一般化主成分分析について
– クラスター検出 –

2002MM011 橋本　登
指導教員　田中　豊

1　はじめに

主成分分析と正準判別分析は行列の固有値問題に定式化される．この問題はただ単に計量のとり方に違いがあるということに注目し, 目的に応じて計量のとり方を変える主成分分析 (一般化主成分分析という) を Cassinus らが提案している．今回の研究では一般化主成分分析を使って，クラスターがよく分かれるような計量のとり方をした，一般化主成分分析の結果と, 通常の主成分分析の結果, 群の情報がわかって正準判別分析をした結果を比較し, 一般化主成分分析は群の情報を用いずにどのくらい上手く群を分類できるかを研究する．なお，このために一般化主成分分析の新しい関数プログラムをRで作成する．

2　固有値問題

$$\text{主成分分析 } (T-\lambda I)\boldsymbol{a_1}=0 \tag{1}$$
$$\text{正準判別分析 } (B-\lambda_1 W)\boldsymbol{a_2}=0 \tag{2}$$
$$\Downarrow * T=B+W$$
$$(T-\lambda_2 W)\boldsymbol{a_2}=0 \tag{3}$$

B:群間平方和積和行列,W:群内平方和積和行列,T:総平方和積和行列
(1) と (3) の式を見てみると異なる部分は I と W である．この事より (1) は単位行列なので群の情報がわからない固有値問題．(3) は群内平方和積和行列なので群の情報がわかる固有値問題である.(3) は個体間の距離を計量行列 $\boldsymbol{W^{-1}}$を用いて$\boldsymbol{(x_i-x_j)'W^{-1}(x_i-x_j)}$ で測るときの個体間のばらつきの情報をできるだけ保存する形で低い次元の空間に射影する方法ともいえる．この事から群の情報を用いずの W に相当する行列を推定しようというのが一般化主成分分析の基本的な考え方である．

3　クラスター検出の一般化主成分分析

データを幾何学的に座標にあてはめた時, ある２点ｉとｊの差$\boldsymbol{x_i-x_j}$の積和行列を変形すると

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_i-x_j)(x_i-x_j)' \tag{4}$$
$$=n\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' \tag{5}$$

のようにデータの平方和積和行列のｎ倍になり, 分散共分散行列はｎ個の点 $\{x_i, i=1,\cdots,n\}$ のすべてのペアの差の積和から計算することができることがわかる．
よりよくクラスターを見つけられるようにするために Caussinus らは式 (5) のように同じ重みで和をとるのではなく，２点間の距離によって変わる重み

$$K(\boldsymbol{u})=exp(-h\boldsymbol{u}) \quad h>0 \tag{6}$$

をつけて和をとる方法を提案している.u は２点間の距離を表し, ｈは距離によって減衰させる程度を制御するパラメータである．ここでｈ=0.50,0.75,1.0 などのいくつかの値を試してみれば十分であるとされているが, 検討する必要があるだろう．そして (6) の重みを用いると以下のようになる．

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K(\boldsymbol{u_{ij}})(x_i-x_j)(x_i-x_j)' \tag{7}$$

クラスターを見つけるための計量Mとして Cassinus らは以下の行列Ｓの逆行列を使っている．

$$\mathrm{S}=\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}K(\parallel \boldsymbol{x_i-x_j} \parallel^{\boldsymbol{2}}_{\boldsymbol{V^{-1}}})(x_i-x_j)(x_i-x_j)'}{\sum_{i=1}^{n}\sum_{j=i+1}^{n}K(\parallel \boldsymbol{x_i-x_j} \parallel^{\boldsymbol{2}}_{\boldsymbol{V^{-1}}})}$$

ここで

$$\parallel \boldsymbol{x} \parallel^{\boldsymbol{2}}_{\boldsymbol{M}}=\boldsymbol{x'Mx} \tag{8}$$

Ｖは$\boldsymbol{x}$の分散共分散行列である．
Ｓ の式の$\parallel \boldsymbol{x_i-x_j} \parallel^{\boldsymbol{2}}_{\boldsymbol{V^{-1}}}=\boldsymbol{(x_i-x_j)'V^{-1}(x_i-x_j)}$ はマハラノビスの汎距離とよばれ, 一変量の場合の標準化距離の多次元への拡張として知られている．この重み関数は互いに距離が近いペアには大きな重みが与えられるようになっており．(5)(6) 式と共に考えると, 群内の分散共分散行列に近い値になると思われる．
そして, Ｍ＝Ｓ$^{-1}$ として固有値問題に定式化してみると

$$(T-\lambda S)\boldsymbol{a}=0 \tag{9}$$

となり正準判別分析の固有値問題と類似したものになり, クラスター検出に適した主成分が求まるのである．

4　"crabs"データの解析

Ｒに入っている"crabs"(200 匹のカニの 5 変量にわたる形態学的計測値データ) を基準化したものに一般化主成分分析を適用し, また通常の主成分分析, 正準判別分析の第 1,2 主成分得点, 判別得点を散布図に表示させ図 1,2,3 を見比べる．ここで, 正準判別の散布図において, ○は 1 から 50, △は 51 から 100, ＋は 101 から 150, ×は 151 か

ら 200 である．このデータに一般化主成分分析を適用するにあたり, パラメータ h=0.5,0.75,1.0 を試した結果を視覚的に見て 1 番クラスタに分かれた h は 1.0 であった．その時, 図 1 と図 3 を見てわかるように, 主成分分析ではまったくクラスタに分かれていなかった散布図が一般化主成分分析ではしっかりクラスタに分かれた．この結果より, 一般化主成分分析を使ってとった計量はかなり群の情報に近いものになっているのがわかる．正準判別分析の図 2 と比べても同等の分類ができていることからも”crabs”において一般化主成分分析は群の情報の代わりになるような計量がとれていることがわかる．

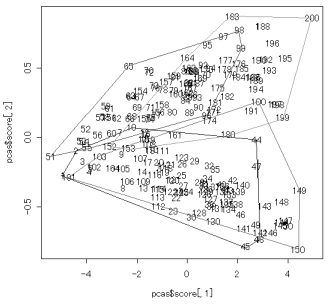

図 1　通常の主成分分析

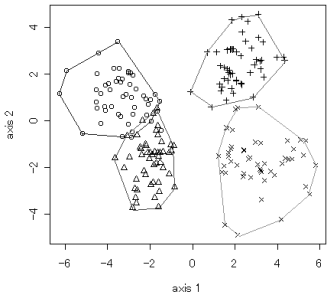

図 2　正準判別分析

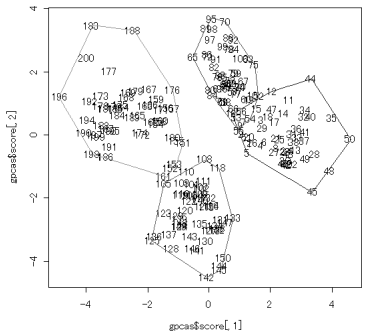

図 3　一般化主成分分析

5　一般化主成分分析の性能の評価

３つの例題の分析で 2 次元に次元を落とす場合を考え, 各群の 2 次元までの得点 (主成分得点, 判別得点) にもう一度正準判別分析を適用した時の固有値の和は $C = \sum_{k=1}^{g} n_k D_k^2$, n_k:k 群の平均から総平均までのマハラノビス距離, を表し, 群間の分離度を表す指標として利用できる．そこで C を用いて通常の主成分分析, 正準判別分析, 一般化主成分分析の結果について数値的に比較した．値が大きいほど群間の距離が大きく, 上手く群に分かれているということである．crabs データの場合についての結果のみ示す. 一般化主成分分析の性能はパラメータｈに依存するので, ｈの値を 0.5,0.75,1.0 に変えた時の一般化主成分分析の主成分得点, 通常の主成分分析の主成分得点, 正準判別分析の判別得点を用いたときの C の値を図 4 に示す. 視覚的に見た結果と同じで, 数値的に見ても h=1.0 の時が 1 番数値が大きく, 上手くクラスターに分かれていることがわかった. 通常の主成分分析の結果と比べると明らかに h=1.0 の数値が大きい. そして, 正準判別分析と比べるとほとんど数値に差がない. それだけ一般化主成分分析は上手くクラスターを検出できていることがわかった.

最後に,h を変化させた時, 固有値の合計が h=1.0 の時が 1 番大きかった．このことよりｈをさらに大きくしてみると, 固有値はさらに大きくなった．h=1.0 より 0.05 ずつ大きくしてみた固有値の和とのグラフを図 5 に示す．そうすると,h=1.55 の時, 固有値の和が最大の 10.48 となる．この時, 最もクラスタ間の距離が大きくなるのだが, 正準判別分析のクラスタ間の距離を越えることはできなかった．

	crabs
一般化主成分分析(h=0.5)	4.351241
(h=0.75)	8.621732
(h=1.0)	10.03174
通常の主成分分析	3.129169
正準判別分析	10.7979

図 4　群間分離度C

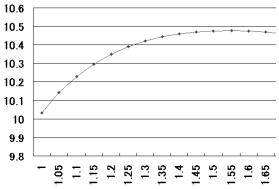

図 5　ｈと群間分離度Ｃのグラフ

6　おわりに

この研究で, クラスタを検出しやすいような計量のとり方をした一般化主成分分析は, 群の情報を使わなくても, しっかりクラスタに分類してくれた. さすがに群の情報を使用した正準判別分析ほどの分類力はなかったものの, 群の情報がないことを考えれば, かなり分類が上手くできていると思われる. 本論文では”crabs”以外に２つのデータに使用してみたところ, 一方では”crabs”と変わらないくらい上手く分類してくれたものの, もう１つではあまり効果的ではなく, 通常の主成分分析の方がクラスタ間の距離が大きいという結果になってしまった. ここで重み関数においてサンプルの分散共分散行列の変わりに新しく出た計量を使い, もう一度行列Ｓをもとめ, それを繰り返すという改良をした関数をこのデータに適用したところ主成分分析よりもクラスター間の距離をとることができた.

参考文献

[1] H.Caussinus. A.Ruiz：Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analyses. COMPSTAT1990,Springer,121-126(1990)