平成 23 年度　　卒 業 研 究 論 文 要 旨

教育研究分野　情報数理学　　2120053　早川 潤

カルバック情報量を用いたボックス・コックス変換法の導出

1　はじめに

Box と Cox はパラメータを介し，正規分布への変換を行うボックス・コックス変換を考案した [1]．Box らはパラメータの推定に最尤法を用いられている．最尤法はサンプルサイズがある程度大きいことを求めており，サンプルサイズが小さい場合のパラメータ推定が重要となる．本論文では Arizono ら [2] が提案した正規性検定統計量に基づく新たなパラメータ推定法を提案した．

2　ボックス・コックス変換

ボックス・コックス変換は正のデータに対して未知パラメータ λ を介して正規分布に近似するデータに変換させる変換で

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log y & (\lambda = 0) \end{cases} \tag{1}$$

と定義される．Box らはパラメータの推定に最尤法を用いた．最尤法とは,“ データがどのパラメータのもとで生じたのが尤もらしいか ”を示す尤度関数に対し，これを最大とするパラメータを推定量とする方法である．Box らは，尤度関数に対数をとった対数尤度関数 (対数変換は単調増加変換なので尤度関数を最大とするパラメータと対数尤度関数を最大とするパラメータは同値となる．)

$$l(\lambda) = -\frac{n}{2}\log\sigma^2 + (\lambda-1)\sum_{i=1}^{n}\log y_i \tag{2}$$

を最大にする λ を変換パラメータとした．

3　ボックス・コックス変換パラメータ新推定法

確率密度関数 $f(x)$，$g(x)$ をそれぞれもつ 2 つの分布の距離尺度であるカルバック情報量 $I(g;f)$ は

$$I(g:f) = \int_{-\infty}^{\infty} g(x)\log\frac{g(x)}{f(x)}dx \tag{3}$$

と定義される．$I(g;f)$ は非負の値をとり，値が 0 のとき 2 つの分布は一致する．Arizono らはこの情報量に基づく正規性検定統計量を提案した．Arizono らの情報量指標は“ 正規性の良さ ”の尺度とも考えることができ，これに基づいた新しいボックス・コックス変換法が確立できるので，本論文では，提案統計量

$$BKL_{mn} = \frac{n}{2mse^{1/2}}\left\{\prod_{i=1}^{n} y_{i+m}^{(\lambda)} - y_{i-m}^{(\lambda)}\right\}^{1/n} \tag{4}$$

を導出した．ただし，$0 \leq BKL_{mn} \leq \sqrt{2\pi}$ であり，$BKL_{mn} \xrightarrow{P} \sqrt{2\pi}$ となる．BKL_{mn} が大きいほど正規分布に近いとみなせるので，BKL_{mn} を最大にする λ を探索し，λ を求めている．なお，本論文では式 (2) 及び式 (4) を最大にする λ の探索法として黄金分割法を採用した．

4　数値検証

各サンプルサイズ n に対し，プロビット変換を施した正規分布 $N(5,1)$ に従うサンプルを 10,000 組作成する．正規分布に対してボックス・コックス変換を行うと $\lambda = 1$ が変換パラメータの真値であることがわかっている．各サンプルについて式 (2)，式 (4) が最大となるパラメータ λ を探索し，推定値 λ の期待値などを求めた. この結果を表 1・2 に示す. また，それぞれの変換後のデータについてシャピロ・ウィルク検定統計量 W(W は 1 以下の値をとり，値が大きいほど正規分布に近い．) を求め,“ 提案手法での $W(W_k$ とする)> 既存手法での $W(W_l$ とする) ”となった数をカウントした．その結果を表 3 に示す．すなわち，カウント数 5,000 を基準とし，5,000 より大きければ提案手法による変換が有効であると考えられる．

表 1: 最尤法でのボックス・コックス変換パラメータ

n	期待値	標準偏差	平均二乗誤差	変動係数
8	0.710	1.993	4.054	2.806
10	0.762	1.692	2.918	2.219
15	0.813	1.260	1.623	1.550
20	0.861	1.031	1.082	1.197
30	0.902	0.801	0.651	0.887
40	0.920	0.674	0.461	0.733

表 2: 提案法でのボックス・コックス変換パラメータ

n	期待値	標準偏差	平均二乗誤差	変動係数
8	0.790	2.175	4.776	2.753
10	0.824	1.813	3.319	2.202
15	0.848	1.316	1.756	1.552
20	0.885	1.062	1.142	1.200
30	0.916	0.815	0.671	0.890
40	0.929	0.682	0.470	0.734

表 3: $W_k > W_l$ のカウント数

n	$W_k > W_l$	n	$W_k > W_l$
8	6970	20	6725
10	6985	30	6049
15	6878	40	5379

推定値の期待値に注目すると提案手法のほうが既存手法よりも真値に近く，データ数が小さい場合にはより顕著である．さらに，変動係数に注目すると，データ数が小さい場合には提案手法が既存手法より良い結果が得られた．また，表 3 の結果よりサンプルサイズが小さい場合に対して提案手法が有効であることがわかった．以上の結果よりサンプルサイズが大きいときには既存手法の適用が望ましく，小さいときには提案手法の適用が望ましいと思われる．

5　おわりに

数値検証の結果，サンプルサイズが小さい場合に関して提案手法の有効性を示すことができた．よって，サンプルサイズがあまり大きくないことが通常である品質工学の分野，特にデータの正規性を要求しているタグチ・メソッドなどへの適用が考えられる．

参考文献

[1] G.E.P.Box and D.R.Cox：“An analysis of transformations, ” Journal of the Royal Statistical Society. Series B， Vol. 26，No.2，pp.216–252，1964.

[2] I.Arizono and H.Ohta：“A test for normality based on Kullback-Leibler information, ” The American Statistician，Vol. 43，No. 1，pp.20–22，1989.