PCA と判別分析を用いた判別変数の推定

仁平　智子　(指導教官 : 吉田 裕亮)

1　はじめに

本研究では, 判別に用いられるであろう隠れて見えない変数 (判別変数) に対して, その判別に影響が大きい変数, つまり相関が高い変数を推定する手法を提案し, 実データに応用して相関の高い変数を判断し, 隠れた変数と同じような判断ができるかを検証する.

ある隠れた変数により 2 群に分かれている多変量データを用意し, それらを主成分分析を用いてデータを可視化可能な 2 次元に縮約する. 縮約された合成変数が判別変数とどれくらいの相関があるかを判別する方法としてマハラノビス距離を用いる. 本来のグループと違うグループであると判断された数を誤判別率とする. 変数を 1 つずつ取り除いたデータに対し主成分分析を施し誤判別率を比較することにより, 相関の大小を判断する. ある変数を取り除いて誤判別率が上がる場合, その変数は判別変数と相関が高いと言えるであろう. また, ある変数を取り除いても誤判別率が変わらなかった場合, その変数は判別変数と相関がほとんどないと言えるであろう.

2　主成分分析 (PCA)

主成分分析 (以下 PCA) とは互いに相関関係のある多次元情報を少数の成分に縮約し, その多次元情報の総合力や特性を少数の成分で表す方法である.

X を $n \times k$ のデータ行列とし, X の縦成分 (変数) ごとに平均と標準偏差を求め標準化し, その行列を X_0 とする. このとき相関行列 R は

$$R = \frac{1}{n} X_0 {}^t X_0$$

で与えられる (相関行列 R は正定値行列である).

R の固有値を大きい順に $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ とする. また, λ_i に対応する固有ベクトルは第 i 主成分と呼ばれ, より大きな λ_i に対応する主成分に情報が縮約されている.

本研究では優固有値 · 固有ベクトルを求めるために, 比較的簡単な手順で求められる累乗法を用いた.

3　累乗法

適当な単位ベクトル $\overrightarrow{x}$ を選び, 規格化し行列 A に掛ける. 再び規格化し A に掛けることを何度も繰り返すことにより, A の絶対値最大の固有値に属する固有ベクトルに収束する.

A が正定値行列であり, 固有値 $\lambda_1, \lambda_2, \cdots, \lambda_n$ がすべて異なるならば, 固有ベクトル $\overrightarrow{x_1}$,$\overrightarrow{x_2}$, $\cdots$,$\overrightarrow{x_n}$ は互いに直交し,

$$A = \lambda_1 \overrightarrow{x_1}^t \overrightarrow{x_1} + \lambda_2 \overrightarrow{x_2}^t \overrightarrow{x_2} + \cdots + \lambda_n \overrightarrow{x_n}^t \overrightarrow{x_n}$$

が成り立つ. したがって, 最大固有値 λ_1 と固有ベクトル $\overrightarrow{x_1}$ が見つかったならば, $A - \lambda_1 \overrightarrow{x_1}^t \overrightarrow{x_1}$ に累乗法を再び適用すれば, 次に絶対値の大きい A の固有値 · 固有ベクトルが得られる. 以降, 望むだけの優固有値 · 固有ベクトルも同様である.

4　マハラノビス距離

多変数間の相関に基づき, 算出される距離としてマハラノビス距離がある. 2 次元データで各クラス ($j = 1, 2$) の平均が $\mu_j = {}^t(\mu_{1j}, \mu_{2j})$ で表される各 j 群のマハラノビス距離 D_j は,

$$D_j = \sqrt{{}^t(x - \mu_j)\Sigma_j^{-1}(x - \mu_j)}$$

で与えられる.

ここで, $\sum_j^{-1}$ は j 群の分散共分散行列 $\sum_j$ の逆行列である.

5　数値実験

5.1　実験の概要

判別変数 Y_0(データ数:2000 個) と相関のある変数 X_1, X_2,X_3(相関は $\rho(Y_0, X_1) > \rho(Y_0, X_2) > \rho(Y_0, X_3)$) と相関のない変数 X_4, X_5 を用意し, 1 つずつ変数を除いて誤判別率を比較する.

相関は

$$\rho(Y_0, X_i) = \frac{Cov(Y_0, X_i)}{\sqrt{V(Y_0)}\sqrt{V(X_i)}}$$

で与えられる. Y_0, Y_1, Y_2, Y_3 は互いに独立に $N(0, 1)$ に従うとき,

$$X_i = \cos\theta_i Y_i + \sin\theta_i Y_0$$

とおくと, $\rho(Y_0, X_i) = \sin\theta_i$ となる. この性質を用いて実験データを作成する. また, Y_0 の数値が 0 以上のとき 0, 0 未満のとき 1 と 2 群に分ける. 本研究では, 相関は以下のように設定した.

$$\rho(Y_0, X_1) = 0.8,$$
$$\rho(Y_0, X_2) = 0.5,$$
$$\rho(Y_0, X_3) = 0.1.$$

5.2 実験結果

(1) 変数:X_1, X_2, X_3 誤判別率 16.0%
(2) 変数:X_1, X_2, X_3, X_4, X_5 誤判別率 16.6%
(3) 変数:X_2, X_3, X_4, X_5 誤判別率 27.1%

(1) と (2) より X_4, X_5 を除いても誤判別率がほとんど変わらないので, X_4 と X_5 は判別変数との相関は低いといえる. また (3) より X_1 を除くと誤判別率が上がるので X_1 との相関は高いといえる.

上記の誤判別率を比較すると, 最初に設定した相関の大小関係と同じ傾向が読み取れ ($\rho(Y_0, X_1) > \rho(Y_0, X_2) > \rho(Y_0, X_3)$), X_4, X_5 は相関が見られないとも判断される.

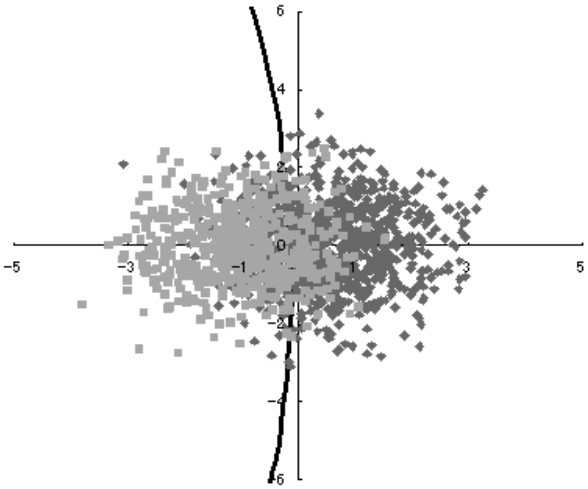

図 1: 変数:X_1, X_2, X_3 の PCA と判別曲線

6 実データへの応用

6.1 健康診断データ

糖尿病判定データが隠れて見えないとき, 年齢, BMI, 血圧, BUN, クレアチニンの 5 成分を用いてこの判定に有効な成分の推定を行った. これら 5 成分は血液検査なしで入手可能なデータを選んだ.

6.2 推定結果

全 5 成分で PCA を施したとき, 誤判別率は 12.4% であった. ここから各々1 成分取り除き 4 成分で PCA を施し, 誤判別率が高いとき, 除いた成分の相関が高いと判断される. 以下の結果よりクレアチニン, BUN, 年齢, 血圧, BMI の順に相関が高いことが分かった.

BMI を取り除いたとき	7.4%
血圧を取り除いたとき	8.3%
年齢を取り除いたとき	10.3%
BUN を取り除いたとき	13.8%
クレアチニンを取り除いたとき	28.6%

上記のことから BMI をノイズであると考え, BMI を取り除いた 4 成分から 1 成分ずつ取り除き PCA を施し, 誤判別率を求めた. 結果は以下の通りである.

血圧を取り除いたとき	3.7%
年齢を取り除いたとき	4.5%
BUN を取り除いたとき	12.3%
クレアチニンを取り除いたとき	31.5%

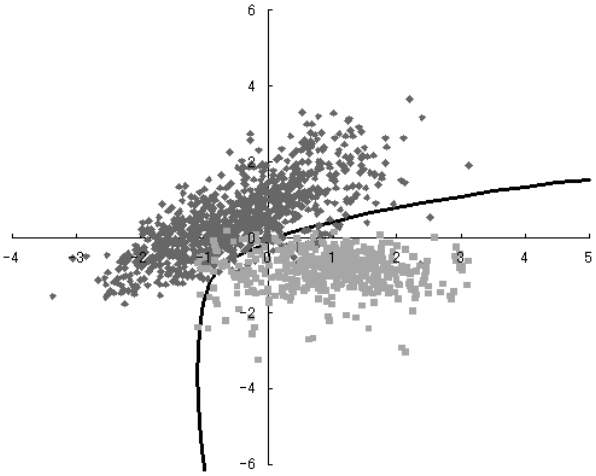

図 2: 年齢,BUN, クレアチニンでの結果

これより, クレアチニンが糖尿判定に大きく関連があると推定できる.

7 まとめ

本研究の手法を用いると, 多次元の変数の中から判別変数と相関の高い変数を判断することが可能であるといえる.

参考文献

鈴木義一郎 : 情報量基準による統計解析入門 講談社