多変量行動データの統計解析法の研究開発

研究室紹介

足　立　浩　平*

Developments of Statistical Techniques
for Analyzing Multivariate Behavioral Data

Key Words：Statistics, Multivariate Analysis, Matrices, Sparse Analysis, Behavioral Science

1．注目される統計学の中で

理科・文科系の相違を問わず，データから結論を導く研究では，統計学に基づくデータ解析を使うことが定番になっています．学究の世界だけでなく，世に溢れるデータを統計解析すればビジネスにも役立つため，産業の世界でも統計学が着目されつつあり，その一端は，好評を得た西内氏の著書[1]のタイトルにも見受けられます．

著者は行動統計科学という研究室に所属し，人間行動に関するデータを念頭においた統計解析法の研究開発に携わっております．ここで，強調したいのは，開発される解析法は数学に基づきますので，種々の分野で利用可能な汎用性を持つことです．本研究室は，統計学の中でも特に多変量データの統計解析法を扱っていますが，そこで，行列（matrix）の数理を扱う線形代数が重要な基礎となります．著者は，英語で行列ベースの多変量解析入門の授業を担当していますが，その資料をまとめた英文の拙著[2]が昨秋発刊されました（図1）．

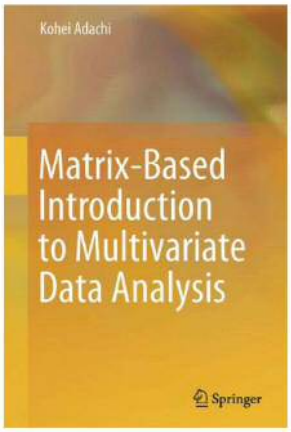

図1．多変量解析の英語授業のためのテキスト[2]

2．研究開発の流れ

「科学とは何か」という大げさな問いに簡単に答えれば，「現象＝理論＋誤差の誤差を小さくする理論を見出すこと」と言えます．著者の分野では，下線部の式の現象がデータに，理論がモデルに相当し，「データ＝モデル＋誤差」と書き換えられます．この式から出発して，新たな統計解析法の開発の流れを以下に概説します．

モデルは，一般に，未知の数の集まりΘ（パラメータ）の関数$f(\Theta)$表せ，上の式は，「データ＝$f(\Theta)$＋誤差」と書き換えられます．$f(\Theta)$を具体的な式で表すことをモデル構成と呼び，研究の出発点となります．どのような$f(\Theta)$を考えるかは，データ解析の目的によります．データが個体×変数の行列$\mathbf{X}$で，変数をより少数の成分で説明したい場合には，$\mathbf{X}$より列数の少ない行列$\mathbf{F}$, $\mathbf{A}$を使って，$f(\mathbf{F}, \mathbf{A}) = \mathbf{F}\mathbf{A}^{\mathrm{T}}$とすることが考えられ，これは主成分分析のモデルです．

モデル構成の次に，「データ＝$f(\Theta)$＋誤差」すなわち「誤差＝データ$-f(\Theta)$」の誤差の大きさを最小化するパラメータΘを求める方法，つまり，解析法を考えます．ここで，誤差の大きさϕ（データ$-f(\Theta)$）を定義には，最小二乗法や最尤法が使われます．しかし，ϕ（データ$-f(\Theta)$）を最小にするΘを数式で表せないことがあり，反復計算を考えるという課題が生じます．この課題は，t回目の反復

* Kohei ADACHI
1958年11月生
京都大学　文学部　哲学科（心理学専攻）（1982年）
現在、大阪大学　人間科学研究科
行動生態学講座　行動統計科学研究分野
教授　博士（文学）　多変量データ解析
心理統計学
TEL：06-6879-4040
FAX：06-6879-4040
E-mail：adachi@hus.osaka-u.ac.jp

で得られるΘの値をΘ_tと表すと，ϕ（データ−$f(\Theta_t)$）≥ϕ（データ−$f(\Theta_{t+1})$）となるように，Θ_tをΘ_{t+1}に更新する式を見出す作業に帰着し，この作業をアルゴリズム構成と呼びます．上記の不等式は，更新のたびに誤差の大きさが単調減少することを表し，いずれはΘの解にたどり着けることが期待できます．

前段までの紙とペンによる作業の次は，計算機上で解析法をプログラミングし，その挙動をシミュレーションによって評価します．その手続きは，Θを特定値Θ_{TRUE}に定めて，モデル構成の式「データ＝$f(\Theta_{\mathrm{TRUE}})$＋誤差」の誤差を乱数で生成して，人工データを発生させ，それを解析するというものです．解析結果のΘの解がΘ_{TRUE}に近ければ，解析法は信頼できるといえます．その後は，解析法を実際のデータに適用して，有用性を例証します．

3．スパース多変量解析法の研究開発

パラメータの解の行列がスパース，すなわち，多くの0要素を含むような解析法が最近注目され，ここでは，考案した3種のスパース解析法を紹介します．

最初は，入力×出力×ソースの三相テンソル・データの解析法[3]です．この方法を，例えば，30名の人（ソース）が11種の色（入力）の印象を9つの形容語（出力）によって答えた結果に適用すると，図2の4層ネットで表せる解が得られます．ここで，中間の2層がいわば心に潜在する構造を表し，その左の層が入力に応答するセンサー，右が出力を生成するモチーフに当たります．解析法の基礎になるのは，ソースk（＝1, …, K）の入力×出力のデータ行列$\mathbf{X}_k$を，それよりサイズが小さい$\mathbf{H}_k$を使って，$\mathbf{AH}_k\mathbf{B}^\top$とモデル化する三相主成分分析ですが，問題は，$\mathbf{AH}_k\mathbf{B}^\top=\mathbf{AS}^{-1}\mathbf{SH}_k\mathbf{T}^\top\mathbf{T}^{\top-1}\mathbf{B}^\top$のように，$\mathbf{AS}^{-1}$，$\mathbf{BT}^{-1}$，$\mathbf{SH}_k\mathbf{T}^\top$もそれぞれ$\mathbf{A}$，$\mathbf{H}_k$，$\mathbf{B}$に相当する解と見なせることです．そこで，$\mathbf{S}$と$\mathbf{T}$を同定するため，$\mathbf{AS}^{-1}$と$\mathbf{BT}^{-1}$がスパース行列を近似すると同時に，$\mathbf{SH}_k\mathbf{T}^\top$（$k$＝1, …, K）をソースに共通の$\mathbf{H}$が要約するというのが，開発手法の骨子です．

次は，個体×変数のデータ行列$\mathbf{X}$の変数間の因果関係の発見を目ざすスパース・パス解析で[4]，オフィス評価データ[5]に適用すると，図3のパス図の

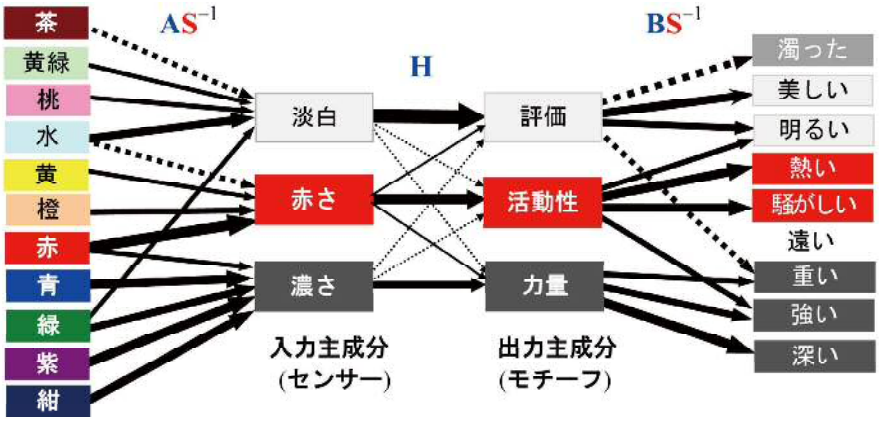

図2．色×形容語×人の印象評定データに対する三相主成分分析の解
（パスの太さが解の絶対値に比例し，実線・点線は正負に対応）

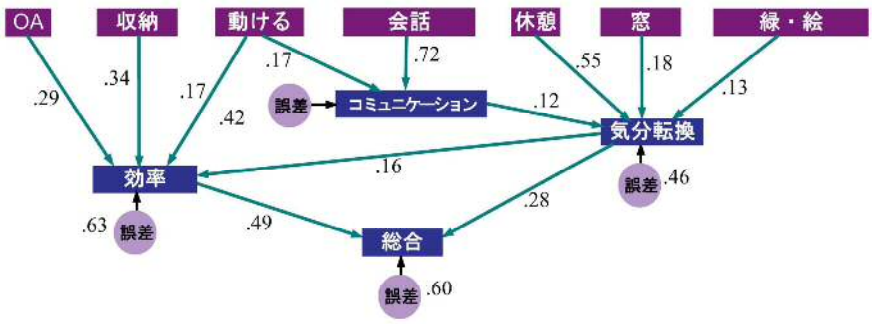

図3．オフィス評価データに対するスパース・パス解析の解

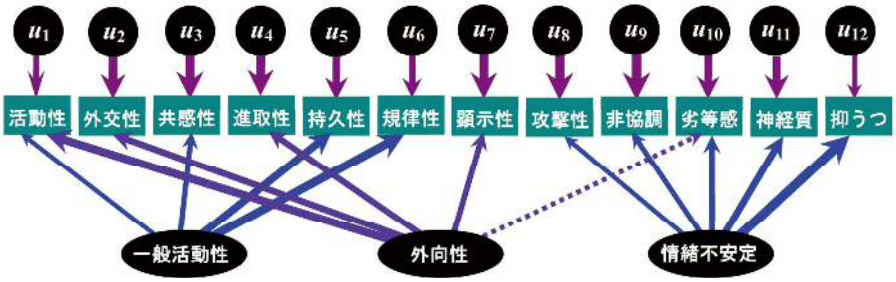

図4. 性格テストデータに対するスパース因子分析の解（パスの太さが解の絶対値に比例，実線・点線は正負に対応し，上部の●が独自成分）

解を与えます．この解析法のモデルは，$\mathbf{E}$ を誤差行列とすると，$\mathbf{X} = \mathbf{XB} + \mathbf{E}$ と書け，$\mathbf{B}$ が原因変数から結果変数へのリンクを表す係数行列ですが，その要素に零がなければ，全ての変数どうしがリンクする無意味な因果モデルとなります．そこで，$\mathbf{B}$ の中の零要素の位置をユーザーが決めていたのが，従来のパス解析ですが，零要素の位置と非零要素の値を同時推定するのが考案した解析法です．その解法は，誤差の大きさ ϕ $(\mathbf{X}-\mathbf{XB})$ に，$\mathbf{B}$ が非零要素を持つことを罰則するペナルティ関数を加えた関数を最小にする $\mathbf{B}$ を求めることと定式化できます．

最後に，ペナルティ関数を使わないスパース因子分析[6]を紹介します．因子分析は，複数変数とそれらの背後にある因子のリンクを表す行列 $\mathbf{\Lambda}$ と因子とは関係しない変数独自の成分を表す行列 $\mathbf{\Psi}$ を求める方法ですが，その目的関数が，$\mathbf{\Lambda}$ に関係しない行列 $\mathbf{C}$ を用いて，$\phi\,(\mathbf{X},\ \mathbf{\Lambda},\ \mathbf{\Psi}) = \phi\,(\mathbf{X}, \mathbf{C}, \mathbf{\Psi}) + n\|\mathbf{C}-\mathbf{\Lambda}\|^2$ のように分割できることに，開発手法では着目します．$\mathbf{\Lambda}$ に関わる項は，単純な関数 $n\|\mathbf{C}-\mathbf{\Lambda}\|^2$ だけですので，スパースにしたい $\mathbf{\Lambda}$ の零要素数を好きな整数に設定した上でのアルゴリズム構成が可能になります．以上の解析法を性格テストのデータ[7]に適用した結果を図4に示します．図の楕円のように解釈できる因子が，テストの項目の背後にあることがわかります．以上と同様の着想に基づくスパース主成分分析[8]も開発しています．

4．著者の興味の変遷

著者は高校時代，歴史が好きだったため文学部を志望しましたが，受検に失敗した後の浪人中に，同じ文学部でも心理学専攻に志望を変えて，入学後に，使う側の立場として統計学に出会いました．そのうち，心理データの解析法を開発する心理統計学に興味が移った後，いつの間にか，どっぷり統計学の世界に入り，今は日本計算機統計学会という理工学系の学会で副会長をしています．「興味は変わる」のか「自分の興味の発見に時間がかかる」のか，答えは出ません．

引用文献：

1) 西内　啓 (2013). 統計学が最強の学問である．ダイヤモンド社
2) Adachi, K. (2016). *Matrix-based introduction to multivariate data analysis. Springer.*
3) Adachi, K. (2011). Three-way Tucker2 component analysis solutions of stimuli × responses × individuals data with simple structure and the fewest core differences. *Psychometrika*, **76**, 285-305.
4) Adachi, K. (2014). Sparse path analysis : Computational identification of causality between explanatory and dependent variables. 日本計算機統計学会第28回シンポジウム講演論文集, 223-226.
5) 小島隆矢 (2003). Excel で学ぶ共分散構造分析とグラフィカルモデリング．オーム社．
6) Adachi, K., & Trendafilov, N. T. (2014). Sparse orthogonal factor analysis. In M. Carpita, E. Brentari, & E. M. Qannari (Eds.), *Advances in latent variables*, pp. 227-239. Springer.
7) Yanai, H., & Ichikawa, M. (2007). Factor analysis. In C.R. Rao & S. Sinharay (Eds) *Handbook of statistics vol. 26*, (pp. 257-296). Elsevier.
8) Adachi, K. & Trendafilov, N. T., (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, **31**, 1403-1427.