PCA と非線形クラスタリングを用いた判別変数の選択

岸本華奈　(指導教員：吉田裕亮)

1　はじめに

日頃, 我々は様々な情報を基に判断を行っており, 情報を効率よく処理することで的確な判断を下す必要がある. 多変量のデータから判別するにあたって大きく関連している要因を知ることは重要となる. そこで, 本研究では主成分分析と非線形クラスタリングを用いて, 複数の変量から判別に有効な変量を推定する手法を提案する.

ある多変量データが与えられたとき, いくつかの変量を取り除いたデータに対して主成分分析を施し可視化可能な 2 次元に縮約する. その 2 次元データに対し非線型手法であるスペクトラルクラスタリングを用いて 2 群に分類する. クラスタリングの良さを計ることで, 2 群判別に有効な変量を推定することを試みる.

ある変量を取り除いても 2 群への分かれ方が取り除く前とそれほど変化がない場合, その変数は判別に大きな影響を与えないと考えられるだろう. また, ある変数を取り除いて 2 群の分かれ方に改善が見られた場合, その変数は判別に影響を与えるものと考えられるだろう.

本研究では, 2 群判別に用いた変量が既知の多変量データに対し, 幾つかの関連する変量を加えた多変量データを構成し, 変量を削除して PCA を施し, スペクトラルクラスタリングで非線型 2 群判別を行う. これより手法の有効性を見る. この 2 群判別の分類結果が正しいかどうかは, 本来のグループと違うグループであると判断された数, 誤判別率を求めることで評価することにする.

以下に, まず本研究で用いる基本的なツールの概説を述べておく.

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_1 に対応する主成分に情報が縮約されている.

また, 優固有値・固有ベクトルを求める手法として, 比較的簡単な手順で求められる累乗法を本研究では用いた.

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　スペクトラルクラスタリング

データをまとまりごとに集めてグループ分けをすることをクラスタリングと呼び, スペクトラルクラスタリングとは, クラスタリングを類似度行列の固有値問題に帰着させる手法のことである.

まず, サンプル点を頂点とするグラフ構造として考える. 枝にはサンプル点同士の近さを表すために, 互いに近いデータには 1, 遠いデータには -1 という離散値をわりあてる. クラスタリングは各サンプルに対して, グループに対応するラベル, $\beta_1 = +-1$ を割り当てる問題としてとらえることが出来る. グラフが 2 分割され, 分割されたグループ間のサンプル点同士を結ぶ枝の式は,

$$\min_{\beta} \sum_{i,j} K_{ij} (\beta_i - \beta_j)^2 = 2\beta^T P \beta$$

になる. $K_ij : i, j$ は成分を結ぶ枝の重みで, 本研究では非線型クラスタリングを行うためにガウスカーネルを用いることにする. ここで

$$\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$$

と定義すると,

$$P = \Lambda - K$$

とかける. β は 2 値ベクトルという制約があり整数計画化問題とよばれるが, 一般には解くのが難しいため, 任意の実数ベクトルに制約をゆるめて行う. こうすることで固有値問題の最小固有値に対応する固有ベクトルを求めることに帰着でき, 固有ベクトルの符号が正になるか負になるかどうかで 2 群に分類する. （ただし ${}^t\beta\Lambda\beta = 1$ とする）

5　実データに基づく解析例

5.1　健康診断データ

約 60 万人の成人女性の健康診断データから 1 万人選出されたデータを利用し, 血圧判定に影響のある変

量の選択に, この手法が有効であるかを見てみる. 健康診断における血圧の異常の有無は最高血圧と最低血圧のみに依存して判定されている. この判定がどの変量を使って２群に分類されているかの推定を行う.

まず, 血圧の判定が正常とみなされた人の中から２００人, 正常でないとみなされた人の中から２００人をランダムに抽出し, 計４００人のデータを用意する. それぞれ, 最高血圧, 最低血圧以外に, 血圧判定に影響を与えない余分な血液データである赤血球数, 白血球数, ヘマトクリット, ヘモグロビンの４つを加えた合計６つの変量をもっている. これらを用いて血圧判定にどの変量が使われて２群に分類されているか、有効な成分の推定を試みた.

5.2　クラスタリング結果

全６成分で PCA を施しクラスタリングしたときの結果は, 以下のようになった. このときの誤判別率は 50.8%であった.

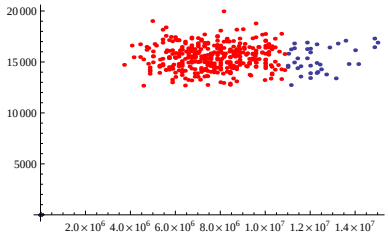

図 1: 全６成分の結果

ここから数成分を取り除き, 同様に求めたいくつかの結果を以下に記す.

(1) ヘマトクリット, ヘモグロビンを取り除いたとき

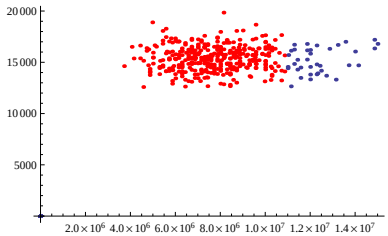

図 2: Ht,Hb を除いた結果

(2) 赤血球数, ヘモグロビン, 最高血圧を取り除いたとき

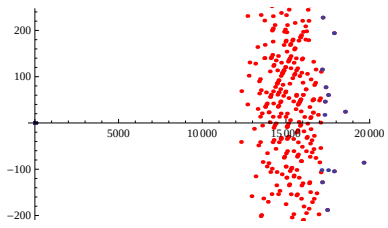

図 3: RBC,Hb, 血圧 MAX を取り除いた結果

(3) 赤血球数, 白血球数, ヘマトクリットを取り除いたとき

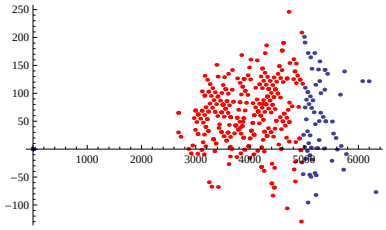

図 4: RBC,WBC,Ht を取り除いた結果

(4) 赤血球数, 白血球数, ヘマトクリット, ヘモグロビンを除いたとき

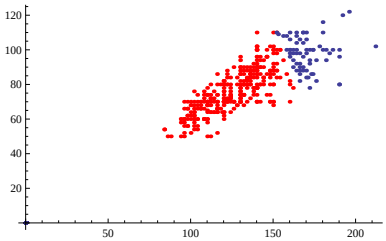

図 5: 血圧 MAX, 血圧 MIN での結果

5.3　推定結果

それぞれのの誤判別率をそれぞれ求めると（１）50.8%（２）53.0%（３）33.3%（４）32.0%となり, 最高血圧と最低血圧以外の成分を取り除いた場合が誤判別率が低くなっていることがわかる. また, 最高血圧と最低血圧でクラスタリングを行った結果（図 5）と血圧判定の正確な結果（図 6）を比較すると似た様な散布図となっている.

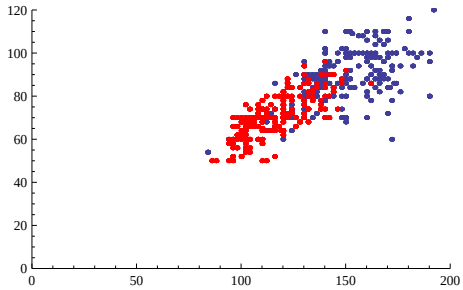

図 6: 血圧判定の正確な結果

よって, 最高血圧と最低血圧がこの判定に必要な変量であるという推定が正しい結果であったと言える.

6　まとめ

本研究の手法を用いると, 複数の変量から２群判別に有効な変量を推定することが可能であるといえる.