正　誤　表

〈経営総合科学　第67号〉

本文中に誤りがありましたので、下記の通り訂正いたします。

頁・行	正	誤
6　頁 上から13行目	無相関から$\frac{1}{n}F'F = I,$	無相関から$\frac{1}{n}Z'Z = I,$

[研究ノート]

多変量統計解析手法の覚え書き

神頭広好

I　はじめに

今日に至っては，コンピューターやソフトの開発は目覚ましいものがあり，研究分野にもよるが，理論的な基礎が理解されていないと，中々アウトプットが解釈できない事態が多々ある。筆者は，10年位まえからMacintoshを利用してきたが，当時の多変量解析ソフトStat80において，100サンプル50変数の計算が10～15分位かかったものが，今では（機種にもよるが）Systat，StatView，Jump，Statistica，SPSSなどのどのソフトを使っても，ほぼ同じことが数秒で計算され，かつグラフが容易に描出される。ここで，上記ソフトの特徴を述べさせて頂くと，解析手法の多さやグラフの種類から総合的に判断すれば，Systatがベストなソフトであろう。しかし，日本語化されていないので，和文で論文などを作成する場合統計英語を訳す必要がある。Jumpについても同様である。このソフトは主として回帰分析に強く，素早く同時に異なる形状の推計を行い，視覚化を試みる。StatisticaはSystat同様に統計解析手法やグラフなどの種類も多いが，バージョンアップが比較的多く，信頼性についての不安が少し残る。しかし，このソフトは日本語化されているのでそのまま日本語の論文にペーストすることができる。古くからのMacユーザーの間で最も日本で使われているソフトはおそらくStatViewであろう。これには頻繁に使われている最小限度の解析手法（回帰分析，分散分析，因子分析など）は含まれており，Jump同様に素早いグラフを描出するため無駄な時間を減らすことができ

る。また，これは日本語化されており，特に初心者において最も使いやすいソフトである。最後に従来から世界的にも有名であるSPSSについては，Systatおよび Statistica 同様に解析の種類は豊富である。とくに時系列分析や非計量的なアンケート調査を含む分析（コレスポンデンス分析等）など，Excelとともに力を発揮する。特に6.0Jにバージョンアップしてからは，他のソフトとの互換性も高まり，使いやすくなっている。ただし，3次元などのグラフの種類が少ないことや，Windows版SPSSには共分散構造分析などができるアプリケーションがあるがMac版については今のところない。

Ⅱ　多変量統計解析手法と文献

以下では,主に利用されている統計解析手法とそれに関わる文献を見ていく。ここで，重複を避けるために特に初心者向けの文献を紹介しておくと，[12]，[18]および[44]などがあげられる。また,手続き的なものとしては，[13]，[16]，[24]，[25]および[35]などがある。ついで，統計学の基礎学力が備わった中級者向けのものとしては，[2]，[3]，[4]，[10]，[17]，[20]，[21]，[22]，[31]，[38]，[39]，[49]および[50]など多数ある。比較的高度な統計および数学的理解力のある上級者向けのものとしては，[15]および[47]などがある。また，地理研究者や経営，マーケティング研究者向けのものとしては，[8]，[19]，[33]，[34]，[38]および[39]などがある。

1　重回帰分析

この分析手法は,主として予測や推計に用いられるが,クロスセクションデータを使って経済社会構造を分析することも可能である。

(1)線形モデル

線形関数を次のように設定する。

$$y = bX + e$$

この関数から最小二乗法（誤差 e の二乗和を最小にする方法）を用いて整理すると，

$$b = (X'X)^{-1} X'y$$

が導かれ，b が推計される。なお，ここでは検定等の説明は省略する。このモデルについては，統計学や計量経済学においてほとんどどの書物を見ても十分説明がなされている。ここでのほとんどの参考文献において，統計との関わりで重回帰分析が平易に説明されているが，とりわけ，[4]，[20]および[21]などが分かりやすい。

(2)非線形モデル

変数を対数や乗数などに変形しても最小二乗法が使えない場合の非線形モデルでは，最尤推定法（[4]，第6章）やシンプレックス法などの探索法などによって係数を求める分析手法がある。これについては[9，第4章]を参照せよ。また，線形および非線形の適合性について検討する際には Box-Cox モデルなどが有効である。因に，Systat ではシンプレックス法が採用されている。

2　数量化Ⅰ類

この分析手法は，質的データに対する回帰分析手法であり，変数がダミー変数であることから，方程式が一意的に決まらないため，通常各アイテムの第1カテゴリーを削除することよって，上記の最小二乗法を用いて係数が推計される。なお，これについては，[12，第6章]，[21，第6章]および[27]などを参照せよ。

3　判別分析

この分析手法は，2つ以上あってもかまわないが，個体あるいはサンプルがどちらの群に含まれるかを分析するのに用いられる。

(1)マハラノビスの距離

$$D_k^2 = (x - \mu^{(k)})' \Sigma^{-1} (x - \mu^{(k)})$$

ただし，Σ は分散共分散行列を示す。

もし x が $D_1{}^2 < D_2{}^2$ ならば1群に属し，$D_1{}^2 > D_2{}^2$ ならば2群に属することになる。これについては，[12，第4章]および[17，第3章]，地域分析への応用例としては[33，第4章]などで説明されている。

(2)多重ロジスティック分析による方法

確率 p_i がロジスティック曲線によって表されることを仮定すると，対数オッズ（log odds）は次のようになる。

$$\log \frac{p_i}{1-p_i} = \sum_{j=1}^{t} \beta_i x_{ij}$$

ただし，p_i はサンプル i が選択する確率，x_{ij} はサンプル i の j 変数をそれぞれ示す。

ここで θi を実現値とすれば，サンプル n に関する尤度関数は

$$L = \prod_{i=1}^{n} (p_i)^{\theta_i} (1 - p_i)^{1-\theta_i} = \exp(\sum_{j=1}^{t} \beta_j x_{ij} y_i) / \prod_{j=1}^{n} (1 + \exp(\sum_{j=1}^{t} \beta_j x_{ij}))$$

と表される。この関数を最大化するために，この関数を β_j で偏微分することによってロジスティック曲線が得られる。これについては[48，p.73]を参照せよ。

(3)正準判別分析による方法

合成変数 $f = \sum_{j=1}^{n} x_j a_j = Xa$ の相関比（級間分散/全分散）を最大にするように重み係数ベクトル a を求める。したがって，$\frac{a'S_B a}{a'S_T a}$ を最大化するために，次の固有方程式 $S_B a = \lambda S_T a$ を解くことになる。ただし，S は分散共分散行列を示す。その結果，固有値の数だけ個人またはサンプルに対して正準判別関数が

計算され，これらを各郡ごとに計算を行い，各郡の重心を求め，各個体はそれとの距離から，最も短い距離の群に判別される。なお，これについては［48,pp.73−76］参照せよ。他にベイズ流の判別規則などがある。また，判別関数適用上の注意などが，［49,pp.216-226］に掲げられている。

4　数量化Ⅱ類

これは，質的データにもとづく判別分析手法である。

なお，同手法の詳細については，基礎的な説明は，［12，第7章］，［22,第13章］および［27］などによってなされている。

5　主成分分析

この分析手法は，多くの変数が有する情報をできるだけ簡略化して，そこから得られた主成分の解釈を容易にするために用いられる。

基本的な関数は $z = a'X$ と表され，これらの係数 a' が極端な大きさになることを防ぐために，$a'a = 1$ の制約を付ける。

標準化された分散行列 (V) は，相関行列 (R) を意味する。そこで，これを

$$V\{z\} = V\{a'X\} = a'Va = a'Ra$$

と表示して，ラグランジュ法を用いて整理すると，

$$Ra = \lambda a \text{ または } (R - \lambda I)a = 0$$

次の特性方程式

$|R - \lambda I| = 0$ から λ が計算され，これを $(R - \lambda I)a = 0$ へ代入することによって，a が導出される。

同分析手法の初歩的な説明は，［32，第3章］，［3，第5章］，［17，pp.68-86］および［21，第2章］などによってなされている。また，多変量解析用のソフトにおいて，主成分分析は因子分析に含まれているものが多い（例えば，Systat，Statistica，StatView などほとんどのソフト）。なお，主成分因子分析手法およ

び因子分析手法のプログラムを用いる主成分分析手法については，[3, pp.67-74]を参照せよ。

6　数量化Ⅲ類

これは，質的データにもとづく主成分分析手法である。

この分析手法については，[12，第8章]が初歩的に説明されており，[17，第6章]，[22，第14章]，[27]および[49，pp.185-203]などが平易に説明されている。

7　因子分析

この分析手法は，ある仮説のもとで，現存する情報の背後に潜む要因をグループ化し，それらのグループに対して意味付けを行うために用いられる。まず，以下の線形の関数を設定する。

$$Z = FA' + UD$$

相関行列は，$R = \frac{1}{n} Z'Z$ となる。

また，因子得点間の無相関から $\frac{1}{n} Z'Z = I$，独自因子間についても $\frac{1}{n} U'U = I$ 共通因子と独自因子間についても $F'U = 0$ が仮定される。したがって，$R = A'A + D^2$ また，それから D の対角要素を引くと，$R^* = R - D^2 = A'A$ として表される。後は，主成分分析同様に $|R^* - \lambda I| = 0$ の解 w を用いて因子負荷量と λ が導出される。ただし，$w'w = 1$である。

〈因子軸の回転〉

因子の解釈をしやすくするために因子軸を回転させる必要がある。例えば，次の $TT' = T'T = I$ となる直交行列 T を用いると，$B = AT'$ という因子負荷行列ができるが，$BB' = (AT')(AT') = AT'TA = AA'$ となり，もとの因子負荷量に変化を与えない。

〈因子得点の導出〉

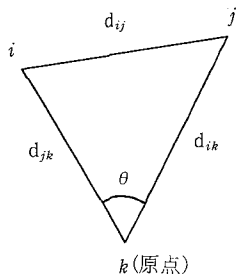
$E = Z - FA'$ から E を最小にするように F を導出する。すなわち，最小二乗法を用いて，$F = ZA(A'A)^{-1}$ が推計される。

同分析手法を理解するために，初心者向きのものとしては[12]および[18]，一般向けのものとしては[1]，[21]，[36]および[43]，少し踏み込んだものとしては[46]などがある。また，特に主成分分析との相違については，[49,pp.182-185]で説明されている。なお，同分析手法は，マーケティングの分野においてSD法（Semantic Differential Method）とともに用いられている（[11]を参照）。

8　多次元尺度構成法（MDS）

この分析手法は，類似性データにもとづき，潜在する少数の次元に簡略化し，背後にある構造を明らかにするために用いられる。

(1)距離データのMDS



上図から $d_{ik}{}^2 + d_{jk}{}^2 - 2\ d_{ik}d_{jk}cos\theta = d_{ij}{}^2$，また三角形の余弦定理から $b_{ij} = d_{ik}d_{jk}cos\theta$ が成り立つ。これより，b_{ij} は $b_{ij} = \frac{1}{2}(d_{ik}{}^2 + d_{jk}{}^2 - 2\ d_{ij}{}^2)$ で表される。また，個体i，jの原点からの内積 b_{ij} は，$b_{ij} = \sum_{t=1}^{T} x_{it}x_{jt}$ で表される。したがって，これを行列表示すると，$B=XX'$ と書かれ，これを因子分解するのと同様な方法で X（x が因子負荷量 A に相当する）を導出することができる。ここで，$B = TLT'$ から $X = TL^{\frac{1}{2}}$ となる。ただし T は固有ベクトルであり，L は固有値行列である。したがって，固有ベクトルが距離を，固有値が次元をそれぞれ示している。

なお，この導出方法については，[23，第5章]および[50，第10章]で説明さ

れている。

(2)非計量 MDS

距離が順序尺度である場合の多次元尺度法

a) ミンコフスキー距離

s_{ij} は順序データであるため距離データに変換する必要があり，次のように表示される。

$$s_{ij} = \hat{d}_{ij} \cong d_{ij} = \left[\sum_{t=1}^{T} \left| x_{it} - x_{jt} \right|^{r}\right]^{\frac{1}{r}}$$

b) 適合の基準（ストレス）

$$S = \sqrt{\frac{\sum_{i<j} (d_{ij} - \hat{d}_{ij})^2}{\sum d_{ij}^2}}$$

この S の最小値を求めるためには，最急降下法を用いて S の値を最も減少させるような方向に x_{it} を逐次的に計算する。ここで最小となる x_{it} が決まったところで，Kruscal の方法を用いて，まず $\hat{d}_{ij}$ において隣り合った数の大小関係を調べ，少なくとも1つは等号が成立していない部分を捜し，その部分にその大小関係の異なる部分の平均値をあてることによって，S_{ij} と d_{ij} の順序関係を満たすような点間距離（$\hat{d}_{ij}$）が推計される。

なお，この導出方法については，[23，第5章]で説明されている。また，同分析手法の解説書としては，[6]および[7]などがある。

9 数量化Ⅳ類

これは，多次元尺度法の融通的分析手法である。

非類似度から多次元ユーグリッド空間の配置を求めるために，次の関数を設定する。

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{m} e_{ij} d_{ij}^2$$

ただし，$d_{ij} = \sqrt{\sum_{t=1}^{p} (x_{it} - x_{jt})}$ ，　$e_{ij} = -S_{ij}$ であり，S_{ij} は非類似度を示す。

ここで，Q の大小を比較するために，x_i の平均をゼロ，分散1として，この制約のもとで Q を最大化することを考え，固有方程式を解くことによって，x_i が導出される。なお，この分析については，初歩的な解説書としては，[12, 第9章]，一般的なものとしては，[22, 第15章]，[23, pp.103-105]および[27, 第5章]などがある。

10　クラスター分析

この分析手法は，各個体の類似性にもとづいて，いくつかにグループ化するときに用いられる。大きくは，以下の2つに分析方法が分けられる。

(1)階層的方法

この方法は，主として次の6つの方法に分けられる。

最短距離法：　2つのクラスターに属する個体間の距離の最大値を採用

最長距離法：　2つのクラスターに属する個体間の距離の最大値を採用

群　平　均：　2つのクラスターにまたがる個体間の距離の二乗の平均値を採用

メジアン法：　2つのクラスターにまたがる個体間の最短と最長の中間距離を採用

重　心　法：　メジアン法に個体数を考慮したもの

ウォード法：　クラスター内平方和が最小となるように融合

なお，上記分析手法については，初歩的なものとして[3]，[14]および[17]，一般的なものとしては[22]などによって説明されている。

(2)非階層的方法

この方法は，最適化型手法とも呼ばれ，代表的なものとしては，k-means法

などがある。この手法は，多変量解析ソフト Systat，Statistica などにも採用されている。なお，これについては[22，pp.41-43]および[49，pp.243-247]を参照せよ。また，[49，pp.247-249]にはクラスター分析を使用する場合の注意が記してある。

11　正準相関分析

同分析手法は，2つの特性値のグループ間の関連とその強さについて分析するために用いられる。

まず，合成変数を $f = w'x$，$g = v'y$ と表示すると，x と y の相関係数は $R_{xy}v$ と表され，f と g の相関係数は $w'R_{xy}v$ となる。ついで，これを最大にするような w および v を求めると，

$\left| R_{xy}R_{yy}^{-1}\ R_{yx} - \lambda R_{xx}w \right| = 0$ から，w は最大固有値に対する固有ベクトルであり，

$v = \frac{1}{\sqrt{\lambda}}\ R_{yy}^{-1}\ R_{yx}w$ が導出される。

なお，同分析とコレスポンデンス分析との関係については，[11, pp.119-121]に記されている。

Ⅲ　おわりに

本研究ノートでは，忘れかけた分析手法を思い出すことと（そのため，数式および変数などの詳細な説明は省略してある），それに対してどの文献を読むとすぐに要点がとらえられるかが，一目で分かるようにまとめたものである。多変量解析に関する筆者所有の文献（以下の参考文献）を見ると，様々なレベルがあり，筆者には理解できても，読者には理解しにくいところが，または逆も多々あると思われる。各々の文献はそれぞれ特徴を持っており，おそらく著者のメインとする分野に依拠しているのであろう。もちろん，これについて，各分析に使われたページ数や統計用語数，研究分野などのデータを調べて，それらのデータに対して多変量解析を応用すると興味深い結果も得られるであろう。文献を速読して分かったことであるが，高度な数学や統計手法を使ったものも理解しがたいが，逆に数学や統計的手法もほとんど使わず，入門的なものも理解するのに頭を痛めた。本を書くのは実に難しいものだということが分かった。今日では，特に高価な統計ソフトを購入しなくとも，表計算ソフトExcelやWingzなどを用いて統計解析などできるようになった。Excelにおいては，多変量解析のアプリケーションやマクロ解析のプログラム（例えば，エクセル統計（(株) 社会情報サービス）（解説書としては[16]））および[32]）など市販されているものがある。一方Wingzは行列の計算などが容易なため，連立方程式モデルや産業連関分析などに即利用できる。道具は整いつつある。今後は，筆者を含め，若い研究者の理論モデル構築への熱意が期待される。

参考文献

[1] A.L.Comrey, *A First Course in Factor Analysis*, Academic Press, 1973（芝祐順『因子分析入門』サイエンス社，1979）

[2] B.Flury and H.Riedwgl, *Multivariate Statistics*, Gustav Fischer Verlag, 1983（田畑吉雄訳『多変量解析とその応用』現代数学社，1990）

[3] B.F.J.Manly, *Multivariate Statistical Analysis*, Chapman and Hall, 1986（村上正康・田栗正章共訳『多変量解析の基礎』培風館, 1992）
[4] B.W.Bolch and C.J.Huang, *Multivariate Statistical Methods for Business and Economics*, Prentice-Hall, 1974（中村慶一訳『応用多変量解析』森北出版, 1976）
[5] D.F.Morrison, Multivariate Statistical Methods,2ed., McGraw-Hill, 1976
[6] J.B.Kruskal and M.Wish, *Multidimensional Scaling*, Sage, 1978（高根芳雄訳『多次元尺度法』朝倉書店, 1980）
[7] P.Arabie,J.D.Carroll and W.S.DeSarbo, *Three-Way Scaling and Clustering*, Sage, 1987（岡太彬訓・今泉忠『3元データの分析ー多次元尺度構成法とクラスター分析法』共立出版, 1990）
[8] R.J.Johnston, *Multivariate Statistical Analysis in Geography*, Longman. 1978
[9] S.L.S.Jacoby,J.S.Kowalik and J.T.Pizzo, *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, 1972（関根智明訳『非線形最適化問題の反復解法』培風館, 1976）
[10] S.M.Kendall, *Multivariate Analysis*, 2nd, Charles Griffin, 1975（奥野忠一・大橋靖雄共訳『多変量解析』培風館, 1981）
[11]朝野熙彦『入門多変量解析の実際』講談社サイエンティフィク, 1996
[12]有馬哲・石村貞夫『多変量解析のはなし』東京図書, 1987
[13]石村貞夫『すぐにわかる多変量解析』東京図書, 1992
[14]石村貞夫『グラフ統計のはなし』東京図書, 1995
[15]伊藤孝一『多変量解析の理論』培風館, 1969
[16]内田治『すぐにわかる EXCEL による多変量解析』東京図書, 1996
[17]圓川隆夫『多変量のデータ解析』朝倉書店, 1988
[18]大村平『多変量解析のはなし』日科技連, 1985
[19]奥野隆史『計量地理学の基礎』大明堂, 1977
[20]奥野忠一・久米均『多変量解析法』日科技連, 1971
[21]河口至商『多変量解析入門Ⅰ』森北出版, 1973
[22]河口至商『多変量解析入門Ⅱ』森北出版, 1978
[23]河口至商・水田正弘『多変量グラフ解析入門』森北出版, 1978
[24]菅民郎『初心者がらくらく読める多変量解析の実践（上）』現代数学社, 1993
[25]菅民郎『初心者がらくらく読める多変量解析の実践（下）』現代数学社, 1993
[26]久米弘・高梨一彦『実務的 SPSS による多変量解析法』高文堂, 1993
[27]小林龍一『数量化理論入門』日科技連, 1981

[28]駒澤勉・橋口捷久『パソコン数量化分析』朝倉書店，1988
[29]古屋野亘『数学が苦手な人のための多変量解析ガイド』川島書店，1988
[30]清水利信・齋藤耕二『(改訂増補) 因子分析法』日本文化科学社，1976
[31]新村秀一『パソコンによるデータ解析』講談社ブルーバックス，1995
[32]杉山和雄・井上勝雄『EXCEL による調査分析入門』海文堂，1996
[33]杉山高一・千葉芳雄・吉岡茂『応用多変量解析』インフォメーションサイエンス，1986
[34]鈴木栄一『環境統計学』地人書館，1979
[35]田中豊・脇本和昌『多変量統計解析法』現代数学社，1983
[36]瀧好英『経済分析のための因子分析法』国元書房，1978
[37]林知己夫『行動計量学序説』朝倉書店，1993
[38]本田正久・島田一明『経営のための多変量解析法』産能大学出版，1977
[39]本田正久『インフォメーション・アナリストのための多変量解析の実際』産能大学出版，1993
[40]松田紀之『質的情報の多変量解析』朝倉書店，1988
[41]村上征勝・田村義保編『パソコンによるデータ解析』朝倉書店，1988
[42]安田三郎・海野道郎『改訂２版社会統計学』丸善，1983
[43]安本美典・本多正久『因子分析法』培風館，1981
[44]柳井晴夫・岩坪秀一『複雑さに挑む科学ー多変量解析入門』講談社ブルーバックス，1971
[45]柳井晴夫・高根芳雄『多変量解析法』朝倉書店，1977
[46]柳井晴夫・繁桝算男他『因子分析ーその理論と方法』朝倉書店，1990
[47]柳井晴夫・岩坪秀一他編『人間行動の計量分析ー多変量データ解析と応用』東京大学出版会，1990
[48]柳井晴夫『多変量データ解析法』朝倉書店，1994
[49]鷲尾泰俊・大橋靖雄『多次元データの解析』岩波書店，1989
[50]渡部洋『心理・教育のための多変量解析法入門（基礎編)』福村出版，1988