2002年日本オペレーションズ・リサーチ学会
春季研究発表会

L_p計量ノルムを用いた主成分分析法

現在申請中　日本大学　＊　畑澤　文祐　HATAZAWA Fumihiro
01100500　日本大学　　大澤　慶吉　OSAWA Keikichi
01205220　日本大学　　篠原　正明　SHINOHARA Masaaki

1　はじめに

私たちが意思決定をする時、多変量データ解析 (multivariate data analysis) が多く使われる。その代表的な手法に主成分分析 (Principal Component Analysis:PCA) がある。主成分分析とは、相関のある多くの変数の値を、1つまたは少数の合成変量（主成分）であらわす方法である。つまりは、データから情報を縮約する事が、この分析の目標である。本研究では、従来型の主成分分析法 (L_2-PCA) に距離関数を導入し、新たなる主成分分析法 (L_1-PCA) の提案をし、かつそのメリットを紹介する。なお、本実験計算には「EXCEL多変量解析」(エスミ)、「NUOPT」(数理システム) を用いた。

2　距離関数とL_2-PCA

2.1　距離関数

$a,b\in\mathbb{R}^n$とする。このとき2点a,b間の距離を

$$d^{(p)}(a,b) = \left(\sum_{i=1}^{n}|a_i-b_i|^p\right)^{\frac{1}{p}} \quad (1)$$
$$(1\leqq p<\infty)$$

で定める。これをp乗距離、またはL_p計量ノルムと言う。

2.2　L_2-PCA

解析対象のm次元n個のデータを以下の行列S、

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mn} \end{pmatrix} \quad (2)$$
$$= (S_1,S_2,\cdots,S_n)$$

と表す。

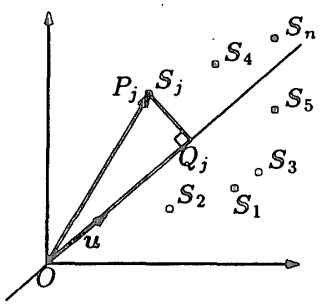

図1: n個のm次元データベクトルと主成分ベクトルu

また図1より

$$|\overrightarrow{OQ_j}| = |u^TS_j|$$

$$|\overrightarrow{OQ_j}|^2 = u^TS_jS_j^Tu$$
$$\therefore \sum_{j=1}^{n}|\overrightarrow{OQ_j}|^2 = u^T\left(\sum_{j=1}^{n}S_jS_j^T\right)u$$
$$= u^TSS^Tu$$

以上より、L_2-PCAを定式化すると、

L_2-PCA	
目的関数	$u^TSS^Tu\to\max$
制約式	$u^Tu=1$

となる。

3　L_1-PCA

本論文で提案するL_1-PCAは、$\sum_{j=1}^{n}|\overrightarrow{OQ_j}|^2\to\max$の代わりに

$$\sum_{j=1}^{n}|\overrightarrow{OQ_j}|\to\max$$

とする。定式化すると次式を得る。

L_1-PCA 第1主成分	
目的関数	$\sum_{j=1}^{n}\lvert u^TS_j\rvert\to\max$
制約式	$u^Tu=1$

ここで、目的関数は補助変数等を導入することにより、線形に等価変換できるだろうが、制約式は二乗和 = 1 であり、本質的に非線形性が残ってしまい、線形計画問題への帰着は無理である。そこで、本研究では、NUOPT の NLP 版を (絶対値関数を利用し) 直接適用して、最適解を求めた。

一方、第 2 主成分ベクトルは第 1 主成分ベクトルと、直交するという性質があるので、以下のようにして求めることができる。

$L_1 - PCA$ 第 2 主成分

目的関数 : $\sum_{j=1}^{n} |\boldsymbol{v}^T \boldsymbol{S}_j| \to \max$

制約式 : $\boldsymbol{v}^T \boldsymbol{v} = 1$

$\boldsymbol{u} \cdot \boldsymbol{v} = 0$

4 例題

比較実験として、文献 [1] の例題を解く。

表 1: 例題

データ No.	身長	体重	胸囲	座高
1	182.0	67.5	85.1	93.7
2	176.3	63.5	83.5	93.0
3	179.0	63.0	87.0	96.3
4	172.8	88.0	101.0	94.1
5	170.0	61.5	83.0	91.8
6	164.5	58.5	83.5	88.5
7	170.6	57.0	83.0	88.1
8	170.4	79.0	98.0	92.5
9	165.0	63.3	89.5	89.7
10	172.5	66.0	91.4	91.4
11	173.6	62.5	88.2	92.0
12	176.3	64.0	87.0	97.0
13	170.5	60.5	84.0	88.5
14	161.0	63.0	86.0	87.2
15	173.0	65.0	84.0	91.0
16	166.8	70.0	93.5	86.2
17	171.0	62.0	84.0	88.4
18	175.5	69.5	92.5	94.5
19	176.3	73.5	102.8	91.0
20	169.0	77.5	94.5	95.4

表 2: 例題の第 1 主成分

項目	L_2 第 1 主成分	L_1 第 1 主成分
身長	0.101523452	0.008800329
体重	0.780282378	0.745315269
胸囲	0.603256464	0.663994916
座高	0.130131453	0.059484945

表 1 の例題に関して、第 2 主成分を求める。

表 3: 例題の第 2 主成分

項目	L_2 第 2 主成分	L_1 第 2 主成分
身長	0.89628917	0.910619958
体重	−0.058768414	−0.004568795
胸囲	−0.162892818	−0.043751038
座高	0.408262074	0.410896971

5 制約付き $L_1 - PCA$

評価者の意図で、評価ベクトルに制約を付けたい場合がある。制約式は、線形であるとは限らない。しかしながら、数理計画法パッケージの計算を用いれば、身長 ${u_1}^2$+ 体重 ${u_2}^2 \geqq 0.8$ といった条件の解析も簡単実現できる。

表 4: ${u_1}^2 + {u_2}^2 \geqq 0.8$

項目	L_1 非線形制約付き第 1 主成分
身長	−0.010560238
体重	−0.894364848
胸囲	−0.445429719
座高	−0.039904465

${u_1}^2 + {u_2}^2 = 0.800006496$ であった。

6 考察

従来までの L_2 つまり、2 乗距離最大化と同様に、L_1 つまり、絶対距離最大化での PCA も簡単に解くことができた。また、その結果は第 1,2 主成分共に、L_2 と L_1 で、同傾向の係数が導き出されることが分かった。また、線形制約はもちろん、非線形制約を与えても、それを満たす最適解を求めることができた。

7 おわりに

p 乗距離にもとづく主成分分析法を提案し、制約条件を考慮することができることを示した ($p = 1$)。$p = 1, 2$ 以外の場合についても意味のある分析法があるが、今後の課題である。

参考文献

[1] 木下栄蔵, 「わかりやすい数学モデルによる 多変量解析入門」 近代科学社

[2] 高根芳雄, 「制約つき主成分分析法 新しい多変量データ解析法」 朝倉書店 (1995)