（7）統計的グラフィックスとその応用

多変量解析の感度分析
− influential subsets の探索とそのグラフ表現

文　勝浩 (岡山大学・大学院)
垂水共之 (岡山大学・教養部)
田中　豊 (岡山大学・教養部)

回帰分析において、複数個の影響の大きい観測値 (influential subsets of observations) が存在する場合、通常の診断法では、それらの検出ができないことがあり、マスク効果と呼ばれている。マスク効果を防ぐため Rousseeuw(1984) は最小自乗法の $min\frac{1}{n}\sum r^2$ での $\sum r^2$ の平均をとる代わりによりロバストな $median$ をいれて $min\ med\ r^2$ とする推定法とそれにもとづく診断法を提案した。彼の提案した診断尺度 RD_i(Resistant Diagnostic) は通常の方法ではマスク効果の生じる場合にも外れ値の検出に非常によい結果を出している。本報告では主成分分析 (PCA) において influential subsets を探索する問題を扱うが、Rousseeuw の考え方にならってロバストな推定法–Campbell(1980) の提案したロバスト主成分分析 (PCA)–を用いる診断法を提案し、その方法を数値例に適用してみる。

PCA において 1 つずつの個体の影響を考える場合、ある分布関数 F に対して、

$$F \to (1-\varepsilon)F + \varepsilon\delta_x \quad (\delta_x : \text{cdf with a unit point mass at } x) \tag{2.1}$$

のような摂動を考える。このとき、対応する平均、分散行列の変化は、ε の 1 次の項までを取り上げれば、

$$\mu \to \mu + \varepsilon\mu^{(1)}, \quad \Sigma \to \Sigma + \varepsilon\Sigma^{(1)} \tag{2.2}$$

$$\mu^{(1)} = x - \mu, \quad \Sigma^{(1)} = (x-\mu)(x-\mu)^T - \Sigma \tag{2.3}$$

で与えられ、これに対応して、主成分分析においての固有値 θ_s と固有ベクトル v_s は

$$\theta_s \to \theta_s + \varepsilon\theta_s^{(1)}, \quad v_s \to v_s + \varepsilon v_s^{(1)} \tag{2.4}$$

$$\theta_s^{(1)} = v_s^T((x-\mu)(x-\mu)^T - \Sigma)v_s \tag{2.5}$$

$$v_s^{(1)} = \sum_{r\neq s}(\theta_s - \theta_r)^{-1}[v_s^T((x-\mu)(x-\mu)^T - \Sigma)v_r]v_r, \quad 1 \le r \neq s \le p \tag{2.6}$$

と変化し、$\theta_s^{(1)}$と$v_s^{(1)}$は分散行列の一次微係数$\Sigma^{(1)}$の線形関数の形で表せることがわかる。

複数個の個体の影響を考えるためには、分布関数 F に対して、

$$F \to (1-\varepsilon)F + \varepsilon G, \quad G = \frac{1}{k}\sum_{i \in S} \delta_{x_i} \tag{2.7}$$

(S : $i_1, i_2, \cdots, i_k$ の index subset 、δ_{x_i} : cdf with a unit point mass at x_i) のような摂動を考える。対応する平均、分散行列の変化は、

$$\mu \to \mu + \varepsilon\mu_S^{(1)}, \quad \Sigma \to \Sigma + \varepsilon\Sigma_S^{(1)} \tag{2.8}$$

$$\mu_S^{(1)} = \bar{x} - \mu = \frac{1}{k}\sum_{i \in S} \mu_i^{(1)} \tag{2.9}$$

$$\Sigma_S^{(1)} = \frac{1}{k}\sum_{i \in S}(x_i - \mu)(x_i - \mu)^T - \Sigma = \frac{1}{k}\sum_{i \in S}\Sigma_i^{(1)} \tag{2.10}$$

のようになり、複数個の個体の集合 S に対する、平均ベクトル及び分散共分散行列の一般化された影響関数の値は、それぞれの点について求められる通常の影響関数の平均で与えられることがわかる。θ_sと v_sの影響関数が、Σの影響関数の 1 次関数の形で与えられることから、個体の集合 S に対する、θ_sと v_sの一般化された影響関数もμやΣの場合と同じく、

$$\theta_S^{(1)} = \frac{1}{k}\sum_{i \in S}\theta_i^{(1)}, \quad v_S^{(1)} = \frac{1}{k}\sum_{i \in S} v_i^{(1)} \tag{2.11}$$

により得られる。

標本分散共分散行列 Vあるいは標本相関行列 Rの固有値、固有ベクトルにもとづく主成分分析の主要な結果は、少数個の外れ値の混入により大きく変化する可能性のあることが知られている。このような外れ値の影響をなるべく受けずに、大部分のデータの特徴を反映するロバストな方法が Campbell(1980) や Delvin et al.(1981) などによって研究されているが、ここでは Campbell(1980) の提案した二つの方法に基づくロバストな主成分分析を用いる。

非対称類似性データのベクトル場表現

九州大学理学部　仁木直人
九州大学総理工　宿久　洋

心理学・行動学・社会学等においては、しばしば非対称な類似性(あるいは疑似距離) データを取り扱う必要がある。 この主題については多くの研究がなされているが、Niki (1990) 提案による力学的解釈は、自然なモデルの採用による理解の容易さと発展性において、他の方法にない特長を持っている。この方法の基本となるアイデアは、「各対象をそれぞれ一点に対応づけ、対称成分からその位置ベクトルを、また歪対称成分から運動（あるいは作用）ベクトルを決定する」というものである。

宿久・仁木 (1991) では、その発展のひとつとして、「ベクトルの湧出し」を測ることによる系の分散・収縮傾向の抽出について議論するとともに、各種の行動学上意味のある量が、系全体ばかりでなく、その任意の部分（跳び離れていてもよい）についても計算できることを指摘した。

本研究集会では、結果全体をひとつのベクトル場として捉え、そのスカラー・ポテンシャルを推定することにより、系の全体像をより見易くする工夫について、主に議論を行った。ただし、ポテンシャルに関しては、実用的な結果表示が可能である２次元ベクトル場についてのみ考えることにしている。

系に属する点を全て含む２次元矩形領域を定め、その上で定義された「滑らか」なスカラー・ポテンシャルの各格子点での値を推定

することが目的である。

事前の解析 (Niki, 1990) により与えられた歪対称成分を表す各ベクトルがスカラー・ポテンシャルの勾配 (gradient) とベクトル・ポテンシャルの回転 (rotation) の和で書けると考える。前者をその点を含む単位格子上の立体四辺形の「勾配」で近似することとし、後者をランダム・ベクトルのように見なすとともに、スカラー・ポテンシャルの滑らかさに関して、その2次元の「2階差分」が0に近いという条件を課す。このような構造を仮定して、各点それぞれにおける勾配の近似の良さと全体にわたる平滑条件の満足度を同時に高めることを考えると、赤池の Bayesian モデル (Akaike, 1980) の一形態が導かれる。

近似の良さと条件の満足度は互いに相反する面があるが、その調整に赤池の Bayesian 情報量基準 (ABIC) を用いれば、十分に実用的なスカラー・ポテンシャルの推定が可能であることを研究集会では示した。なお、「勾配」や「2階差分」の定義、あるいは、いかなるランダム・ベクトルとするか等には任意性があり、この考え方に基づく多くのモデルが構築可能である。その一例を研究集会の報告集に掲載している (Niki and Yadohisa, 1992)。

参考文献

Akaike, H. (1980). Likelihood and the Bayes procedure, *Bayesian Statistics* (Bernardo, J. M. et al., eds.), University Press, Valencia, Spain, 143-166.

Niki, N. (1990). Kinetic interpretation of asymmetric proximities, *Statistical Methods and Data Analysis* (Niki, N., ed.), Scientist, Tokyo, 29-32.

宿久 洋, 仁木 直人 (1991). 非対称類似性のベクトル表現, 第59回日本統計学会講演報告集, 72-73.

Niki, N. and Yadohisa, H. (1992). Vector field representation of asymmetric proximities, 「統計グラフィックスとその応用」シンポジウム報告集 (編集: 白旗慎吾).

外的基準のないデータに対する曲線あてはめについて

北大・工学部・情報科学　水田正弘

1．はじめに

各種データに滑らかな曲線を当てはめることは、データの有する情報を視覚的に表現する一つの有効な方法である。曲線当てはめの状況を二つに大別することができる。

第1の状況は多変量データにおいて、一つの目的変量と複数の説明変量から構成されている場合である。このときには、目的変量を説明変量の関数として表すことが基本的な事項である。古典的には回帰分析がこの目的に合致している。さらに、各種の平滑化の手法が開発されている。lowess(Cleveland(1979))や、平滑化スプライン、移動平均法なども含まれる。1989年10月の応用統計学会のシンポジウムでは、「平滑化とその周辺」がテーマとして開催された。

第2の状況は、多変量データの各変量を目的変量と説明変量に分けることが出来ない場合、または複数の説明変量の相互関係を検討する場合である。本報告では、これらを簡単に「外的基準がないデータ」と呼ぶことにする。このようなデータに対して、密度関数の推定を試みることが考えられる。得られた密度関数を3Dプロットなどで表示することは、一つの統計的グラフィックスと言える。また、データに半順序付けを行う研究もなされている（松原・余田・後藤(1989)）。データがある曲線上に存在する可能性がある場合、または曲線上に存在することが期待されている場合も少なくない。

本報告では、外的基準のないデータに対して曲線を当てはめる方法をいくつか紹介する。

2．曲線当てはめの方法

以下では、説明の都合上、2変量データ（x,y）の場合について述べるが、計算時間や記憶容量の問題を無視すれば、多変量でも同様に適用できる。

2－1．一般化主成分分析

陰関数による曲線当てはめ、すなわち、「代数曲線」による曲線当てはめとしては、Gnanadesikan & Wilk(1969)による一般化主成分分析が初期の研究としてある。これは、データの次元pを${}_{p+r}C_r-1$（$r=2,\cdots$）に上げ、その空間で通常の主成分分析を行う手法である。r=1の場合には、通常の主成分分析と一致する。Kendall(1975)による教科書的な本にも同様な手法が記述されている。

2－2．最短2乗距離を有する2次曲線

一般化主成分分析およびその変形した手法では、「散らばりが小さい」合成変量 z を利用しているが、データ点と曲線との（集合としての）距離の2乗和が最小になっているわけではない。すなわち、これらの手法によって、データに曲線を当てはめることは可能であるが、得られた曲線は必ずしも垂直な距離の2乗和を最小にしていない。

そこで、水田(1987)は、2次元データに対して垂直な距離の2乗和が最小となる2次曲線を求めるアルゴリズムを提案した。

2－3．Principal Curves

陰関数ではなく、媒介変数（パラメータ）を利用した曲線（x(t),y(t)）による当てはめを考えることができる。この種の興味深い手法としてHastie & Stuetzle(1989:JASA) による Principal Curves がある。

2－4．Shape Polynomials と Kernel function

Shape Polynomials は Taubin(1989)により、パターンマッチングのために開発された手法である。

西邑(1991)は、Shape Polynomials を利用した、曲線当てはめおよびクラスタリングを提案している。

2－5．その他の方法

パターン認識などの分野では、上記以外の手法が使われている。曲線（折れ線）当てはめの手法としては区分的直線近似法などがあり、さらに関連する手法として線追跡、フーリエ記述子など数多く使われている

Principal Curves を拡張した Principal Sets を利用して、枝分かれのある曲線によってデータを当てはめる方法も提案されている(水田・馬場(1990B))。これは初期曲線群を与えて、繰り返しによって曲線群を求める方法である。

3．数量化Ⅲ類への応用

外的基準のない場合の曲線当てはめの実際的な応用として現在、数量化Ⅲ類の結果の解釈への応用を進めている。

アイテム・カテゴリ間に何らかの順序が付く場合に数量化Ⅲ類を適用すると、多次元空間内の曲線（2次

曲線、3次曲線、…）の近くにアイテム・カテゴリあるいは個体が付置されることが多い。また、数量化Ⅲ類を n-way に拡張した手法においても、同様な曲線が得られることがある(岩坪(1987))。そこで、この曲線に沿った順序でアイテムやカテゴリを並び替えることが考えられる。

4．ソフトウェアとアルゴリズム

コンピュータ指向な手法では、その実際的なアルゴリズムによって有効性が大きく変わる。適切な言語・システムで手法を実現することが重要である。

◎一般化主成分分析

これは、通常の主成分分析のプログラムがあれば、入力データを変換すればよい。ただし、通常の主成分分析とは異なり、最小固有値に対応する主成分が曲線当てはめとしては意味がある。また、Gnanadesikan(1977)は、相関行列を使うことを推奨しているが、多くの数値計算例では分散共分散行列を使う方が妥当と思われる曲線が得られる。さらに前述したように、座標の平行移動に対して不変ではないので、あらかじめデータの各変量の平均を0にしておいた方がよい。

◎Principal Curves

HastieによるSのPrincipal Cureves が近々、公表される予定である。lowess も PROJECTION STEP も単純に計算すると時間がかかるので、このプログラムがどのようなアルゴリズムになっているか、特にPROJECTION STEP における工夫がなされているかに興味がある。

◎Shape Polynomials

これは、データを変換した後、分散共分散行列を求め、逆行列と行列の演算により簡単に得られる。しかし、むしろ、Shape Polynomials を３Ｄで表示したり、細線化するためのプログラムの方が大変である。

5．おわりに

外的基準のないデータに対する曲線当てはめの研究は外的基準のあるデータの場合より少ない。その理由として、データの一次変換による結果の不変性を保つことが困難なこと、一般に計算が困難であることが考えられる。

計算時間については、アルゴリズムの工夫およびコンピュータの進歩によってかなり回避できる。しかし、座標系の一次変換については、慎重に考察する必要がある。すなわち、外的基準のないデータに対して、安易にある変量の尺度を変えることは危険であり、また各変量の尺度を無視したければ、各変量を標準化してから解析すべきである。したがって、座標系の変換として、回転（直交変換）および平行移動だけを考え、それらの変換に関する不変性を検討すれば十分であろう。

参考文献

[1]Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons. （和訳　丘本　正・磯貝恭史(1979). 統計的多変量データ解析、 日科技連）

[2]Hastie, T. & Stuetzle, W. (1989). Principal Curves. J. Amer. Statist. Assoc., 84, pp. 502-516.

[3]Mizuta, M. (1983). Generalized Principal Components Analysis Invariant under Rotations of a Coordinate System. J. Japan Statist. Soc., 14, pp. 1-9.

[4]Taubin, G., Bolle, R. M. & Cooper, D. B. (1989). Representing and Comparing Shapes Using Shape Polynomials, Proc. IEEE Computer Soc. Conference on Computer Vision and Pattern Recognition, pp. 510-516.

[5]岩坪秀一(1987). 数量化法の基礎. 朝倉書店.

[6]西邑嘉人(1991). Shape polynomials によるデータ解析法に関する研究、平成２年度北海道大学大学院工学研究科修士論文.

[7]松原義弘・余田明夫・後藤昌司(1989). 多変量データの順序付けとその応用. 第57回日本統計学会講演報告集. pp. 228-229.

[8]水田正弘(1987). データ点との最小２乗距離を有する２次曲線の導出法. 第15回日本行動計量学会発表論文抄録集. pp. B4-2-1～B42-2.

[9]水田正弘(1988). 一般化主成分分析法とその改良について. 北海道大学工学部研究報告. 第141号. pp. 209-215.

[10]水田正弘・馬場康維(1990A). Principal Curvesについて. 第58回日本統計学会講演報告集. pp. 244-246.

[11]水田正弘・馬場康維(1990B). 枝分かれがある場合の曲線あてはめ. 第7回分類の理論と応用に関する研究会研究報告予稿集. pp. 28-31.

[12]水田正弘(1991). Shape Polynomialsについて. 第59回日本統計学会講演報告集. pp. 210-211.

正規性への変換

磯貝恭史（大阪大学・教養部）

§．序

適当な条件を満たす任意の連続分布 $F_\theta(x)$ を、正規分布 $\Phi(z)$ に変換することを論じる。変換は、$z=T(x)=\Phi^{-1}(F_\theta(x))$ なる関数 $T=\Phi^{-1}\cdot F_\theta$ で与えられるが、T について解析的に調べることは難しい。

ここでは、T の性質を、Kaskey 等(1980)らのアイデアに従って２階常微分方程式を用いて、それを数値的に解き、図示表現することにより考察する。

§１．数値解法

$F_\theta(x)$ を区間［A，B］上で定義されている任意の連続分布とし、その密度関数 $f_\theta(x)$ は、x について１回連続微分可能とする。さらに、端点 A，B は母数 θ に依存しないと仮定する。

$\Phi(z)$ を標準正規分布の分布関数とすれば、変換 $z=T(x)=\Phi^{-1}(F_\theta(x))$ は２階常微分方程式

$$\frac{d^2z}{dx^2}=z\left(\frac{dz}{dx}\right)^2+\left[\frac{\partial}{\partial x}\log f_\theta(x)\right]\frac{dz}{dx}$$

の解になっている。この解を初期条件

$$z(x_{0.5})=0,\quad z'(x_{0.5})=\sqrt{2\pi}\,f_\theta(x_{0.5})$$

の下で、４次のルンゲ・クッタ法で数値解法により求める。ここで x_α は

$$F_\theta(x_\alpha)=\alpha\qquad(0<\alpha<1)$$

で定義され、$x_{0.5}$ は $\alpha=0.5$ に対応する F_θ の中央値である。

§2．変換 $z=T(x)$，およびそれの近似

ルンゲ・クッタ法により、変換 $z=T(x)$ の関数形は $(x, z(x))$ のグラフによって図示表現される。

$z=T(x)$ の関数形を探すために、Tarter 等(1972)は

$$R(x)=\phi(\Phi^{-1}F_{\theta}(x))/f_{\theta}(x)$$

のプロットを提案しているが、これは、$(x, 1/z'(x))$ のプロットと同値である。

§3．変換 $z=T(x)$ の診断

Efron (1981) は、特に変換 $z=T(x)$ が次の形

$$z(x)=(g(x)-\nu_{\theta})/\sigma_{\theta}$$

で書け、$g(x)$ が母数 θ に依存しないようなものが存在するかどうかを調べる道具を提案した。

変換 $z=T(x)$ が、θ について微分可能であると仮定して、その微分係数 $\frac{\partial}{\partial\theta}z$ を z_{θ} とした時、Efron (1982) の診断法は

$$(z, z_{\theta}/z_{\theta}(x_{0.5}))$$

のプロットである。これは、$F_{\theta}(x)$ が対称な場合には用いることができない。そこで、少し修正して、

$$(z, z_{\theta}-z_{\theta}(x_{0.5}))$$

を用いることを提案する。さらに、新しい診断法として、

$$(z, \frac{\partial}{\partial\theta}z'/z')$$

のプロットが提案できる。

§4．引用文献

Efron (1982). A.S., 10, pp.323～339.

Kaskey et al.(1980). Handbook of Statistics, Vol.1, pp.321～341.

Tarter et al.(1972). Technometrics, 14, pp.735～744.

糖尿病ファジィ診断

－耐糖能ダイナミックグラフからの試み－

有田清三郎（川崎医科大学数学）・米田正也（川崎医科大学内分泌内科）

１．はじめに

糖尿病(ＤＭ)は成人病の中でも、大きな役割を占め、我が国では３００万人以上の患者がいると予測され、現在まさに飽食・過食の時代の文明病ともいわれている。

糖尿病診断では７５ｇのグルコースを経口摂取させ、糖負荷による糖代謝能力(耐糖能)を中心に判定している。例えば血糖２時間値が２００ｍｇ/ｄｌ以上ならば糖尿病、２００ｍｇ/ｄｌ未満ならば糖尿病以外(境界型、正常型)と判定している。この判定法に従えば、２時間値が２００ｍｇ/ｄｌは糖尿病、１９９ｍｇ/ｄｌは非糖尿病と判定されることになるが、糖負荷テストによる個人の反応パターンは連続性を持ち、かつ多様性を持っているものである。日本糖尿病学会がＷＨＯを参考に作成した糖尿病診断基準は、観察時刻の血糖値がある値を越えるか、越えないかの判定(クリスプ判定)で糖尿病型、境界型、正常型の３つに分類している。この分類方法には、次のような問題点がある。

１）血糖２時間値は連続的な変化量であり、この血糖値を人工的なクリスプ判定で３群に分けることに無理がある。

２）個々の耐糖能には多様性が存在するため血糖値だけの単一尺度で評価するのは適当ではない。

今回、我々は糖尿病の病態をＦｕｚｚｙ ｓｅｔと考え、ファジィ理論を用いた糖尿病診断の試みを行ったので報告する。

２．耐糖能のための新しい表現法－ダイナミックグラフ

我々は、血糖値に加えて通常の経口糖負荷テストで得られたインスリン分泌量のデータを基に、個々の耐糖能の反応パターンを血糖値とインスリンの両軸で表現した。我々は、このグラフを耐糖能のダイナミックグラフと名付けた。

このグラフは血糖２時間だけにこだわらず、血糖値とインスリンが時間経過

と共にどのように変化していくかを動的に描いたものである。このグラフで、糖尿病型は高血糖値を反映して横長のグラフを示すのに対して、境界型では糖尿病型に近いものから、インスリンの分泌が高い縦長のものまで、また縦長のグラフから正常型近くのものまでと実に多種多様な反応パターンを示す。従来から糖尿病における境界線でははっきり分けられないことが指摘されていたが、これを明瞭にした。このダイナミックグラフをもとに、糖負荷テストの血糖値とインスリンの両方のデータを用いて、「血糖値がやや高く、インスリンの分泌量がきわめて低い」などのファジィ情報としてとらえ、ファジィ理論を使って糖尿病診断ロジックを開発した。この診断システムは、コンピュータのディスプレー上にダイナミックグラフとファジィ理論による結果を帯状グラフで糖尿病型、境界型、正常型のどのあたりかの位置が表示される。

またこのダイナミックグラフによって境界型ではインスリン分泌が多いものから少ないものまでの多様性が示された。このグラフはインスリン投与、食事療法などの治療経過中に個々の耐糖能がどのようなパターンを経るかがわかり、治療効果判定に有益な情報を与えるであろう。

［参考文献］

1) WHO Diabetes Mellitus. Report of a WHO study group. Technical Report Series 727, WHO. Geneva, 1985.

2) Jarret, R. J., Keen, H., Fuller, J. H., and McCartney, M.: Worsening to diabetes in men with impaired glucose tolerance ("borderline diabetes"). Diabetologia 16: 25, 1979.

3) Yoneda, M., Arita, S., Nishida, S. and Horino, M: New index for the diagnosis of diabetes. Peptide Hormones in Pancreas, Biomedical Res. Fund. Tokyo: 96-101, 1990.

4) Arita, S., Yoneda, M. and Hori, Y.: Diagnostic system for diabetes mellitus based on the response of glucose tolerance test using a fuzzy inference. Proc. of the International Congress of Biomedical Fuzzy Systems & the Third Annual Meeting of Biomedical Fuzzy Systems Association: 67-70, 1991.

時系列データ解析の予備解析のためのモデル判別について。

大分大学工学部　永井武昭

時系列データの解析システム構築において、まず、解析対象となる時系列のプロフィールと潜在構造について、明確な理解を得るための予備解析としての前処理プロセスが必要である。その手段として、通常ルーチーン的に計算する基本統計量、平均値関数、分散関数および自己相関関数やスペクトル密度に現われる時系列の潜在構造特性を利用し、それによって、簡単な統計理論と（ある種の専門家的な）直感や経験に基づいた時系列モデルの判定方式を構築することとした。この判定にしたがって、前もって組み込まれている適切な変換を対象時系列データに施すことで、最終的には、より高度な解析を必要とするＡＲＭＡモデル当てはめに耐えられる正規定常系列データを得ようとしているのである。この一連の過程から成る予備解析としての前処理モジュールについて、その基本的なアイデアと具体的な処理の流れと、いくつかの実施例について紹介する。この前処理段階の大きな目的は、先ず第一に解析者が目下取り扱っている時系列データの統計的構造とそのプロフィールについて、できるだけ明確な理解を得ることと、第二に、これに続く次のステージでのＡＲＭＡモデル当てはめの準備として、解析対象時系列データを正規定常化データに変換しておくことにある。その第２ステージはこれら前処理によって正規定常化された、変換時系列データにたいし精巧なパラメータモデル（ＡＲＭＡモデル）を当てはめることである。そして最終ステージではこれらのパラメータモデルを用いた制御、予測、あるいは、時系列の潜在構造についての解釈等への実際的な応用を行うのである。

このような予備解析システムを前提とした時系列データの予測システムについては、平成３年２月の大分統計談話会で、すでに発表を行なっている。そこでは、データの前処理について時系列データグラフ、コレログラムおよびスペクトル図等の視察によってモデルの判断を

行い、かつ適切な変換を選択する、という、一種の時系列データ解析の経験者向きの、ad hocな手法を紹介したのであったが、今回の方式はモデル判定のためのルール構築に、比較的簡単な統計理論と、今までの時系列データ解析で積んできた経験や直感をできるだけ生かし、これらを統計解析ソフト"S"を用いて関数化する事とした。その判定関数には時系列データの区分時間上の変動挙動とその根底にある時系列モデルの構造的特徴とを数値化し、それらに基づいて判定を行なうようにした。その実行は、まず、判定結果の表示、その判定に至った理由、そして取るべき変換処置(とその実行)とが指示できるように工夫している。

今回紹介する判定方式は、時系列データ解析の予備解析システムの一環として構成した方式であり、時系列データ全般に亘っての包括的なものとして位置づけているので、できるだけ多種多様な時系列データを検証的に用いて実施を試みることとした。それで、諸種の文献に現れた時系列データや、統計データ集から収集した実際データについて、どの程度、この判別ルールが有効に働くか、いくつか実験を行なったので、その結果をまとめて「時系列モデル判定属性指標数値表」に提示した。

もちろん、比較的的確に判定を行なうことはできたのであるが、ランダムトレンドモデルのように、局所的にはトレンドモデルのようであり、しかし大局的には彷徨性をもつ準定常系列であるデータ(化学プラント制御数値データ等)については二者択一的判定は困難で誤判別となった。しかし、ある意味ではどちらに判定してもいくぶんかは正しい判定になっているのであり、完全な正確さもないが、全くの間違いともいえない。時系列データの前処理においては、ややもすると、あまり重要でないトレンド成分や、ノイズ成分の陰に隠れているが、内容的に重要な情報をもつ(成分パワーは小さい)特性を有効に取り出すためには、的確なモデルに関する認識とそのモデルに施す変換関数の選択、そしてそれらを施す順序の決定とが解析目的にとって重要であると解った。

データ省察用グラフィクスの開発

塩野義製薬(株) 解析センター 余田明夫
松原義弘
後藤昌司

データ省察は統計的データ解析(SDA)過程の個々の構成要件の妥当性を保証する観点から重要である(後藤他, 1979). 実地では, そういったデータ省察に卑近なグラフィカル手法が汎用されている. ここに, 卑近なデータ省察用グラフィカル手法とは, 手近でたやすく用いることができ, あまり高尚でないありふれた方法で, 具体的にはヒストグラム, 確率プロット, 散布図, ボックス・プロットなどを指す(余田 他, 1990).

我々の周辺で, 最近までの約10年間のデータ解析で用いたグラフィカル手法の使用頻度の調査結果では, 実地に適用されたグラフィカル手法の7割以上が卑近な方法である. これらのうち, ほぼ半数がデータ省察用グラフィカル手法で占められている. 因に, 我々の「計算環境」には約450種のSDA標準プログラムが整備されている(後藤, 1991). 卑近なグラフィカル手法のうちでは, ボックス・ウィスカー・プロット(箱ひげ図)やそれらの変法の使用頻度が高い. とくに, スキーマティック・プロット, スキーマティック・プロットと散布図を組み合わせてデータの周辺分布と同時分布の視察を可能にするスライディング・スクエア・プロット, データの密度情報を箱の側面に盛り込むヒスト・プロットとヴェイス・プロットなどが特異的である. このとき, 実地での使用経験から, 次の適用上の「経験則」が浮かびあがる.

①曲線よりも傾きのある直線, 傾きのある直線よりも水平線の順が視覚的に(とくに比較を目標とするとき)理解しやすい.

②標本サイズ(n)が大きい場合に個体の表示プロットを適用すると, 生産的知見を得にくいことが多い. 例えば, 散布図やAndrewsプロットなどで注意がいる.

③観測値に重複が多い場合の散布図などでは「ジッタリング」が有用である. 「ジッタリング」とは, 例えばカテゴリカルな観測値に-0.25から0.25までの区間の乱数を加えるといったことを指す. また, カテゴリカル変数の散布図や2値応答のあてはめ結果の表示で注意がいる.

さらに, 個々の諸法については以下の点に留意したい.

ヒストグラム：①相対的に大きな標本サイズの場合に適している. ②級区間の選択を慎重に行うことが必要である. 適切でない級区間を選択すれば, 観測値の本来の分布の形状が正しく表現されないこともありうる.

ボックス・プロット：①相対的に小さな標本サイズの場合に適している. ②市販の統計ソフトウェア・パッケージに組み込まれているものでは, 四分位点の定義が異なっている場合もある. ③分布の2峰性が示唆されない場合もある. ④比較に用いると, 誤解を生じることもある.

確率プロット：①相対的に大きな標本サイズの場合に適している. ②分布形状を直観的に把握しにくい. ③代表的な分布についてQ-Qプロットの形式を理論的あるいは経験的に把握しておくことが望ましい. Q-Qプロットの典型例のみを豊富に収集した成書もある(Fowlkes, 1987).

散布図・散布図行列：①相対的に小さな標本サイズの場合に適している．②変量の個数(p)が大きい場合に，散布図行列を適用してもデータのくせを発見することに寄与せず，情報の見落しが増すことも多い．③変量の組み合わせで表示する場合に，その表示する個数は${}_pC_k$になり，視察が困難になる．例えば，$p=10$，$k=5$の場合に，${}_{10}C_5=252$個の散布図行列が得られる．

とくに，「一般化Q-Qプロット」(Bacon-Shone & Fung, 1987)は単一変量データや多変量データのいずれにも適用でき，しかも1個ないし複数個の外れ値を視覚的に検出することができる．ただし，「一般化Q-Qプロット」には，一見して外れ値の個数を判断しにくい欠点がある．このとき，このプロットの中に3本の密度の異なる点線で回帰直線のあてはめた結果を表示することで，この欠点を補うことができる．すなわち，最も密な直線はすべての点，中程度に密な直線は最高の1点のみを除く他のすべての点，疎な直線は最高点とその次の点の2点を除く他のすべての点に回帰直線をあてはめた結果に対応させる．例えば，外れ値の個数を4個以下と想定した場合に，4個の改良図が得られる．これらの中から，3本の直線の交わる角度が最も大きい(すなわち，これらの回帰直線の変動が最も大きい)ものを見い出すのは容易である．

データ省察で用いられるグラフィカル手法は「その場しのぎ(ad hoc)」的なものであることが多い．したがって，「標準手法」はその場に適応して修正が加えられるものでなければならない．このことはその手法の形骸化した利用を防ぐことにもなる．また，「見方」のわかる使い慣れた手法，すなわちマイ・ツールを一つでも多く利用者自身の道具箱の中に用意することが必要である．そのような身近かな工夫と対処が，利用者の場面に応じた適切なデータ省察用グラフィカル手法の臨機応変の使い分けを提供するに違いない．

参考文献

Bacon-Shone, J. & Fung, W. K. (1987). A new graphical method for detecting single and multiple outliers in univariate and multivariate data. Appl. Statist., 36, 153-162.

Fowlkes, E. B. (1987). A Folio of Distributions. Marcel Dekker.

後藤昌司・上坂浩之・浅野長一郎(1979)．NISANシステムにおけるdata investigation．文部省科学研究費特定研究A-4班「情報システムの形成過程と学術情報の組織化：統計プログラム・パッケージの研究」(研究代表者丘本 正)，101-110.

後藤昌司(1991)．統計的グラフィクスの顧客の創造に向けて：最近の発展．日本統計学会60周年記念：統計学の現状と将来展望に関するシンポジウム：パネル討論「統計的グラフィクス」，21-27.

余田明夫・松原義弘・後藤昌司(1990)．統計的グラフィクスにおける卑近な方法：最近の発展と適用上の留意点．SHI-Preliminary Research NO. 187，塩野義解析センター［松原義弘・余田明夫・後藤昌司(1989)．統計的グラフィクスにおける卑近な方法：最近の発展と適用上の留意点．第17回日本行動計量学会大会発表論文抄録集，119-122］.

統計的グラフィクスの最近の展開

塩野義製薬(株) 解析センター 後藤昌司
松原義弘
田崎武信

統計的データ解析過程の諸種の側面で終始一貫して効用を発揮しているツールは統計的グラフィクスをおいて他にない．一般に，統計的方法は，場面や状況に応じて適応的に用いられないと，その有用性を発揮しない．その意味では，データのグラフィカル表現を臨機応変に工夫する姿勢そのものが統計的データ解析での方法論の本領であるといえるかもしれない．

統計的グラフィクスの呼称のもとに視覚表現に訴える多くの諸法が開発されている．これらの諸法が世の中に受け入れられ，有用性を発揮するには提案者の力倆(実績・思想)，時代の好尚(背景)，成熟度(状況による適応力)が必要のようである．最近の統計的グラフィクスの隆盛を裏で支えているのはコンピュータの発展とそれに伴う計算環境の整備である．すなわち，現時点で開発されている大半の統計的グラフィクスには，本質的にコンピュータの利用が不可欠である．最新のコンピュータ・グラフィクスを用いると，程々の費用で会話性向の高い統計的グラフィクスを開発することが可能である．とくに，マウスなどの画面入力装置を用いることで，データ解析者はグラフィクスの要件を指定し，それらの装置を操作し，表示画面を即時に連続して変化さすことができる．それらは過去における静的グラフィクスに対する媒体とは有意に異なる．現在では，それらの高性能・高品質のハードウェアとソフトウェアが比較的に安価に利用できるようになっている．今後にかけて，統計的グラフィクスを研究するには，これらの計算環境に注目し，それに纏る「情報リテラシー」を身につけることが不可欠である．とくに，著名な統計ソフトウェアについては従来のメイン・フレーム版がパソコンやワークステーション版に移植され，それらの中に含まれている統計的グラフィクスはディスプレイ表示できるように改良されている．また，最近に開発されている殆どの統計ソフトウェアには動的グラフィクスが組み込まれつつある．このことは統計的グラフィクスが定型的方法の補完的役割を果たし，統計的データ解析がグラフィクスだけで完遂できないことも物語っている．

統計的方法論の最近の研究・開発動向[BMS, 1989a, 1989b]からもみれるように，統計的グラフィクスは今後にかけて統計的方法論の研究主題の主役として一段と脚光を浴びそうである．このとき，次の点に留意しておきたい．

①最近の統計的方法論の発展を固有技術的側面でみるとき，規模が大きく輻輳する諸種のデータを処理・解析することのニーズに応えて，人的要件を主とする深耕した技と人工知能を加味したエキスパート・シ

ステムの開発が要請されつつある．また，共通技術的側面で眺めるとき，そこでは多種・多彩なツールが要請されている．我々が統計的グラフィクスにとくに望むのは前者の側面では難度を緩和することへの貢献，および後者の側面では動的標準化と諸種インターフェイスへの働きである(後藤，1991)．

②人的側面から考えるとき，人にとって「見る」ということは一種の自己投影であり，人には自らの内に関心を抱いているものしか見えないのが普通である．人のこの特徴に留意し，統計的グラフィクスの過去・現在・未来について徹底した総合省察(review)を行うことは埋もれている「文化(方法論)」に光をあてる点で非常に重要である．また，人間は「忘れる動物」であることに注意すると，積極的に忘れる努力としてできるだけ見通しの効く形式で財産(文献・作品)の整理を試みることは，本能的に行動を起こすことのできる「潜在能」に蓄積をはかるうえでも有意義であろう(後藤 他，1988：Goto et al.，1991)．

③統計的グラフィクスが統計的データ解析過程で果たす特徴(長所)として豊潤性，融通性，聡明性，有用性，簡明性が強調されることが多い(Goto, 1987)が，実際の適用では主観的(独断的)，最適性基準の欠如，再現性の疑問などの短所も指摘されている．とくに，一つの数表ないしグラフの明瞭な表示が正確に何を明らかにしようとし，その情報の基盤がどこにあるかという根本的な問題は常に忘れてはならない．

参考文献

Board on Mathematical Sciences(BMS), Commission on Physical Sciences, Mathematics, and Resources, National Research Council, Washington, D. C. (1989a). Statistical Sciences:Some research trends-statistics. The IMS Bulletin, 18(2), 181-193.

Board on Mathematical Sciences(BMS), Commission on Physical Sciences, Mathematics, and Resources, National Research Council, Washington, D. C. (1989b). Statistical Sciences:Some research trends-probability. The IMS Bulletin, 18(4), 397-411.

Goto, M. (1987). Statistical graphics: Discussion. Proceedings of the 46th Session, ISI, Tokyo, 401-402.

Goto, M., Matsubara, Y., Yoden, A., Tsuchiya, Y. & Wakimoto, K. (1991). Statistical graphics : A classified and selected bibliography. J. Japan Statist. Soc., 2(1), 97-121.

後藤昌司・松原義弘・脇本和昌(1988)．グラフィカル接近法の最近の発展．行動計量学，15(2)，45-70，1988

後藤昌司(1991)．統計的グラフィクスの顧客の創造に向けて：最近の発展．日本統計学会60周年記念シンポジウム：パネルディスカッション「統計的グラフィクス」，1991-7-25．神戸(SHI-Preliminary Research NO. 198，塩野義解析センター)．

多変量データの順序付けとその応用

塩野義製薬㈱ 解析センター　松原義弘
後藤昌司

順序統計量は統計的方法の種々の適用場面で重要な役割を演じている．とくに，順序付き観測値や順位に基づく接近法は簡便性と頑健性の長所を有しており，実地で繁用されている．しかし，「多次元空間には自然な順序がない」ことから，限定した形式での順序付け(所謂，半順序付け)原理が提案されている(Barnett, 1976)．本報告では，解析目標に依存しなく，観測値の散布状況だけに依存する多変量観測値の順序付けを試みた．その方法は凸包の構成とその皮むき(ピーリング)に基づいており，その順序付け概念は Barnett(1976)の部分順序付け(P法)の範疇に属する．ここでは，２変量の場合に注目し，実地に応用の効く形式でグラフィカル接近法との連結をはかった．

多変量データに対する凸包はすべての観測値を含む最小の凸多面体を構成することで得られる．このとき，多変量観測値の同時あるいは周辺での極値はその標本に関する凸包の頂点として定義できる(Barnett, 1976)．これらの凸包の頂点をすべて捨て去り，残りの観測値について新たな凸包を構成することが凸包ピーリングと呼ばれる(Eddy, 1982)．これは単一変量での刈り込み概念の一般化になっている．これらの基本概念を組み合わせ，さらに標本の重心から観測値までの Euclid 距離，あるいは観測値の散布方向などを一つの基準とすることにより，多変量観測値のいくつかの順序付けが行える．このときの順序付けを利用することにより，極値，中央値，範囲，４分位点などの単一変量での順序統計量の概念をより高次に拡張することができる．とくに，２次元の場合にはボックス・ウィスカー・プロットと類似なグラフィカル表現が可能である．

いま，２変量確率変数Xのn個の観測値の確率標本を$x_1, \cdots, x_n$で表し，新たに次の集合を定義する．すなわち，全観測値集合をQ_nとする．これらの観測値をすべて含む最小凸包を構成し，全観測値の重心から凸包の頂点に対応する観測値までの距離の長い順に観測値を順序付ける．これらの順序付き観測値を順位の低い順に捨て去り，そのたびに構成される凸包に含まれる観測値の集合を$Q_{n-1}, Q_{n-2}, \cdots$で表す．これらの集合は包含関係

$$Q_n \supset Q_{n-1} \supset Q_{n-2} \supset \cdots$$

にある．このとき，任意の$k(1 \leq k \leq n)$に対して，$\{x_i\}$がQ_kに含まれる割合は

$$G_n(Q_k) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i \in Q_k\}$$

である．ここに，$I\{x_i \in Q_k\}$は$x_i \in Q_k$が真のときに1，偽のときに0をとる指標関数である．したがって，$G_n(Q_k)$は階段関数であり，単一変量の場合の経験累積分布関数に相当する．また，ボックス・ウィスカー

・プロットのヒゲの先端にあたるデータの最小値と最大値はそれぞれ$G_n^{-1}(1/n)$と$G_n^{-1}(1)$に対応する点と凸包，およびボックスの上辺と下辺を構成する第1四分位点と第3四分位点はそれぞれ$G_n^{-1}(0.25)$と$G_n^{-1}(0.75)$に対応する凸包に相当する．ここでは，このプロットを凸包プロットと呼ぶ．より一般的に，$G_n^{-1}(Q_k)$は2変量分布の標本確率等高線図を与える．

ピーリングだけに基づく上記の順序付けは，単一変量の場合での順序の一つの概念，すなわち「小から大」，「軽から重」，あるいは「低から高」などのような方向性をもたない．ここでは，2変量の場合の方向性の基準として，相関係数の符号を利用した．このときの順序付けは次の手順で行える．

①規準化した2変量対(x_1, x_2)の相関係数rを求める．

②$r>0$のとき，直線$x_2=-x_1$，$r<0$のとき直線$x_2=x_1$を境界とする半平面をそれぞれA，Bとし，これらの半平面に属する観測値の集合をそれぞれS_A，S_Bとする．

③凸包ピーリングまたは点ピーリングによる順序付けを行う．このとき，対応する観測値が集合S_Aに属するときには1から昇順に，対応する観測値が集合S_Bに属するときにはnから降順に順序付けを行う．

上記の方法を人工データと文献データに適用し，その性能を評価し，さらに2変量順序と周辺順序(単一変量順序)の対応関係を小規模シミュレーションで吟味した．その結果，得られる順序と周辺変量の順序との対応の強さは原データの相関の強さに依存していた．比較や分類を意図する場合の順序付けにはこの方向性を考慮した順序付けが適切であることが示唆される．ただし，上記の順序付けには二三の問題が残されている．第1に，この方法では尺度不変性が満たされないことである．多変量順序統計量についての尺度の評価を別途に検討することが必要である．便宜的には変量毎に規準化をはかることが考えられる．第2に，この方法での同順位のとり扱いの問題である．これについては，単一変量の場合の同順位のとり扱いが適用できると考えられる．第3に，この方法を3次元以上に直截的に拡張すること，とくにグラフィカル表示が困難なことである．この場合には主成分解析や平面射影による次元の縮約が考えられる．このとき，多変量観測値の順序付け後に「ノンパラメトリック・グラフィクス」の諸法も利用できる(Matsubara, et al., 1987：松原 他, 1989).

参考文献

Barnett, V. (1976). The ordering of multivariate data. J. Roy. Statist. Soc., A139(3), 318-355[後藤昌司・土屋佳英(1985). 多変量データの順序付け. SHI-SEMINARY NOTE NO. 3, 塩野義解析センター].

Eddy, W. F. (1981). Comment on the paper of Friedman & Rafsky(1981). J. Amer. Statist. Assoc., 76, 287-289[後藤昌司・松原義弘(1988). Friedman & Rafsky(1981)の論文に対する討論. SHI-SEMINARY NOTE NO. 46, 塩野義解析センター].

Matsubara, Y., Tsuchiya, Y. & Goto, M. (1987). Graphical comparisons of multivariate data. Computational Statistics & Data Analysis, 5(2), 103-112.

松原義弘・余田明夫・後藤昌司(1989). 多変量データの順序付けとその応用. SHI-Preliminary Research NO. 177, 塩野義解析センター.

多重比較検定とグラフィクスの対峙と調整

塩野義製薬(株) 解析センター 田崎武信
財前政美
後藤昌司

一般に，検定と区間推定は随伴関係にある．そして，区間推定はグラフィカル表現に直結する．しかし，多重検定あるいは多段階検定の場合には事情が異なる．不確定項のプロットに伴う諸種の工夫(例えばLenth (1988)を参照)はグラフを検定に適合させようとする試みであり，グラフからの検定への懸命のあゆみよりである．一方，検定からグラフへのあゆみよりは2者択一の論理にp値を添えること，有意水準を数段階で変化させることぐらいであろう．グラフに探索的あるいは仮説創成的な役割をもたせない限り，グラフと検定の調整では検定が主体になるであろう．したがって，無理に調整するよりも相補的な役割をもたせるのが望ましいかもしれない．例えば，検定にはグラフの視察から得られる知見の確かさを評価させ，一方，グラフには検定の結果を常識に基づいて点検させることが考えられる．

本報告では，多重比較検定が適用される場の本来の目的(例えばTukey (1949)を参照)を再考することで，問題を検定よりもむしろ分類におきかえる統計的方法，すなわち標本平均クラスタリング法をとりあげ，そしてそれらの方法に伴うグラフィカル表現を議論した．実際にはTasaki et al. (1987)でとりあげられた6個のクラスタリング法とMcLachlan & Basford (1988)の混合分布モデルの接近法とを主に議論した．これらの方法のうち階層的なクラスタリング法では，その結果をデンドログラムに要約して表現するのが便利である．ここに本報告では，標本平均の比較に適したデンドログラム表現の修正を提案した．また，混合分布モデルのあてはめ結果を表現するためのグラフィカル手法を提案した．最初の提案では，例えば凝集型の階層的クラスタリング法をとりあげると，クラスターとクラスターとの併合可能性が棄却される段階を実線でなく点線で表現した．同じ精神に従い，分岐型の階層的クラスタリング法では初めて有意でない分割を，そして段階的B法(Bozdogan, 1986)を凝集型のクラスタリング法と見做して得られるデンドログラムでも最小AICの次の段階の併合を点線で表示した．第2の提案では，正規混合分布モデルをあてはめて得られた(各標本平均の各クラスターへの所属の)事後確率の推定値を$\{\hat{\tau}_{ji}: j=1,2,\cdots,m; i=1,2,\cdots,k\}$とする．ここに，$k$は標本平均の個数，$m$はクラスター(成分分布)の個数である．これらの推定値を次のように半径1の円の上半分の内側で表示した．最初に，標本平均を大きさの順に並べ，最も小さなものを正の方向，最も大きなものを負の方向の半円上に布置させた．残りの標本平均は比例配分した角度の半円上に布置させた．このとき，原点から半円上の標本平均の点へ伸ばした座標軸がk本 構成される．次に，成分j ($j=1,2,$

$\cdots, m$)について，上記のk座標系で$\{\hat{\tau}_{j1}, \hat{\tau}_{j2}, \cdots, \hat{\tau}_{jk}\}$の点を連結線分で表示した．

標本平均のデンドログラム表現には問題も残されている．本質的な問題として，１変量の場合に，本来１次元に配置できる平均値の集合にデンドログラムのような超次元の枠を強制することに疑問が残る．また表層的な問題として，Jolliffe(1975)とCalinski & Corsten(1985)にそれぞれ記載されているデンドログラムには微妙な相違点がある．横軸に平均値を並べ，縦軸でクラスターの結合を表示することは共通している．しかし，Jolliffe(1975)では平均値を小さいものから大きいものまで等間隔に並べ，結合のｐ値を縦軸にとっている．一方，Calinski & Corsten(1985)では平均値を横軸にとり，結合の順序のみを等間隔に並べている．ここでは後者の表現に倣った．これは趣味の問題であるかもしれない．もちろん，本報告での提案と同じ考え方に基づいて，Jolliffe型の修正を提示することは容易である．

正規混合分布モデルのあてはめにおける推定事後確率のプロットでは最初にレーダ・チャートの利用を考えた．しかし，平均値をそれらの大きさの順に等間隔に配置すると，最も小さな平均値と最も大きな平均値が隣接することに気づいた．この不備を除くために，半円を利用することにした．また，等間隔でなく平均値の大きさに従って180°を分配させた座標軸を利用することにした．ここで提案したグラフィカル表現法においては，外に向って最も張りの強い，すなわち各平均が半円周に最も接近する図柄を与えるクラスター数が標本平均を分類するのに適切なクラスター数の候補になる．これは近似分布に基づく尤度比検定の代替になる．

参考文献

Bozdogan, H. (1986). Multi-sample cluster analysis as an alternative to multiple comparison procedures. Bulletin of Informatics and Cybernetics, 22, 95-130.

Calinski, T. & Corsten, L. C. A. (1985). Clustering means in ANOVA by simultaneous testing. Biometrics, 41, 39-48.

Jolliffe, I. T. (1975). Cluster analysis as a multiple comparison method. In Applied Statistics, R. P. Gupta (ed.), 159-168. North-Holland.

Lenth, R. V. (1988). Graphics for multiple comparisons. Proceedings of the Section on Statistical Graphics, ASA, 70-75.

McLachlan, G. J. & Basford, K. E. (1988). Partitioning of treatment means in ANOVA. In Mixture Models :Inference and Applications to Clustering, Chap. 6, 145-171, Marcel Dekker.

Tasaki, T., Yoden, A. & Goto, M. (1987). Graphical data analysis in comparative experimental studies. Computational Statistics & Data Analysis, 5, 113-125.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, 5, 99-114.

順位グラフ上のクラスタリング

統計数理研究所　　馬場康維

1．はじめに

感覚的な計測などで、ランキングは重要な役割を果しているが、ランキングに基づくデータ解析の方法はまだ十分研究されていない。そこで、その一つとして、ランキングに基づく分類法について報告する。

kアイテムをn人の判定者が、順位づけする場合を考える。得られるn組のランキングデータを（$R_{i1}, R_{i2}, \cdots, R_{ik}$；i＝1，…，n）とする。ここで、（$R_{i1}, \cdots, R_{ik}$）は(1,…,k)の置換の一つである。こういうランキングデータを扱う方法としては、ケンドールの一致性係数W，クレーマーの検定、順位グラフなどがある。この報告では、順位グラフをベースにした分類法について述べる。

2．順位グラフ

ベースになる順位グラフ(1987, Baba)の描き方を簡単に説明する。n人の判定者が、k個のアイテムに順位づけする。j番目のアイテムをr順位とした判定者数をf_{jr}とする。相対頻度を

$$p_{jr} = f_{jr}/n$$

とし、順位rに角度

$$\theta_r = \frac{r-1}{k-1}\pi \qquad (1)$$

を対応させる。

順位グラフ

(1)　まず半径1の半円を描き、(1)式で与えられる角度に対応する目盛をつける。θ_rは順位rに対応する方向を表わす。

(2)　次式で定義されるアイテムベクトルを描く。

$$X_j = \sum_{r=1}^{k} X_{jr}$$

$$X_j = (p_{jr}\cos\theta_r,\ p_{jr}\sin\theta_r)$$

アイテムベクトルは

①　長さが、評価の一致度に対応する。

②　方向が平均順位に対応する。

という性質を持っているため、順位グラフにより、平均順位と一致度という2つの面から各アイテムの評価ができる。すなわち、アイテムベクトルの方向によりアイテムの平均順位が読み取れ、アイテムベクトルが円周に近いかどうかで、アイテムの評価が一致しているかどうかが判定できる。

3．順位グラフによるクラスタリング

順位グラフ上の座標系は、（平均順位、一致度）の座標系である。したがって、順位グラフ上の点は通常のユークリッド空間上の点のようには扱えないが、便宜上、この点をユークリッド空間上の点として扱い、順位グラフ上のクラスター分析ができる。

4．モデルケースによる解析

順位グラフによるクラスタリングが、どんな時に有効かを、簡単なモデルケースを比較することによってみてみよう。

Ａ～Ｏまでの１５のアイテムがある。このアイテムは、G_1, G_2, G_3, G_4の４つのグループに分類され、それぞれのメンバーは以下のようであったとしよう。

G_1＝（Ｄ，Ｊ，Ｌ，Ｏ）
G_2＝（Ｇ，Ｎ）
G_3＝（Ｂ，Ｃ，Ｋ，Ｍ）
G_4＝（Ａ，Ｅ，Ｆ，Ｈ，Ｉ）

典型的な３つの場合を考えてみよう。

［ケースⅠ］　４つのグループは明らかに優劣があり、どの判定者もグループ間の区別はつく。しかし、グループ内のアイテム間の区別はできない。すなわち$G_1 < G_2 < G_3 < G_4$となるケースである。

［ケースⅡ］　４つのグループには不完全ながら順序がある。ただしグループ内のアイテムには区別がない。ここでは、$G_1 < G_4$が成り立ち、G_1, G_2, G_3, はランダムに順序がつき、G_2, G_3, G_4もランダムに順序がつくという場合を想定する。

［ケースⅢ］　G_1, G_2, G_3, G_4に優劣がなく、ランダムにグループ内の順序がつけられる。また、グループ内の順序もランダムである。

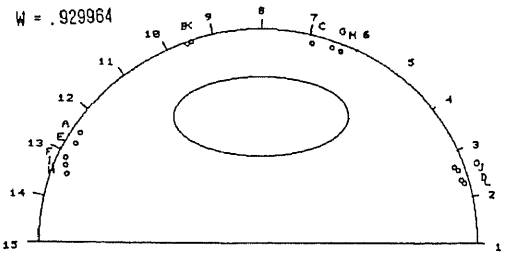

図1　ケースⅠの順位グラフ

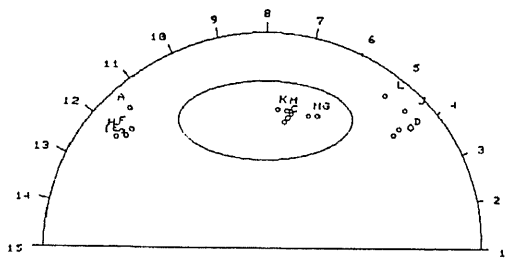
図2　ケースⅡの順位グラフ

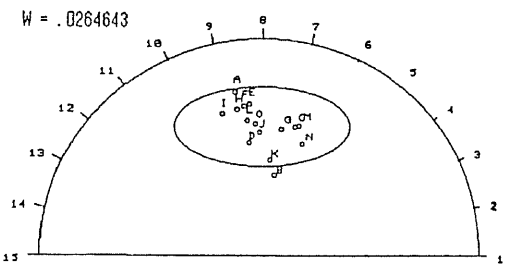

図3　ケースⅢの順位グラフ

これらの３つのケースにつきアイテムベクトルの終点を順位グラフ上に描くと図１～３のようになる。いずれの例も、２０人の判定者を想定したシミュレーションの結果である。

この３つのグラフを比較して次のようなことがわかる。

(1) アイテムベクトルによるアイテムの評価と同じようにグループの評価ができる。

即ち、グループの位置から、平均順位の高いグループか低いグループか、評価の一致度が高いか低いか等がわかる。

(2) 全体として、区別がはっきりしているときは、点は円周の近くに集まり、ほとんど区別がつかない場合は、平均的な位置に集まる。したがって分離の良さが、全体の分布から推察される。

参考文献

Baba, Y. (1987). Graphical Analysis of Rank Data, Behaviormetrika, No. 19, 1-15.

連結ベクトル図による検定統計量

白旗慎吾（大阪大学・教養部）

データをグラフに表現し、それに基づいてデータを解釈することは統計学における重要な一分野である。ただし、グラフ表現から視覚的に解釈するとき、その解釈は多く直観のみに頼り、数理的な厳密さを欠く。予備的な解析の場合にはそれで十分なことも多いが、客観的な評価が必要な場合には、傾向の強さを確率で評価したり、検定統計量を構成しその分布を調べる必要がある。

linked line chart による方法とは，１つのデータに１つのベクトル（有向線分）を対応させ，それらの線分を結んで適当な図形を描く方法である．その図形を見ることによって，視覚的にデータの情報を読みとることが可能となり，またそこから統計量を構成する．linked line chart による方法では、多くの場合図形は多角形にとり，帰無仮説のとき，または対立仮説で最も極端な場合に正多角形（極限では円）や見やすい図形になるように描く．多変量データの場合は，問題によっては各変量ごとにベクトルを対応させそのパターンを見たり，多変量データ１個に１つのベクトルを対応させたりする．正多角形をつくるのは，それからの離れ方が視覚的に捉えやすいこと、および後に統計量を構成することを意識してである。

辺の長さが一定の凸多角形は正多角形のとき面積最大である。この性質から、多くの問題で面積を統計量にすることができる。多変量データで個々の変量に線分を対応させる場合はベクトルを結んだ図形の終端の位置や基準点からの距離を見ることも多い．

講演では、以下の問題に対する linked line chart とそれに基づく統計量、および分布論を紹介した．

(1) 目的変数が k 個の説明変数をもち、データが順位に変換されている場合を考える。このとき、目的変数と各説明変数の間の相関を視覚的に測るためのグラフ表現とそれに基づく相関係数を定義し、その漸近正規性について報告し、グラフの例を表示した。グラフは、相関が 1 なら漸近的に半円、相関が -1 なら円になるように構成され視覚的に分かりやすい。この表現には、星座グラフなどの色々なヴァリエーションがある。

(2) k 変量連続分布からの無作為標本からの順位データを考える。k 個の変量の

間の一致性の強さ、すなわち、変量の間に非負の相関を仮定し、その強さをグラフに表現し、それに基づく統計量を構成した。統計量は非線形順位統計量であるが、その分布は漸近的に正規分布に従う。さらに、Kendall の一致度係数との比較を行なった。すべての変量の間の相関が 1 のとき漸近的に円になるようにグラフ表現される。相関がないという帰無仮説に対する Kendall の一致度係数に対する効率は 1 に近い。

(3)大きさ n の無作為標本がある指定された分布に従うかどうかの適合度検定におけるグラフ表現とそれに基づく検定統計量を紹介した。データを一様分布に変換し、順序統計量をグラフに表現する。指定された分布に適合するとき漸近的に円に近づくように表現される。統計量は退化した核関数を持つU統計量であり、漸近正規性を持たないが、カイ２乗分布の重み付き和で表現され、その漸近分布の数値計算が可能である。さらに、位置母数や尺度母数を推定した場合の性質についても紹介した。

(4)大きさ n の無作為標本に基づいた対称性の符号付き順位検定はノンパラメトリック検定でよく用いられる。符号付き順位をグラフに表現し、符号付き順位検定統計量がグラフのどのようなパターンで表現されるかを紹介した。他の場合と異なり、符号付き順位は正多角形や円では表現されない。さらに、グラフ表現に基づくいくつかの検定統計量の構成とその分布についても考察した。

(5) k 個の標本が同じかどうかを検定する分散分析問題に関するノンパラメトリック検定を考える。順位に基づくグラフ表現を紹介した。表現は k 個のグラフを描き、それぞれが正多角形に近いとき帰無仮説が成り立つと解釈できるようになされる。さらに、各多角形の面積や、頂点の座標を利用していくつかの検定統計量を構成する。統計量は漸近正規性が成立するもの、重み付きカイ２乗分布の重み付き和で表現されるもの等、統一的には扱えない。通常の順位検定、Anderson-Darling 検定、Watson 検定などとの効率の比較も行なっている。

(6) (5) と同じ問題で、ただし、k 個の母集団の間に順序があらかじめ想定されている場合のノンパラメトリックなグラフ表現とそれに基づく検定統計量、およびグラフ表現を紹介した。統計量は線形順位統計量となり、漸近正規性が成立する。この問題での標準的な順位検定は Jonckeere 検定である。Jonckeere 検定は２標本問題での Mann-Whitney 統計量の単純な和であるが、ここでの統計量はその重み付き和になる。Jonckeere 検定との効率の比較が非常に 1 に近いことが分かる。