growth invariant discriminant function について

2003MM103 高橋 友弥
指導教員: 田中 豊

1　はじめに

種類の異なった生物の計測特徴から種類の違いを判別したい場合、ばらばらな成長の段階のデータが混ざっていると群内分散共分散は大きくなり、種類間の特徴が分からなくなる恐れがある。そのような場合に対応するための判別方法としてT.P.Burnabyは成長に対して不変な線形判別関数[1]を提案した。研究の目的はBurnabyの方法を理解してRの関数として作成し、実データを適用してその新しい関数を通常の判別分析と比較してこの方法の特徴を明らかにすることである。以下の２節以降で説明する。

2　線形成長モデル

2.1　導入

p次元の測定値（$x_1, \cdots x_p$ ）があるとき、次のような線形成長モデルを考える。

$$\frac{x_1 - x_{1A}}{\mu_1} = \cdots = \frac{x_p - x_{pA}}{\mu_p} = t \tag{1}$$

分母をはらって各変数の値を時間tの関数で表すと、

$$\begin{aligned} x_1 &= x_{1A} + t\mu_1 \\ &\vdots \\ x_p &= x_{pA} + t\mu_p \end{aligned}$$

となり、成長モデルとしてp次元ユークリッド空間の中の直線を考えることになる。

2.2　一般的な成長モデル

すべての個体の成長の方向（μ）が同一の時、Mを

$$M = \mu(\mu'\mu)^{-1}\mu' \tag{2}$$

と定義すると、M^2は

$$M^2 = \mu(\mu'\mu)^{-1}\mu'\mu(\mu'\mu)^{-1}\mu' = \mu(\mu'\mu)^{-1}\mu' = M \tag{3}$$

となりMはベキ等行列である。次にQを

$$Q = I - M = I - \mu(\mu'\mu)^{-1}\mu' \tag{4}$$

により定義すると、Q^2は

$$Q^2 = (I-M)^2 = (I - 2M + M^2) = (I - M) = Q \tag{5}$$

となりQはベキ等行列である。このときMQは

$$MQ = M(I - M) = M - M^2 = M - M = 0 \tag{6}$$

になりMとQは直交する。以上よりMは一次元部分空間$L\{\mu\}$への射影行列、Qは（p－1）次元部分空間$L\{\mu\}$の直交補空間$L\{\mu\}^{\perp}$への射影行列である。点xの部分空間$L\{\mu\}^{\perp}$への射影は成長過程に対して不変であり、x_A, x_Bの$L\{\mu\}^{\perp}$に射影された点間の距離は成長とは無関係なA，Bの距離を示す。

2.3　方向ベクトルの違う成長モデル

$$\Lambda = (\mu_A, \ldots, \mu_K)$$

ただし$\mu_A, \ldots, \mu_K$は$A, \ldots, K$ 群の成長の方向を表すベクトルとし、Mを

$$M = \Lambda(\Lambda'\Lambda)^{-1}\Lambda' \tag{7}$$

と定義するとM^2は

$$M^2 = \Lambda(\Lambda'\Lambda)^{-1}\Lambda'\Lambda(\Lambda'\Lambda)^{-1}\Lambda' = \Lambda(\Lambda'\Lambda)^{-1}\Lambda' = M \tag{8}$$

となる。またQを

$$Q = I - M = I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda' \tag{9}$$

により定義すると、２．２節と同様に計算すると、
(6) より、MQ=0
(8)より、$M^2 = MM = M' = M$
(5)より、$Q^2 = QQ = Q' = Q$
MとQはいずれもp×p行列、それぞれランクkと（p－k）のベキ等行列であり、Mはk次元部分空間$L\{\Lambda\}$への射影行列であり、Qは（p－k）次元部分空間$L\{\Lambda\}$の直交補空間$L\{\Lambda\}^{\perp}$への射影行列である。2.2節の場合と同様$Q(x_A - x_B)$は成長とは無関係なA、Bの距離を表す。

3　補助定理

x_A, x_Bは群A，Bの平均で$\delta = x_A - x_B$ とおく。Wを集団内の分散共分散行列とする。
G=$(g_1, \cdots, g_k)$ ただし$g_1, \cdots, g_k$ は各群の成長の方向を表すベクトル、rankG=kの時、３．１節と３．２節の補助定理が成立する。

3.1　補助定理１

任意のp次元ベクトルをℓとする。
制約条件　$\ell' W \ell =$一定　　$\ell' G = 0$のもとで

$$\frac{(\ell'\delta)^2}{\ell' W \ell} \leq \frac{(\ell^{*\prime}\delta)^2}{\ell^{*\prime} W \ell^*} \tag{10}$$

をみたすℓ^*は

$$\ell^* = W^{-1}\{I - G(G'W^{-1}G)^{-1}G'W^{-1}\}\delta$$

この定理より$\ell^{*'}x$は点xが部分空間G＝（$g_1, \cdots g_k$）の中で動いても不変な最良の判別関数である。

3.2　補助定理２

W:群内分散共分散行列、B:群間分散共分散行列、Gを成長の方向を表すk個のベクトルからなるp×k行列とする。

制約条件$\ell' G = 0$のもとで

$$\frac{(\ell' B \ell)}{(\ell' W \ell)} \tag{11}$$

を最大にするℓは行列

$$\{W - W^{-1}G(G'W^{-1}G)^{-1}G'W^{-1}\}B \tag{12}$$

の最大固有値に対応する固有ベクトルとして得られる。

すなわちGのランクがｋであり、BとWがともに正則であるなら、ｐ－ｋ個のゼロでない固有値が存在し、各固有値に対応する固有ベクトルを正準ベクトルとして採用すればよい。

4　マハラノビス距離D^2の加法的成分への分解

ℓ ’ｘは母集団Ａ，Ｂの判別関数で、マハラノビスの距離の定義は、

$$D^2 = \delta W^{-1}\delta = l'(x_A - x_B)$$

であり、l^*を使ってD^2を２つの成分、部分空間$L\{G\}^{\perp}$内の距離D_Q^2と部分空間$L\{G\}$内の距離D_M^2に分解する。

$$D_Q^2 = \delta'\{W^{-1} - W^{-1}G(G'W^{-1}G)^{-1}G'W^{-1}\}\delta \tag{13}$$

$$D_M^2 = \delta' W^{-1}G(G'W^{-1}G)^{-1}G'W^{-1}\delta \tag{14}$$

と表されるが、上の式を座標軸の変換によって簡単な形になおすことができる。すなわち平方根法で$R'R = W^{-1}$となるようなｐ×ｐ正則行列Ｒと求めて、

$y = Rx$、$d = y_A - y_B = R\delta$、$\Lambda = RG$

これらで座標変換をする。変換するとD^2、D_M^2、D_Q^2は

$$D^2 = d'd$$

$$D_M^2 = d'Md = (Md)'Md$$

$$D_Q^2 = d'Qd = (Qd)'Qd$$

と変換できる。δ が基底がＲでの平均値の差であるのに対して、ｄは基底がＩでの平均値の差をあらわす。

5　データの解析

5.1　データの内容

薬剤の効果判定のために急性肝炎（２０例)、慢性肝炎（３０例)、肝硬変（１０例）の合計６０例について１２種類の肝機能検査成績を６週間追跡した成績である。

5.2　解析結果

通常の判別分析で判別スコアを出して、plotしたのが図1である。

この図より急性肝炎(α)、慢性肝炎(β)、肝硬変(γ)として、それぞれの群でマハラノビス距離D^2 を出すと

$$\alpha - \beta = 16.01$$

$$\beta - \gamma = 9.58$$

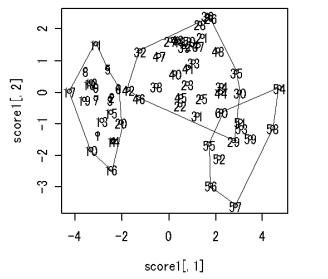

図 1: 正準評点のプロット

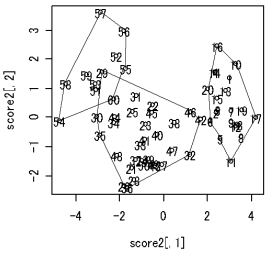

図 2: Ｇを加えた正準評点のプロット

$$\gamma - \alpha = 35.67$$

次にＧを加えた判別関数を行う。Ｇを各群ごとに主成分分析をして求めることにする。今回のデータの場合は第１主成分には形の因子をあらわす主成分、第２主成分に大きさの因子があらわれる。第２主成分の方向を成長の方向（病気の重さ）と考えることにする。それをＧとおく。

その成長の方向Ｇを加えて判別分析で判別スコアを出して、plotしたのが図２である。

図1と点対称の図が図2に見えるのは固有ベクトルの向きの不定性による。通常の判別分析の結果と比較する各群の平均ベクトル間のマハラノビスの距離を求めると下のようになる。

$$\alpha 1 - \beta 1 = 15.99$$

$$\beta 1 - \gamma 1 = 9.62$$

$$\gamma 1 - \alpha 1 = 35.57$$

5.3　考察

成長の方向Ｇを加えることによって、判別係数の値に少し違いがあらわれた。Ｇを考えることにより、病気の重さではなく、タイプで判別されているものと考えられる。判別スコアをもとに通常の場合とＧを加えた判別関数の場合の群間の距離D^2 を見るとほとんど変わりがないが、わずかに距離が短くなる結果になった。Ｇを考慮した判別はＧに直交するという制約をつけて群間の差を最大にする形に定式化になっており、制約のない場合に比べて判別効率は落ちると考えられる。一方成長の段階の異なる新しいデータ集合に対しても適用できるという利点を持つと期待される。

参考文献

[1] T.P.Burnaby：Growth Invariant Discriminant Functiona and Generalized Distances, Biometric,22(1),96-110(1966)

[2] 田中　豊，脇本和昌：多変量統計解析法, 現代数学社(2004)．

[3] 田中　豊，脇本和昌，垂水共之：パソコン統計解析ハンドブック　　多変量解析編, 共立出版(1992).

[4] 田中　豊：現象分析のための多変量解析と実際セミナー　正準分析法．大阪科学技術センター(1971)