウェイトは一意に決まらない．各DMUにとって効率値が1となるウェイトの集まりをウェイト空間における支配領域といい，この支配領域の内部にあるウェイトは強相補性条件が成立しており，多少のノイズが加わってもその効率性評価に影響しない．

4. 効率値の信頼性

ばらつきのあるデータはその平均値などの代表値を用いて解析することになるが，その場合には得られた効率性評価がばらつきに対してどの程度信頼できるものであるかが注目される．一般に，効率的なDMUの効率値は1となるため，効率的なもの同士では優劣がつけられない．効率的なDMUの順序付けの方法としてAndersenらの拡張効率値（super efficiency score）[1]があり，広く使われている方法である．これは評価対象以外のDMUで構成される生産可能集合 P_{-o} に基づいて算出した効率値のことで，効率的なDMUは生産可能集合 P_{-o} からの距離に応じて1以上の効率値を取るようになっている．この値が大きければ，効率的という評価は多少のデータのばらつきに対して影響を受けることはない．拡張効率値がばらつきに対する信頼性を表していると見ることもでき，効率的と評価されたDMUのばらつきに対するロバスト性を示している．しかし，極端なウェイト付けによって効率的となっている特異な効率的DMUは拡張効率値が大きくなりすぎることも知られており，その値の大小によってロバスト性を比較することは必ずしも適切ではない．

データのもつばらつきを直接扱い，ある確率水準 α におけるデータの取り得る信頼領域 S_α を考えることによって効率性評価の信頼性を測ることができる[5]．データが領域 S_α 内の値をとるときの最悪の効率値を確率水準 α における効率値 θ_α とすると，θ_α は次のミニマックス問題から求めることができる．

$$\left|\begin{array}{ll}\min\limits_{x_o,y_o}\ \max\limits_{u,v} & \theta_\alpha = v'y_o\\ \text{subject to} & u'x_o=1\\ & -u'X+v'Y\leq 0\\ & u, v\geq 0\\ & (x_o, y_o)\in S_\alpha\end{array}\right. \tag{3}$$

このとき最適値 θ_α^* を α 効率値（α-efficiency score）と呼ぶ．$\theta_\alpha^*=1$ となる α の最大値 $\alpha_{\max}$ は，確率水準 $\alpha_{\max}$ で起こり得る任意の確率変動に対して常に効率的になることを示しており，この確率水準をもって効率性の信頼度（reliability）と呼ぶことにする．信頼領域は正規性の仮定の下では2次関数として表される．

問題(3)を直接解くことによって $\alpha_{\max}$ を求めるのは難しいので，双対問題に信頼領域制約を付け加え，さらに信頼領域の大きさが確率水準 α と単調な関係にあることから，次の2次計画問題により信頼度を求めることができる．

$$\left|\begin{array}{ll} \text{minimize} & \Delta=(x_o-\bar{x}_o,\ y_o-\bar{y}_o)'\Sigma_o^{-1} \\ & \quad (x_o-\bar{x}_o,\ y_o-\bar{y}_o) \\ \text{subject to} & y_o-Y_{-o}\lambda\leq 0 \\ & x_o-X_{-o}\lambda\geq 0 \\ & \lambda\geq 0 \end{array}\right. \tag{4}$$

入出力データ (x_o, y_o) は正規分布 $N((\bar{x}_o, \bar{y}_o), \Sigma_o)$ に従うものとしている．また，X_{-o}, Y_{-o} は評価対象である DMU_o の入出力データを外して得られるそれぞれ $m\times(n-1)$, $s\times(n-1)$ 行列であり，拡張効率値を求めるときと同じ生産可能集合 P_{-o} を構成することになる．つまりこの問題の最適値は，評価対象以外のDMUで構成される生産可能集合までのマハラノビス距離を求めていることになる．

このときの信頼度 $\alpha_{\max}$ は，問題(4)の最適値 Δ^* が自由度 $m+s$ の χ^2 分布に従うことから，

$$\chi^2_{m+s}(\alpha_{\max})=\Delta^* \tag{5}$$

として求められる．極端な重み付けによって効率的と評価されたものは，特定の入出力項目のみが不確実性の影響を強く受けるようになり，その結果，信頼度は低いものになってしまう．

ここでは入出力データは紙面の都合上掲載しないが，刀根[9]にある病院のデータに適用したところ，14の病院のうち，5つの病院（No. 2, 3, 6, 8, 10）が効率値1をとりD効率的となった．これらの信頼度などを表1に示す．拡張効率値はいずれも1より大きくなっているが，Δ^* とそこから導出される信頼度 $\alpha_{\max}$ より，No. 8は効率的ではあるがその信頼度はきわめて低くわずかな確率的変動に対しても非効率的となってしまうことがわかる．No. 8は残りの13個のDMUから構成される生産可能集合のフロンティアに極めて接近していることを意味している．No. 2とNo. 6の信頼度は高く，ここでの不確実性に対しては，確信をもって効率的であるということができる．詳細な解析結果は文献[5]を参照されたい．

表1　効率的と評価されたDMUの信頼度

DMU	2	3	6	8	10
拡張効率値	1.06	1.02	1.08	1.00	1.04
Δ^*	6.42	0.92	14.6	0.00	4.34
信頼度 $\alpha_{\max}$	0.83	0.08	0.99	0.00	0.64
95%効率値	.988	.955	1	.933	.981
99%効率値	.975	.944	1	.920	.971

信頼領域を構成するにあたってはデータ構造をあらわすパラメータを推定する必要がある．特にデータの分散成分の推定が重要となるが，事前にこれらがわかっている場合はほとんどなく，与えられた入出力データから推定することが多くなる．しかしながら，通常のデータセットでは分散を精度よく推定するために十分なデータ数があるとは限らない．領域限定法などで設定する重みの上下限値と同じように，計画問題の最適解はこれらの値に直接影響される．したがって，その与え方には十分注意しなければならないが，とはいっても精度良く推定するには不十分な情報しかないのが一般的であろう．

5. 確率的計画法とDEA

5.1 機会制約条件モデル

データを確率変数とみると，効率値を求めるための線形計画問題は確率的線形計画問題となる．このときの確率変数は正規分布に従うと仮定されることが多い．DEAはノンパラメトリック手法であるが，ここにパラメトリックな確率モデルを入れることになる．しかしこの仮定はデータ構造に対する正規性の仮定であり，DMUの構造やDEAのモデルに対する仮定ではない．確率的計画法では確率的な制約条件の成立に関して2つの考え方がある．1つは制約条件の成立する確率をあるレベル以上にしようというもので，もう1つは制約条件を満たさない量をできるだけ小さくしようとするものである．前者は機会制約条件モデル（chance-constrained model），後者はリコース付き2段階問題（two-stage problem with recourse）として定式化することができる．

Olesenらの機会制約条件モデル[7]では，制約条件 $-v'X+u'Y\leq 0$ が機会制約条件

$$\Pr(-v'x_j+u'y_j\leq 0)\geq\alpha_j,\ j=1, 2, \cdots, n \tag{6}$$

に緩められている．これは生産可能集合の仮定の1つである「入出力データは生産可能集合に含まれる」を「入出力データは確率 α_j 以上で生産可能集合に含まれる」としたものである．入出力データ (x_j, y_j) は正規分布 $N((\bar{x}_j, \bar{y}_j), \Sigma_j)$ に従うと仮定すると，(6)式は

$$-v'\bar{x}_j+u'\bar{y}_j+\Phi^{-1}(\alpha_j)\sqrt{(u, v)'\Sigma_j(u, v)}\leq 0 \tag{7}$$

と表される．ここで Φ は標準正規分布の分布関数である．目的関数にも確率変数が含まれるが，α%点の最大化という考え方により

$$\max \tilde{\theta}_o = u'\bar{y}_o + \Phi^{-1}(\alpha_o)\sqrt{(u, v)'\Sigma_o(u, v)} \tag{8}$$

としている．この最大値 $\tilde{\theta}_o^*$ を機会制約効率値 (chance constrained efficiency score) と呼んでいる．

この問題は凸制約における凸関数の最大化となり，その最適解を求めるのは容易ではない．Olesen らも大域的最適化のためのソフトを用いて機会制約効率値の上限値と下限値を求めるにとどまっている．

5.2 リコースモデル

不確実性をデータの持つ観測誤差としてとらえるならば，制約を満たさない量が最小となるように生産可能集合を構成しようという考え方もできる．これに対応するのが確率的計画法におけるリコースモデル[3]である．リコースモデルでは2段階法が取られ，まず，問題(1)の制約条件を満たさない量をリコース変数 (r_x, r_y) として導入する．

$$\begin{aligned} \theta(x_o + r_x) &= X\lambda + s_x \\ (y_o - r_y) &= Y\lambda - s_y \end{aligned} \tag{9}$$

このときリコース変数の加重和の最小値を考え，その期待値

$$Q(\theta, \lambda, s_x, s_y) = E[\min\{q_x'r_x + q_y'r_y\}] \tag{10}$$

をリコースコスト (recourse cost) と呼び，制約を満たさなかったことに対するペナルティとして目的関数に課す．次に最小化問題を解き，効率値 θ とリコースコストの値 (q とおく) を求める．

通常の確率的計画法でいうリコースは制約条件の左辺と右辺のギャップを指している．そのようにリコースを定義するなら(9)式の最初の式は

$$\theta x_o + r_x = X\lambda + s_x \tag{11}$$

としなければならないが，ここで考えているリコースでは，ばらつきの原因はデータ (x_o, y_o) にあるものとしているため，データの値にリコース変数を加えて制約を満たさない量を補正することとした．したがって，$(x_o + r_x, y_o - r_y)$ が制約を満たす，つまり，生産可能集合に含まれるようにするためにリコース変数をおいている．

リコースを考えることで，ばらつきを考えないときの生産可能集合と比べて，ばらつきに相当する分だけ生産可能集合を拡大させたことになり，それに伴って効率値も下がることになる．

このリコースモデルでも期待値操作を行うために確率分布を特定することが必要となるが，ノンパラメトリックな確率モデルを適用できる場面として，繰り返し測定が行われるときとか予測確率が与えられているときなどが考えられる．繰り返し測定が行われたとき，それぞれの観測値を母集団からのランダムサンプルとして扱うと，これらは等確率で実現した確率変数の値とみることができる．K 個の観測値が (X^k, Y^k), $k = 1, 2, \cdots, K$ と得られたとき，各々の観測値は確率 $p_k = 1/K$ で実現したものと考えられる．何らかの事前情報によりこの出現確率 p_k が与えられる場合にはその値を使えばよい．このときのリコースコストは

$$Q(\theta, \lambda, s_x, s_y) = \min \sum_{k=1}^{K} p_k(q_x'r_x^k + q_y'r_y^k) \tag{12}$$

と表され，その結果に得られる確率的計画問題は

$$\left|\begin{array}{ll} \text{minimize} & \theta - (e's_x + e's_y) \\ & \quad + \sum_{k=1}^{K} p_k(q_x'r_x^k + q_y'r_y^k) \\ \text{subject to} & \theta(x_o^k + r_x^k) = X^k\lambda + s_x \\ & (y_o^k - r_y^k) = Y^k\lambda - s_y \\ & \lambda \geq 0,\ s_x \geq 0,\ s_y \geq 0,\ r_x^k \geq 0,\ r_y^k \geq 0 \\ & k = 1, 2, \cdots, K \end{array}\right. \tag{13}$$

となる．

問題(13)の最初の制約条件が(11)式のように与えられるときには，その最適解は線形計画問題を逐次的に解くL型法[3]と呼ばれる切除平面法によって求められる．しかし問題(13)の最初の制約条件には変数 θ と変数 r_x^k の積がある非線形関数であり，L型法をそのまま用いることはできない．交互に一方の変数を固定して線形計画問題として解くことを繰り返すことで解を求めることができる．

このモデルで得られるものは効率値 θ^* とスラック s_x^*, s_y^* および目的関数の第3項にあるリコース値 q^* である．効率的なDMUは $\theta^* = 1$, $s_x^* = s_y^* = 0$ であるとともに $q^* = 0$ となる．このリコース値はデータのもつばらつきが効率値へ及ぼす影響の度合いを表しているとみることができる．

数値例を用いてその様子を見てみよう．2入力1出力をもつ6つのDMU (A, B, C, D, E, F) があり，入力値がそれぞれ3回観測されたとする．ここで3回の観測値には対応関係があって，3つのデータセット (X^1, Y^1), (X^2, Y^2), (X^3, Y^3) が各々確率1/3で生起するものとしている．図1に単位出力あたりの入力値の散布図を示しているが，簡単のため対応関係の表示は省略している．

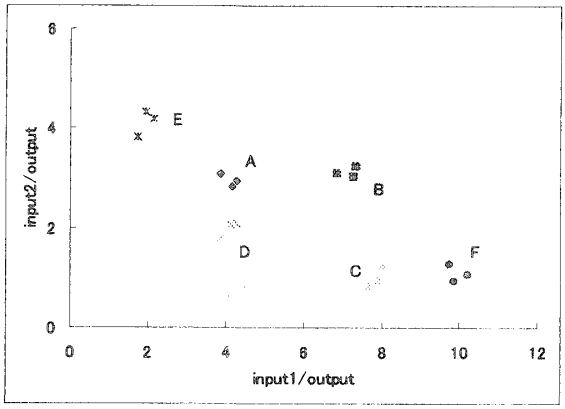

図1　入出力データ

表2　効率性分析の結果

DMU	A	B	C	D	E	F
効率値(平均値)	.859	.638	1	1	1	.936
効率値(リコース)	.782	.528	1	1	1	.756
リコース値	0.35	0.61	0	0	0	0.30

表3　拡張効率値とリコース値

DMU	C	D	E
拡張効率値	1.003	1.141	1.917
リコース値	0.343	0.238	0.192

A，B，Fは非効率的で，Fにはスラックが存在しそうである．C，D，Eは効率的となりそうである．表2に3回の平均値による効率値（第1行）とリコースモデルによる効率値（第2行）およびリコースコスト（第3行）を示す．A，Bは非効率的であり，効率値もばらつきの影響を受けて変化する．Fは効率的フロンティアに近く，確率的な変動によって効率的と評価されることもありうるが，リコースモデルでの効率値はかなり小さくなっている．これはデータのばらつきにより効率性が大きく影響されることが原因と考えられる．C，D，Eはすべて効率的となっており，その度合を見ることはできない．Andersenらの拡張効率値を算出する要領でリコースモデルでの拡張効率値を計算すると表3のようになる．効率値とリコース値から，Cが最もばらつきの影響を受けていることがわかる．

リコース値とばらつきの大きさの関係を見るために，

表4　効率性分析の結果（すべてのばらつきが2倍）

DMU	A	B	C	D	E	F
効率値	.710	.473	.858	1	1	.708
リコース値	0.76	1.31	0.72	0	0	0.51

表5　効率性分析の結果（C，D，Eのばらつきが2倍）

DMU	A	B	C	D	E	F
効率値	.769	.511	.993	.729	1	.723
リコース値	0.42	0.80	0.35	0.42	0	0.44

平均値はそのままでばらつきを2倍にしたときの効率値とリコース値を計算した．表4はすべてのDMUにおいてばらつきの大きさを2倍したときの結果であり，表5は効率的であったC，D，Eのばらつきだけを2倍にしたときの結果である．表2と表4を比べてみると，ばらつきが大きくなったことでCは非効率的となっている．全体的に効率値は下がっており，リコース値は2倍近くになっている．リコース値がばらつきの大きさに応じた指標になっていることがわかる．また，表2と表5を比べてみると，Dの評価が大きく変わっている．Dのばらつきが大きくなっているため，Aとの違いがなくなってきているものと思われる．

これらはいずれも平均値は固定させてばらつきの大きさだけを変化させたものであり，ばらつきを考慮に入れたモデルであるからこそ得られるものである．ひとつのDMUが何回かの入出力値を測定できるときには，従来では平均値を取ることによってばらつきを小さくしてから効率性分析を行うことはできたが，ばらつきの大きさを考慮することはできない．このとき同じ平均値をもっていれば，同じ効率性と評価されてしまう．しかし，同じ平均値をもっていても，ばらつきの大きいデータをもつDMUの効率性は劣るものと見るべきである．複数個の観測データを個々に用いて効率性を評価することにより，データのもつばらつきの大きさを取り入れた効率性分析が可能となる．

また，将来の環境を予測しながらシステム設計などを行うとき，実現する環境によってシステムの効率性は変わる．シナリオ分析で用いられるように将来の環境が予測確率とともに示されている場合，各環境における入出力値は予測確率の大きさにしたがって生起する確率変数と見ることができる．リコースモデルを用いると，複数の環境における効率性をその予測確率に

したがって統合した効率値によって表すことも可能となる．

不確実性が伴うときには複数の観測データを取ることでそのばらつきを捉えることができるが，リコースモデルによりこの複数の観測データからの効率性の評価が可能となる．さらに，機会制約条件モデルなど他の確率的DEAモデルで仮定していたデータに対するパラメトリックな統計モデルが必要なく，DEA本来の特徴であるノンパラメトリックでデータオリエンテッドな評価指標を与えることができる．

6. おわりに

本稿では確率的DEA法に関するいくつかの話題を紹介した．最近の国際会議でもDEAを確率的なデータに適用しようとした報告がいくつか見られるが，簡便法的な内容のものが多く見られる．厳密な統計的な解析をしないままに効率値の信頼区間であると主張しているものもあった．解析的な導出は難しいものも多く，評価値を如何にして導出するかという方が応用上も重要であることは間違いない．理論的な裏づけを与えるとともに，今後多くの適用事例を積み上げていくことが必要であろう．

参考文献

[1] P. Andersen and N. C. Petersen, A procedure for ranking efficient units in data envelopment analysis, Management Sciences, Vol. 39, pp. 1261-1264, 1993

[2] P. W. Bauer, Recent developments in the econometric estimation of frontiers, Journal of Econometrics, Vol. 46, pp. 39-56, 1990

[3] J. Birge and F. Louveaux, Introduction to stochastic programming, Springer Verlag, 1997

[4] W. W. Cooper, L. M. Seiford and K. Tone, Data envelopment analysis: A comprehensive text with models, applications, references and DEA-solver software, Kluwer Academic Publishers, 2000

[5] H. Morita and L. M. Seiford, Characteristics on stochastic DEA efficiency—Reliability and probability being efficient—, Journal of Operations Research Society of Japan, Vol. 42, No. 4, pp. 389-404, 1999

[6] 乾口雅弘, 谷野哲三, ファジィ入出力データによる可能性DEA, 日本ファジィ学会誌, Vol. 11, No. 3, pp. 472-481, 1999

[7] O. Olesen and N. C. Petersen, Chance-constrained efficiency evaluation, Management Science, Vol. 41, pp. 442-457, 1995

[8] L. M. Seiford, Data envelopment analysis: The evaluation of the state of the art (1978-1995), Journal of Productivity Analysis, Vol. 7, pp. 99-138, 1996

[9] 刀根薫, 経営効率性の測定と改善, 日本科学技術連盟, 1995

[10] 上田徹, 包絡分析法DEAとファジィDEA, 日本ファジィ学会誌, Vol. 10, No. 2, pp. 193-199, 1998