岡山理科大学紀要 第43号 A pp. 87-93(2007)

交互最小二乗法による質的データの変数選択への試み

森　裕一・松本喜彦*・飯塚誠也**・黒田正博
岡山理科大学総合情報学部社会情報学科
* 岡山大学大学院環境学研究科
** 岡山大学環境理工学部環境数理学科

(2007年10月1日受付、2007年11月2日受理)

1　はじめに

主成分分析のような外的変数をもたない多変量解析が適用される多くの応用場面において，変数選択は有用である．

これまで，主成分分析における変数選択の研究としては，Jolliffe (1972)，Robert and Escoufier (1976)，Krzanowski (1987)，Mori et al. (2004) などがある．これらの研究は，基本的に，元の全変数から求められる主成分に対して，できるだけ多くの情報をもつように一部の変数から通常の主成分を抽出しようというものである．これに対して，Tanaka and Mori (1997) は，拡張主成分分析 (Modified Principal Component Analysis, M.PCA) と呼ばれる手法を提案している．この手法は，選択された変数だけでなく，選択されなかった変数の情報も再現するものである．この M.PCA は，その計算の中に変数選択手順を自然に含むので，この規準を妥当な変数群を見つけることに直接利用できる．このようにして見つけた主成分は，高い妥当性をもち，実際の適用が容易な多次元の評価指標を提供してくれることになる．

ところで，前述した手法のほとんどは，量的データを扱う手法である．一方では，質的データに対して変数選択を行う場合も実際には多く存在する．これに対して，コレスポンデンス分析における変数選択 (たとえば，Iizuka et al., 2002) などが適用可能である．しかし，選択された変数の情報だけでなく，選択されなかった変数の情報も含むような変数選択規準はまだない．

そこで，M.PCA を質的データに対応できるように拡張することを考える．すなわち，質的データの数量化と M.PCA を同時に実行しようというものである．ここでは，交互最小二乗法 (Alternating Least Squares, ALS) の考え方 (PRINCIPALS, Young et al., 1978) を利用して，数量化と M.PCA を同時に実行する．この手法の一部は，Mori et al. (1997) でも提案されているが，ここではその手法をさらに進め，変数選択手順の考案と選択結果等について考察する．

2　拡張主成分分析

M.PCA（Tanaka and Mori, 1997）について簡単に触れておく．

量的データ Y(n 個体，p 変数) が得られているとする．この Y の p 個の変数のうち，q 個 $(1 \leq q \leq p)$ の変数の線形結合により得られる r 個 $(1 \leq r \leq q)$ の主成分で，元のデータ（Y）をできるだけよく予測する，すなわち，このときの主成分は，q 個の変数の情報を基にしながらも，残りの $p-q$ 個の変数の情報も取り込んだものとして推定しようというものである．

ここで，Y_1 を q 個の変数をもつ Y の部分行列，Y_2 を残りの $p-q$ 個の変数をもつ部分行列とし，$Y=(Y_1, Y_2)$ とする．A を q 変数に対するウエイト行列とすると，Y_1 による r 個の線形結合は，$Z=Y_1A$ となる．これが元の変数（p 変数）を最もよく表すように A を推定する．

A を求める規準として，次の 2 つを用いる．

(C1) 線形結合 Z を用いて Y の予測効率を最大にする (Rao, 1964).

(C2) Y と Z の間の RV 係数 (Robert and Escoufier, 1976) を最大にする．

(C1) では，予測効率の最大値 P は，$P=\sum_{i=1}^{r}\lambda_i/tr(S)$ として得られる．(C2) では，RV 係数は，$RV=\left\{\sum_{i=1}^{r}\lambda_i^2/tr(S^2)\right\}^{1/2}$ として得られる．ただし，λ_i は，$Y=(Y_1, Y_2)$ の分散共分散行列を $S=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$

としたとき，次の一般化固有値問題から得られる i 番目の固有値である．

$$[(S_{11}^2 + S_{12}S_{21}) - \lambda S_{11}]\boldsymbol{a} = 0 \tag{1}$$

ここで得られる q 個の固有値を大きい順に $\lambda_1, \lambda_2, \ldots, \lambda_q$，対応する固有ベクトルを $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_q$ とすれば，目的の A の解は，$A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ として得られる．このとき，寄与率 P は，q 個の変数から求められる最初の r 個の主成分によって説明される元の変動の割合であり，RV 係数は，元の p 変数の布置と r 個の主成分の布置の近さを示すものとなる．

ここでは，(C1) の寄与率 P による M.PCA の規準を変数選択に利用する．すなわち，p 個の変数のうち Y_1 として考えられる q 個の組み合わせの中から，この寄与率 P を最大にする変数群を選ぶもので，そのときの寄与率 P が最適な規準値ということになる．

3 交互最小二乗法による質的データの数量化手法

3.1 最適変換

X を n 個体，p 変数の質的データ行列とする．このとき，j 番目の変数 $\boldsymbol{x}_j$ が $1, 2, \ldots, c_j$ のようにラベル付けされた c_j 個のカテゴリーをもつとする．X を数量化するために $\boldsymbol{x}_j$ を次のような要素をもったダミー変数 G_j に置き換える．

$$g_{ijk} = \begin{cases} 1 & (x_{ij} = k) \\ 0 & (x_{ij} \neq k) \end{cases} \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p;\ k = 1, 2, \ldots, c_j)$$

w_{jk} を G_j の k 番目の最適なカテゴリースコアとすると，$y_{ij} = \sum_{k=1}^{c_j} w_{jk} g_{ijk}$ により，$n \times p$ の数量化行列 Y が X の最適変換行列として得られる．ここでは，この w_{jk} を次の 3.2 節のように推定する．

3.2 交互最小二乗法

質的データに対する数量化の手法，つまり最適なカテゴリースコア w_{jk} と M.PCA におけるウエイト行列 A の推定のために，ALS によるあらゆる尺度に対応した主成分分析（PRINCIPALS, Young et al., 1978）を適用する．

Z を (1) 式で得られる n 個の主成分行列 $(n \times r)$，B を係数行列 $(r \times p)$ として，

$$\hat{Y} = ZB \tag{2}$$

としておく。Y に w_{jk}，Z に A が含まれている状況で，$\hat{Y}$ が Y を最もよく予測するように，この B を決めたい．このとき，最適化の規準を，

$$\theta = tr(Y - \hat{Y})^{\mathsf{T}}(Y - \hat{Y}) \tag{3}$$

とする．これに対して，ALS による 2 つのパラメータの推定手順は以下のようになる．

(step0) 初期値設定：

Y の初期値を与える．元の質的データでもよいし，乱数でもよい．この Y を変数ごとに標準化しておく．

(step1) モデル推定：

$Y = (Y_1, Y_2)$ として M.PCA を適用し，固有値問題 (1) を解き，大きい方から r 個の固有値に対応する固有ベクトルとして A を求める。これより，主成分得点 $Z = Y_1 A$ と係数ベクトル $B = (Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}Y$，さらに (2) 式より $\hat{Y}$ を求める。

(step2) 収束判定：

(3) 式により，前の θ と現在の θ を比較し，実質的変化が見られないときは収束したとみなし，ここで止める．そうでなければ，次に進む．

(step3) 最適変換：

step2 で求めた $\hat{Y}$ を固定して，(3) 式の θ を最小化するようにカテゴリースコア w_{jk} を求め，最適変換行列 Y を求める．この最適変換は，平均 0，分散 1 の条件の下で，各変数ごとに行う．

step1 から **step3** までを収束するまで繰り返す．

4 変数選択手法

最適な変数群は，その時の q に対する全ての変数の組み合わせについて寄与率 P を計算し，その中で最大の P の値を提供する変数群として得られる．しかし，この方法は，全体の変数の数 p が大きくなれば，それだけ計算コストも大きくなってしまう．そこで，変数減少法，変数減増法，変数増加法，変数増減法の 4 つの簡便法 (森 他, 1998) を用いる．

5 質的データに対する変数選択の手順

本研究で提案する質的データに対する M.PCA 規準を利用した変数選択手法は，次のとおりである．

(proc1)

3 節で述べた数量化を X に対して行い，数量化された行列 Y を得る．すなわち，$q := p$ として，w_{jk} と A を推定する．この結果を見て，次元数 r を特定する．

(proc2)

q を決める．変数減少法などの後退系の「選択手順」であれば，$q := p-1$，変数増加法などの前進系の「選択手順」であれば，$q := r$，ある特定の変数の数に着目するときは，その数を q とする．

(proc3)

「選択手順」に応じて，Y_1 の候補を決め，その候補 1 つ 1 つに対して，3 節の数量化を実行し，w_{jk} と A を推定し，寄与率 P を求める．これをすべての候補に対して実行し，その中で最も大きな寄与率 P を与える Y_1 をその q における最適な変数群とする．

(proc4)

q を更新する．後退系の「選択手順」では，$q := q-1$，前進系の「選択手順」では，$q := q+1$ とする．ある特定の変数の数についてのみ着目しているときは，ここで終了する．

(proc5)

前進系，後退系の「選択手順」においては，寄与率 P の値と変数の数 q が事前に決めた値より大きければ，**proc3** に戻り，そうでなければ，終了する．

上記の「選択手順」とは，4 節で述べた 4 つの逐次選択手法のことである．**proc3** では，この逐次選択の手順によって，検討すべき Y_1 の候補となる変数群が決まってくる．また，**proc2** や **proc4** の「ある特定の変数に着目するとき」とは，p 個の変数から q 個の変数を取り出す全ての組み合わせについて寄与率 P を求めることを意味する．

6 数値例

23 変数の検査項目をもつ「軽症意識障害 (Mild Disturbance Of Consciousness, MDOC, 佐野他, 1971)」のデータに，提案の手法を適用した例を示す．このデータは脳障害などの患者の軽症意識を調べようとするものであるが，佐野他 (1971) や Tanaka and Kodake (1981) などでは，解析には冗長である 2 変数を除いた 23 項目の妥当性の検討や検査項目の特定と同時に，23 項目と同等のスケールがより少ない項目で得られないかの研究がなされている．23 項目の検査項目は表 1 のとおりであり，これらの変数が 5 段階，もしくは 2 段階のカテゴリーで評価されている．

この MDOC データに対して，前述した変数選択手順により変数選択を行った．選択手順については，4 つの逐次選択手法を採用した．

それぞれの選択手法による寄与率 P の値の変化を図 1 に示す．4 手順での P の値の比較では，それぞれに

表 1 MDOC データ (解析に使用する 23 変数)

V1	食事	V2	尿失禁	V3	呼び名・挨拶への反応	V4	見当識 (場所)
V5	見当識 (季節)	V6	見当識 (月日)	V7	見当識 (時間)	V8	見当識 (人)
V9	自分の病識の程度	V10	意欲	V11	知識	V12	命令への反応
V13	1 から 20 まで数える	V14	計算力	V15	声の調子	V16	表情
V17	診察中の態度	V18	自発動作	V19	自発発語	V20	注意
V21	保持傾向	V22	生年月日が言える	V23	名前が言える		

大きな差はないが，変数減少法，変数増加法では他の 2 手順と比べて若干小さい値で推移していることがわかる．変数減増法，変数増減法では，選ばれた変数と P の値は同じ結果になった．このことから 4 手順のうち，変数減増法，変数増減法は，より情報を多くもつ変数群を選べているということになる．

次に，変数増減法による変数の選択結果および寄与率 P の変化を表 2 に示す．表 2 を見ると，変数の数が 12 個や 11 個のときの P の値は，全ての変数を使ったときの P の値とそれほど大きな違いはない．これにより，いくつかの変数 (たとえば，11 個や 12 個) を落としたとしても，主成分の情報量には大きな影響はなく，選ばれた変数に基づく主成分は，全 23 変数による主成分と大差ない情報をもつことを示している．実際に表 2 で選ばれた変数を見てみると，たとえば，$q = 10$ のときは，「V1:食事」，「V3:呼名・挨拶への反応」，「V4:見当識 (場所)」，「V5:見当識 (季節)」，「V6:見当識 (月日)」，「V11:知識」，「V12:命令への反応」，「V16:表情」，「V17:態度」，「V22:生年月日が言える」が選択されている．これらの変数は，拡張主成分の規準で最もよく元

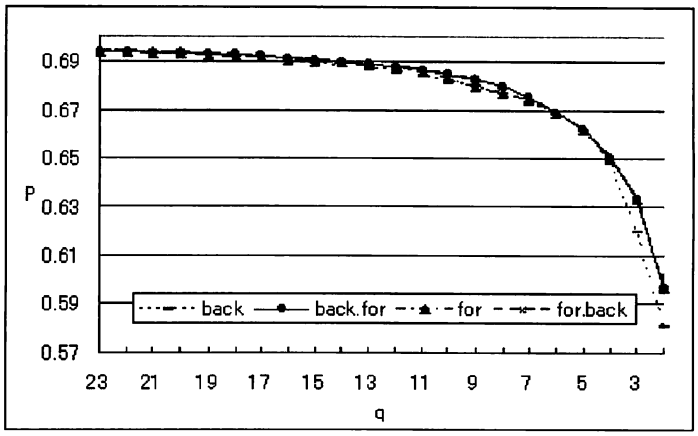

図 1　4 手順での P の値の変化
(MDOC データ, r=2, back:変数減少法, back.for:変数減増法, for:変数増加法, for.back:変数増減法)

表 2　変数選択過程 (MDOC データ，$r = 2$，変数増減法, x が選択された変数)

q	P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
23	0.69389	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
22	0.69353	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
21	0.69322	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	
20	0.69295	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x		x	x	x	x	
19	0.69267	x	x	x	x	x	x	x		x	x	x	x	x	x		x	x		x	x	x	x	
18	0.69233	x		x	x	x	x	x		x	x	x	x	x	x		x	x		x	x	x	x	
17	0.69188	x		x	x	x	x	x		x	x	x	x		x		x	x		x	x	x	x	
16	0.69112	x		x	x	x	x	x		x		x	x		x		x	x		x	x	x	x	
15	0.69018	x		x	x	x	x	x		x		x	x		x		x	x		x		x	x	
14	0.68928	x		x		x	x	x		x	x	x	x		x		x	x		x			x	
13	0.68841	x		x		x	x			x	x	x	x		x		x	x		x			x	
12	0.68732	x		x		x	x			x	x	x	x		x		x	x					x	
11	0.68610	x		x		x	x			x	x	x	x				x	x					x	
10	0.68448	x		x	x	x	x					x	x				x	x					x	
9	0.68216	x		x	x		x					x	x				x	x					x	
8	0.67950			x	x		x					x	x				x	x					x	
7	0.67488			x			x					x	x				x	x					x	
6	0.66834						x			x		x						x		x			x	
5	0.66171						x			x								x		x			x	
4	0.65009									x								x		x			x	
3	0.63336																	x		x			x	
2	0.59701																	x					x	

の変数を再現している 10 変数ということになる.

最後に，MDOC データを数量化せずにカテゴリーをそのままデータ値として変数増減法により変数選択を行ったときの P の値と，今回の手法により求められた寄与率 P の値 (変数増減法による) を図 2 で比較した. 全ての q に対して今回の手法の方が高い値を示している. これより，ALS による最適変換が数量化として適していることがわかる.

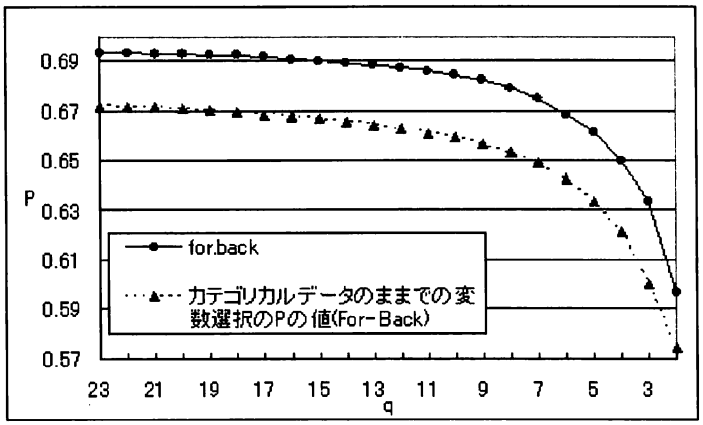

図 2　カテゴリカルデータのままでの変数選択結果との比較
(MDOC データ，$r = 2$，for-back:変数増減法)

7　まとめ

拡張主成分分析における規準を利用した変数選択を質的データに適用するために，ALS によるあらゆる尺度に対応した主成分分析のアイデアを利用して，数量化と M.PCA を同時に適用する手法を提案した. この手法では，選択されなかった変数の情報も含んで全体を再現させるという M.PCA の規準により変数選択を行っており，ALS による数量化も，数値例から，うまくパラメータの推定ができていることがわかった. また，これを実行する反復計算も，4 つの逐次選択の簡便法により，実用時間内に収めることができた.

今後の課題として，4 つの逐次選択手順の有用性を評価するために，全ての組み合わせによる選択結果との比較があげられる. また，計算上の問題として，反復計算における収束の判定およびそのアルゴリズムについて，検討・改良を加えていく必要がある. さらに，この手法の特徴をよりつかむために，先行研究で提案されている選択手法との比較も必要であると考える.

参考文献

Iizuka, M., Mori, Y., Tanaka, Y. and Tarumi, T.(2002). Some new modules in variable selection software VASMM. *Proceedings of the 4th ARS Conference of the IASC*, 166-169.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I. Artificial data. *Applied Statistics*, **21**, 160-173.

Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. *Appl. Statist.*, 36, 22-33.

Mori, Y., Fueda, K. and Iizuka, M. (2004). Orthogonal score estimation with variable selection in multivariate methods. In: Antoch, J (ed), *COMPSTAT2004 Proceedings in Computational Statistics*, 1527-1534, Phisica-Verlag.

Mori, Y., Tanaka, Y. and Tarumi, T. (1997). Principal component analysis based on a subset of variables for qualitative data. *Data Science, Classification, and Related Methods* (*Proceedings of IFCS-96*), 547-554, Springer-Verlag.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied reserch, *Sankhya*, A26, 329-358.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the *RV*-coefficient. *Appl. Statist.*, 25, 257-65.

Tanaka, Y. and Kodake, K. (1981): A method of variable selection in factor analysis and its numerical investigation. *Behaviormetrika*, **10**, 49-61.

Tanaka, Y and Mori, Y. (1997). Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis. *American Journal of Mathematics and Management Sciences*, 17, 1&2, 61-89.

Young, F. W., Takane, Y. and De Leeuw, J. (1978): The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279-81.

佐野圭司 他 8 名 (1977). 軽症意識障害の評価方法に関する統計的研究－断面調査による特徴的臨床像の抽出. 神経進歩, 1052-65.

森 裕一, 垂水共之, 田中 豊 (1998) . 変数の一部に基づく主成分分析－変数選択手法の数値的検討－, 計算機統計学, 11(1), 1-12.

A trial to variable selection for qualitative data using alternating least squares method

Yuichi MORI, Masahiko MATSUMOTO*, Masaya IIZUKA*
and Masahiro KURODA

Department of Socio-Information, Faculty of Informatics
Okayama University of Science
1-1 Ridai-cho, Okayama 700-0005, Japan
* *Graduate School of Environmental Science*
Okayama University
3-1-1 Tsushima Naka, Okayama 700-8530, Japan

(Received October 1, 2007; accepted November 2, 2007)

A variable selection method using criteria in Tanaka and Mori's modified principal component analysis selects a reasonable subset of quantitative variables that provides principal components which are computed using only a selected subset of variables but which represent all of the variables, including those not selected, as much as possible. In this paper modified principal component analysis is extended so that it can deal with qualitative data with unordered or ordered categories. Namely both quantification of qualitative data and modified principal component analysis are performed at the same time. To do this the iteration technique based on the alternating least squares method by Young et al.'s PRINCIPALS is used for quantification at every step of finding the best subset. Example data is analyzed to demonstrate the performance and usefulness of the proposed method, in which four cost-saving selection procedures are compared for real data analysis.

Keywords: principal components; iterative quantification; alternating least squares; cost-saving selection procedures.