カーネル PCA を用いた多変量データの要因変数の推定

中宗由佳　(指導教員：吉田裕亮)

1　はじめに

私たちの身の回りには情報があふれている．その多くの情報の中からなんとなく, これまでの経験から重要と思われるものを取り出し判断するのではなく, 必要な情報を抽出することはできないだろうか.

多変量データの判別は本来計算量が多く時間がかかるものである．しかし, 判別へ影響を与える要因変数を見つけることで, 少ない計算量で判別が行え, 結果の分析にも役立つと考えられる. 本研究では, カーネル PCA を用いて誤判別率を求め変数削減を行うことで, 判別の要因変数を推定する手法の一つを提案する.

2　カーネル PCA

2.1　PCA(主成分分析)

PCA とは, 多変量データのもつ情報をできるだけ保つような低次元空間に, 情報を縮約する方法. 分散はできるだけ大きくし, 縮約したデータを元のデータの近似とみなしたとき, その 2 乗誤差ができるだけ小さくなるように低次元に射影する.

まずデータの変数間の共分散行列または相関行列を用いて, この行列の固有値問題を解く．この固有値は, その主成分がどの程度元のデータの情報を保持しているかを表し, 固有値の大きい方から第 1 主成分, 第 2 主成分と定め, 相関係数を最大化するような少数の合成変数を取り出す．ただし, PCA(線形) は非線形構造のデータがとらえにくい欠点があり, 線形構造のみでの分析は不十分であることが多い.

2.2　カーネル法

カーネル法とは, カーネル関数を用いたデータ解析の方法論. データを非線形変換することで, 非線形特徴や高次モーメントを抽出し解析しやすいデータに変換し, 変換後のデータに線形の解析手法を用いる．カーネル法を用いると, 複雑な関数を表現でき, 非線形な関係を考慮できる．また文字列やグラフ構造など複雑なデータ構造も実数と同じ処理が可能となる.

カーネル関数により写像された空間は, 再生核ヒルベルト空間の性質をもつ. そのため, カーネル関数を用いてなんらかの特徴ベクトル間の内積とみなせる．これを一般に「カーネルトリック」と呼ぶ.

カーネル関数 $k(x, x')$ とは, データ変数の集合の 2 つの要素 x, x' に対し, x, x' のそれぞれの特徴ベクトル $\phi(x), \phi(x')$ どうしの内積

$$k(x, x') = \phi(x)^T \phi(x')$$

として定義される．また再生核ヒルベルト空間には, 関数の値が定まり, カーネルが連続の場合その空間は連続となる性質がある．カーネル関数を用いることで, 計算の煩雑さを抑え, 内積に基づく線形解析手法を高次元ベクトル空間へ拡張し, 実質的に非線形な解析を行うことができる．本研究ではラプラシアンカーネルをカーネル関数に用いた．よく使用されるカーネル関数は以下のようなものがある.
ラプラシアンカーネル

$$k(x, x') = \exp\left(-\beta \sum_{m=1}^{d} |x_m - x'_m|\right) \quad (\beta > 0)$$

ガウスカーネル

$$k(x, x') = \exp\left(-\frac{1}{\sigma^2}||x - x'||^2\right)$$

多項式カーネル

$$k(x, x') = (c + x^T x')^d \quad (d : \text{自然数}, c \geq 0)$$

2.3　カーネル PCA

カーネル PCA は, 特徴抽出した空間で分散が最大となる高次元の特徴ベクトルに変換してから、通常の PCA を行い低次元の線形部分空間を求めるもので, 非線形な方向でのデータのばらつきを扱うことができる. しかし結果はカーネルの選び方に依存し, カーネルの選び方は必ずしも明確でないという欠点がある.

本研究では第 2 主成分まで求めてプロットし, 最適と考えられるカーネル関数及びパラメータの値を見つけることとする.

3　カーネル回帰

カーネル回帰とは線形回帰にカーネル関数を用いて拡張したモデルで

$$y = \sum_{m=1}^{d} \alpha_m k(x_m, x)$$

と定義される．与えられた x に対して各点 x_m との近さを測った関数 k を 1 つの成分とみて, それらを α_m の重みで足し合わせる．直線との 2 乗誤差の総和を最小にする重み α_m を見つける．本研究では, カーネル PCA で処理した後, 2 次元に可視化したデータに 0 群と 1 群の 2 値を割り当て, カーネル回帰を用いて 2 群の判別曲線を引く．判別曲線は回帰関数の値が 0.5 となる等高線で与えられると考えられる.

4　提案手法

4.1　データの 2 群判別

多変量データにカーネル PCA を用いて, 2 次元に可視化したプロットデータを構成する．プロットデータに, 先に述べた 0 群と 1 群を割り当て, カーネル回帰を用いて判別曲線を引き 2 群判別を行う．まず全変数での 2 群判別から始めて逐次変数を減少させ, 結果より要因変数の推定を行う.

4.2 削減する変数の選択方法

簡単のため q 次元データに対し 2 群判別を行う操作を $F(q)$ とする．Y は q 変数で構成されるとする．この q 個の変数の 1 つを削減して得られる q 個の $F(q-1)$ のうち誤判別が最小となる変数を削減する．$q \leftarrow q-1$ として，同様に削減する変数を選んでいく．以後これを繰り返す．

変数を削減し，誤判別率が一定以上になる，または 2 群に分けられない場合，その変数を Y の構成するには削減不可能な変数と判断する．

5 実データへの応用

成人女性約 60 万人の健康診断結果から約 1 万人がランダムに抽出されたデータがある．このデータの脂質代謝 (T-Cho, HDL, TG) の判定が正常 (A,B,C) な約 9500 人から 400 人，異常 (D,E,F) な約 500 人から 400 人をランダムに抽出し，合計 800 人分の年齢，身長，体重，肥満度，総コレステロール (T-Cho)，善玉コレステロール (HDL)，中性脂肪 (TG) の 7 次元データを用意した．判定が A, B, C の人を健常者，D, E, F の人を非健常者として 2 群判別を行い，その結果から変数を削減する．

5.1 実験結果 (赤:非健常者，緑:健常者)

全 7 次元データでの結果，誤判別率は 0.375%

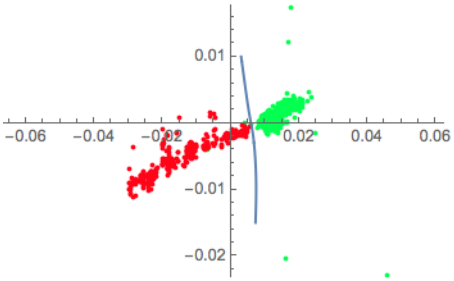

図 1：全 7 変数の結果

次に，1 変数ずつ削減する．誤判別率が体重，身長，年齢，肥満度の順に最少となり変数の削減候補．残りの 3 次元での結果は以下のようになる． 誤判別率は 0.375%で全 7 次元データと同じ精度で 2 群判別が行えた．

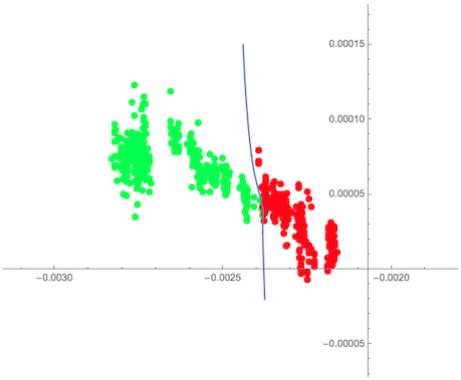

図 2：3 変数の結果

さらに 1 変数削減し，要因変数を推定する．その結果は，これ以上 1 変数でも削減すると主成分値が非常に小さい，または 2 群に分けられなかったためカーネル回帰は行えなかった．

図 3 は判別曲線が引けるほどには 2 群に分けられなかった．図 4 は群としてまとまりはあり，図 5 は 2 群がほぼ重なった．

以上より 3 変数 (T-Cho,HDL,TG) から 1 変数でも削減されると判別不可能となるため，この 3 変数は削減不可能な変数と考えられる．判別への影響度合いは結果より，2 群がほぼ重なった T-Cho が最も高く，次に TG, HDL の順に影響を与えているとも推定される．

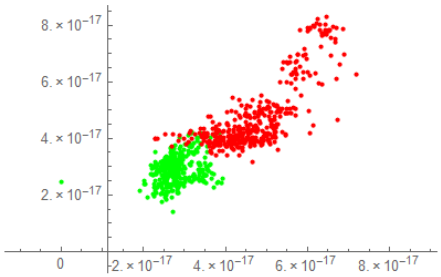

図 3：T-Cho, TG の結果

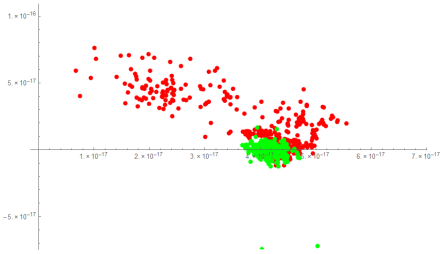

図 4：T-Cho, HDL の結果

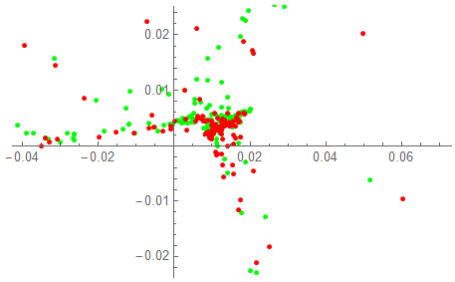

図 5：HDL, TG の結果

6 まとめ

カーネル PCA を用いた 2 群判別は，カーネル PCA と SVM を用いて判別を行った先行研究より高い精度となり有効な手法の一つと考えられる．結果より要因変数の推定も可能である．しかし，判定の違いを与える数値の境界を推定することはできない点は今後の課題である．またカーネル関数やパラメータの取り方は明確ではない難しさがある．

参考文献

[1] 赤穂昭太郎, カーネル多変量解析〜非線形データ解析の新しい展開〜, 岩波書店, 東京, 2008

[2] カーネル法　正定値カーネルを用いたデータ解析, http://www.ism.ac.jp/~fukumizu/ISM_lecture_2004/Lecture2004_kernel_method.pdf

[3] 竹内友美, カーネル PCA と SVM による高次元データの変数削減, お茶の水女子大学情報科学科,2011