拡張主成分分析における総合指標の評価

飯塚誠也*, 森　裕一**
*岡山大学法学部, **岡山理科大学総合情報学部

1. はじめに

調査や検査において, 調査項目や検査項目をその全体的な様相をできるだけ保持した形で表現したい場合がある。このような状況において, 拡張主成分分析（Modified Principal Component Analysis, 以下 M.PCA と略す）を適用することができる。M.PCA とは, 元のデータの一部を用いて元の全変数を最もよく代表するような主成分を推定するという目的で開発されたものである。言い換えれば, 少ない次元でデータの特徴を測りとれる指標を作るが, そこでは, 全変数のうち少数の変数だけを利用するものである。

M.PCA の先行研究には, 以下のようなものがある。

- Tanaka and Mori (1997) ：M.PCA の提唱とその性質の考察や感度分析などを行っている。
- 森 他(1998) ：感度分析の方法を規準値を変数選択に利用した場合の数値的な検討がなされている。
- 森 他, (1998)； 森, (1998)：次節で述べる寄与率 P や RV 係数値に関して, M.PCs の方が O.PCs よりも高い値を得ることを示している。
- Mori at el. (1999), Iizuka at el. (2000) ：M.PCA や主成分分析における変数選択で用いられる手順（逐次変数選択法）や規準（寄与率 P や RV 係数）の特徴や安定性について評価を行っている。

M.PCA により求められる主成分得点（M.PCs）は, 線形結合としては, 元の変数 Y の一部の変数 Y_1しか使わないが, 実際は利用していない残りの変数 Y_2の情報も含んでいる。したがって, Y_1に対して通常の主成分分析（Ordinary PCA, O.PCA と略す）を適用して得られる主成分得点（O.PCs）と比較すると, 元の変数に関する情報をより多く保持した主成分得点が得られていることが期待される。しかしながら, 元の変数がもつ情報や潜在的な構造を M.PCs がどれだけ再現しているかについては, まだ正確な評価はされていない。

そこで, M.PCA の性能を評価するために, 人工データによるシミュレーションを行う。具体的には, r因子モデルを数パターン作り, これに対して, M.PCs と O.PCs がどれだけ本来の潜在因子を再現できているかについて調べることにする。これにより, 因子負荷行列のパターン, 選択される変数の数などによる M.PCA の性能の特徴を, シミュレーションを用いて明らかにし, どのような場合で M.PCA が有効に作用するかなどを考察する。

2. 拡張主成分分析

2.1 拡張主成分分析と通常の主成分分析

Yをn個の個体とp個の変数をもつデータ行列とする。この Yを q 個の変数をもつ $n\times q$ 部分行列 Y_1と残りの $p-q$ 個の変数をもつ $n\times(p-q)$部分行列 Y_2に分割したとき, M.PCA は, この Y_1による r 個の線形結合 $Z=Y_1A$ が元の p 個の変数を最もよく代表するように $A = (\boldsymbol{a}_1, ..., \boldsymbol{a}_r)$ を推定するものである$(1<r<q)$。ここで, その規準として,

(1) 線形結合 $\boldsymbol{z}$ を用いて $\boldsymbol{y}$ の予測効率を最大にする
(2) Y と Z の RV 係数を最大にする

という 2 つの規準が採用できる。この規準値は，$Y = (Y_1, Y_2)$ 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，一般化固有値問題 $[(S_{11}^2 + S_{12}S_{21}) - \lambda_j S_{11})]\boldsymbol{a}_j = 0$ を解いて得られる大きい方から r 個の固有値 $\lambda_1, \lambda_2, \ldots, \lambda_r$ を用いて，(1)の規準値はその和で表される寄与率 $P = \sum_{i=1}^{r} \lambda_i / tr(S)$ と，(2)の規準値はその 2 乗和で表される RV 係数 $RV = \left\{ \sum_{i=1}^{r} \lambda_i^2 / tr(S^2) \right\}^{1/2}$ となる。また，求める係数ベクトル A は，この固有値 $\lambda_1, \lambda_2, \ldots, \lambda_r$ に対応する $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r)$ であり，M.PCs は $Z=Y_1A$ で求められる。これを Z_M と表しておく。

一方，O.PCA では，Y_1 が得られている場合，O.PCs は，固有値問題 $(S_{11} - I)\boldsymbol{v}_j = 0$ を解いて得られる固有ベクトル $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r)$ を用いて，$Z=Y_1V$ で求められる。これを Z_O と表しておく。

2.2 変数の選択

元のデータ Y を Y_1, Y_2 に分割するためには変数を選択し，Y_1 に割り当てる変数，Y_2 に割り当てる変数を決めなければならない。M.PCA は，M.PCs を得るだけではなく元の変数全体の情報を出来るだけ保持して Y_1 として最適な変数群を求める。つまり，p 変数のうち，与えられた q において，規準値である寄与率 P または RV 係数を最大にする q 個の変数の組合せを見つければよい。

その q 変数を見つけ出すためには，すべての組合せを調べる all possible を行うことが望ましいが，それを行うためには計算コストが非常にかかる。シミュレーション回数を増やす場合や，変数が多い場合，実行することが困難である。

幸い，このような高い計算コストがかかってしまうような場合でも，（森 他，1998）により，簡便な方法が提唱されており，

- 変数減少法(Backward)
- 変数増加法(Forward)
- 変数増減法(Backward-Forward)
- 変数減増法(Forward-Backward)

などを用いることにより，特定の q における変数の組合せを得ることができる。また，これら手順間の挙動の数値的な検討は森 他(1998)で行われており，all-possible とほぼ同様の結果が得られるということがわかっている。よって，計算コストが高くつくような場合では，これらの逐次選択手法を用い，変数を選択していくこととする。これにより，各 Y_1 においては，各規準にもとづいた最適な変数群が選ばれていることになる。

3. シミュレーションによる評価

3.1 シミュレーションの方法

M.PCA により，得られる主成分得点と.PCA により得られる主成分得点が潜在的な構造をどれだけ再現しているかを評価する方法として，次のようなシミュレーションを行う。

まず，シミュレーションにおいて，データ作成は，n 個体 p 変数の r 因子モデル $Y=FL'+E$ によって作成することとし，次のようにして作成する。

(1) 共通因子 F（$n \times r$）は $N(0,1)$の正規乱数により作成する。

(2) 独自因子 E $(n\times p)$ は, j 番目の列の分散を d_j^2 として, $N(0, d_j^2)$の正規乱数から生成する。
(3) Y の分散を σ_i^2 $(i=1,2,\ldots,n)$ で与えたとき, 因子負荷行列 L （$p\times r$） は, $\sigma_i^2=h_i^2+d_i^2$, $h_i^2=\Sigma a_i^2$
$Y'Y=LL'+E'E$ の関係より決定する。つまり, このシミュレーションにおいては Y の分散と独自分散を与えて因子負荷量行列 L を決定する。

上記により作成した F , L , E を用い, $Y=FL'+E$ の関係から Y を計算する。この方法で Y を繰り返し作成し, シミュレーションを繰り返す。

次に Y を Y_1, Y_2 に分割するために q を決め, 2.2 の方法で寄与率 P または RV 係数などの規準で変数選択を行い, Y_1 を決める。ここでは, 変数の数が多い場合にも対応できるよう, 変数選択手法として, 逐次選択手順の中で変数減増法を用い変数選択を行う。

ここで得られた Y_1 及び Y_2 を用い, M.PCA においては一般化固有値問題, O.PCA においては固有値問題をそれぞれ解き, 係数行列を求め, Z_M と Z_O を計算する。

そして, 得られたそれぞれの主成分得点が潜在構造をどれだけ再現しているかを見るために, F と Z_M の関係と F と Z_O の関係を評価する。ここで, この関係を評価する指標として,

- $r=1$ のときは相関係数
- $r\geq 2$ のときは, 行列間の近さを見る指標

を採用する。$r\geq 2$ の場合, 行列の近さを見る指標としては, 一般化決定係数, ベクトル相関係数, RV 係数, 一般化されたコーシー＝シュワルツの不等式による係数などがあるが, ここでは scale に依存せず, 回転も考慮されている RV 係数により判断することにする。ここで用いる RV 係数とは, $\tilde{F},\tilde{Z}$ は F と Z を中心化した行列であるとすると

$$RV(F,Z)=tr(\tilde{F}\tilde{F}'\tilde{Z}\tilde{Z}')/\{tr(\tilde{F}\tilde{F}')^2\cdot tr(\tilde{Z}\tilde{Z}')^2\}^{1/2}$$

で表されるもので, 行列 F と行列 Z の空間上の布置を測る規準である。

この評価する指標の絶対値の大小をみることになるが, 差が小さな値の場合, どちらも同様に潜在構造を再現していると考え, 判断の基準となる指標を C とすると, 大小の判断には $C(F,Z)$ が 0.0001 より大きい場合 M.PCs の方が潜在構造を再現しているとし, -0.0001 より小さい場合 O.PCs の方が再現しているとして, そして-0.0001 以上 0.0001 以下の場合, どちらも同様に再現している, と判断することにする。

以上の流れを一定のシミュレーション回数分 N 回繰り返し, 結果をみる。
したがって, シミュレーションとしては, 因子数 r を決めたとき,

- 因子モデルのパターン
- 利用する変数の数(q)

を変えることによって場合分けができ, それぞれの特徴を考察していく。

3.2 シミュレーション結果

3.2.1 １因子モデル

シミュレーションは個体数 n=1000, 変数 p=10, シミュレーション回数 N=100 で行うこととし,
$dif=|Cor(F,Z_M)|-|Cor(F,Z_O)|$とおく。
因子負荷行列が一定でない場合

因子負荷量	共通性
0.949	0.9
0.894	0.8
0.837	0.7
0.775	0.6
0.707	0.5
0.707	0.5
0.632	0.4
0.548	0.3
0.447	0.2
0.316	0.1

まず，独自分散 d_i^2 (i =1,2,…,p) を 0.1,0.2,0.3,0.4,0.5,0.5,0.6,0.7,0.8,0.9 で与える。そのとき，因子負荷量，及び共通性は左のようになる。つまり，共通性の値が 0 から 1 の間で均一の間隔で変化している場合，つまり各変数のもつ情報が一定ではない場合である。この場合において，シミュレーションを行うと以下のような結果が得られた。

	q							
判断	2	3	4	5	6	7	8	9
M.PCsの方が再現性大	100	100	100	100	100	100	100	91
	0	0	0	0	0	0	0	7
O.PCsの方が再現性大	0	0	0	0	0	0	0	2

共通性が一様でない場合では，M.PCs が O.PCs より潜在的な構造を再現しているということがわかる。q=9 の場合は，わずかに O.PCs が大きい場合もあるが， q=2,…,8 では，すべて M.PCs の方が大きな値である。このときの，変数の選択状況を見ると，共通性の小さい変数から，順に Y_2 に加えられている。言い換えると，Y_1 の変数から共通性の小さい変数が削除されている。q=9 では，Y_1 の変数から削除されている変数は共通性が小さく 0.1 であり，情報をあまり保持していないため，M.PCs で Y_2 の情報を取り込んだとしても，いくつかのものについては情報量として大きいものではなく，結局 O.PCs の方が大きな値をとったということが考えられる。

因子負荷量が一定の場合

次に独自分散 d_i^2 を 0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1 で与えシミュレーションを行う。そのとき因子負荷量，共通性は次のようになる。どの変数もほとんど同様の情報をもっている場合，つまり Y の分散の 9 割をすべての変数が説明している場合である。この場合では

因子負荷量	共通性
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9
0.949	0.9

	q							
判断	2	3	4	5	6	7	8	9
M.PCsの方が再現性大	3	3	1	0	0	0	0	0
	97	97	99	100	100	100	100	100
O.PCsの方が再現性大	0	0	0	0	0	0	0	0

という結果が得られる。

q=5,...,9 においては，M.PCs，O.PCs はすべて差がなく，どちらも同様の潜在的な構造を再現しているということがわかる。q=2,…,4 においても，わずかながら，M.PCs の方が大きい値をとっている場合もあるが，97 から 99 の場合で，ほとんど等しいという結果になっている。これは，すべての変数が同じような情報を保持していることから，M.PCA により Y_2 の情報を含んだ場合でも，O.PCA のように含まない場合でもどちらもほとんど同じ結果が得られるということを意味している。

このように，1 因子モデルの場合は，シミュレーションが示すように，データ解析の場面のような状況では，M.PCs の方が O.PCs よりも潜在的な構造を再現するという結果を示していることから，このようなデータに対して M.PCs が有効であるといえる。実際のデータ解析の場面では，このようなケースは考えにくいが，共通性が一定である場面では，M.PCs，O.PCs ともに同様の結果を示していることから，どちらを使っても差し支えはない。

3.2.2 2因子モデル

シミュレーションは個体数 n=1000, 変数 p=10, シミュレーション回数 N=100 で行うこととし, dif=｜$RV(F,Z_M)$｜-｜$RV(F,Z_O)$｜とおく。1 因子モデルの場合, M.PCA で得られた主成分得点が O.PCA で得られた主成分得点より潜在因子との相関の値は大きいという結果が得られたが, 2 因子モデル以上の場合, それらの大小の値は因子負荷量のパターン, およびその値に依存することが予想される。そこで, 2因子モデルでは, 因子負荷量のパターン, およびその値を変化させ, また, そのときの Y_1 に選択される変数を観察し, M.PCA 及び, O.PCA の挙動を見ることにする。

まず, 次のような因子負荷量のパターンを考える。各因子に 1 つずつ, 共通性の高いものがあり, 第 1 因子と第 2 因子が直交するケースである。

因子負荷量		共通性
1	2	
0.949	0.000	0.9
0.447	0.000	0.2
0.447	0.000	0.2
0.316	0.000	0.1
0.316	0.000	0.1
0.000	0.949	0.9
0.000	0.447	0.2
0.000	0.447	0.2
0.000	0.447	0.2
0.000	0.316	0.1

この場合,

	q						
判断	3	4	5	6	7	8	9
M.PCsの方が再現性大	100	100	100	100	100	100	100
	0	0	0	0	0	0	0
O.PCsの方が再現性大	0	0	0	0	0	0	0

という結果が得られ, このようなケースの場合でも M.PCs の方が高い値を得る。

また, 各軸において徐々に共通性が変わるような場合でも,

因子負荷量		共通性
1	2	
0.949	0.000	0.9
0.837	0.000	0.3
0.707	0.000	0.5
0.548	0.000	0.3
0.316	0.000	0.1
0.000	0.949	0.9
0.000	0.837	0.7
0.000	0.707	0.5
0.000	0.548	0.3
0.000	0.316	0.1

という結果が得られる。100 回繰り返したデータに対して, 各 q にお

	q						
判断	3	4	5	6	7	8	9
M.PCsの方が再現性大	100	100	100	100	100	100	98
	0	0	0	0	0	0	1
O.PCsの方が再現性大	0	0	0	0	0	0	1

いてほとんど, M.PCs の方がいい結果を示している。このように, 共通性が一定でない場合は 1 因子モデルの場合と同様に, 拡張主成分得点は潜在的な構造をより再現しているということができる。

ただし, 特に興味深い結果は次のようなパターンの時に起こる。

因子負荷量		共通性
1	2	
0.949	0.000	0.9
0.949	0.000	0.9
0.949	0.000	0.9
0.949	0.000	0.9
0.949	0.000	0.9
0.000	0.949	0.9
0.000	0.949	0.9
0.000	0.949	0.9
0.000	0.949	0.9
0.000	0.949	0.9

共通性が次のような場合である。このような場合について, シミュレーションを行うと結果に大きな違いが現れる。1因子モデルの場合と同様, 共通性が一定のときである。結果は, 表にあるように,

	q						
判断	3	4	5	6	7	8	9
M.PCsの方が再現性大	100	50	100	44	100	14	100
	0	49	0	56	0	86	0
O.PCsの方が再現性大	0	1	0	0	0	0	0

q が偶数の場合, M.PCs の持つ潜在的構造の情報と O.PCs の持つ潜在的構造の情報がほぼ等しくなる。また, q が奇数の場合, M.PCs の持つ潜在的構造が O.PCs よりも高くなるのである。各 q

において，Y_2 に加えられる変数を調べていくと，どのシミュレーションのケースでも 1 軸にある変数，2 軸にある変数が交互に選択されている。

下のような因子負荷量のパターンでは，シミュレーション結果は

因子負荷量		共通性
1	2	
0.794	0.520	0.9
0.700	0.458	0.7
0.592	0.387	0.5
0.458	0.300	0.3
0.265	0.173	0.1
0.520	0.794	0.9
0.458	0.700	0.7
0.387	0.592	0.5
0.300	0.458	0.3
0.173	0.265	0.1

	q						
判断	3	4	5	6	7	8	9
M.PCsの方が再現性大	100	100	100	100	100	99	95
	0	0	0	0	0	0	2
O.PCsの方が再現性大	0	0	0	0	0	1	3

となり，このような場合でも M.PCA が有効であるということが示唆される。通常のデータ解析で起こりうるような場合では，M.PCs は O.PCs よりもよりよい結果を得ることができるということができる。

参考文献

Iizuka, M., Mori, Y., Tarumi, T., Tanaka, Y. (2001.9), Computer intensive trials to determine the number of variables in PCA, Proceedings of the International Conference on New Trends in Computational Statistics with Biomedical Applications, 251-258

Mori, Y., Iizuka, M., Tarumi, T., Tanaka, Y. (1999.8),Variable Selection in “Principal Component Analysis Based on a Subset of Variables”, Bulletin of the International Statistical Institute (52nd Session Contributed Papers Book2) , 333-334.

Rao, C. R.（1964）. The use and interpretation of principal component analysis in applied research, Sankhya Ser. A, 26, 329-358.

Robert, P. and Escoufier, Y.（1976）. A unifying tool for linear multivariate statistical methods: The RV-coefficient. Applied Statistics, 25, 257-265.

Tanaka, Y and Mori, Y.（1997）. Principal component analysis based on a subset of variables: Variable selection and sensitivity analysis, Amer. J. Math. Management Sci., 17(1&2), 61-89.

森　裕一（1998）. 変数の一部に基づく主成分分析－RV 係数規準による数値的検討－，岡山理科大学紀要, 34（A）, 383-396.

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.

森　裕一，飯塚誠也，垂水共之，田中　豊 (2000). 変数の影響分析を利用した変数選択，日本行動計量学会第 28 回大会発表論文抄録集, 301-302.

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.

連絡先：飯塚誠也（岡山大学法学部） masa@law.okayama-u.ac.jp