R. ニャナデシカンの一般化主成分分析について

新 井 宏 尚

1. 初 め に

本ノートでは K. Pearson (1902) により最初提案され，後に H. Hotelling (1933) が現在ある形に発展させた主成分分析をさらに拡張したと言う R. ニヤナデシカンの一般化主成分分析法について考察してみる。

2. 通常の主成分分析と一般化主成分分析

通常の主成分分析法の概要をここで説明すれば次のようになろう。いま p 個の特性を持つ変量 $x_1, x_2, \cdots\cdots x_p$ に関して n 個のサンプルが与えられるならばベクトル $\boldsymbol{x}$，マトリックス $\boldsymbol{X}$ は次のようになり

$\boldsymbol{x}'=(x_1, x_2, x_3, \cdots\cdots x_p)$

$\underset{n\times p}{\boldsymbol{X}'}$，$\boldsymbol{X}'$ の行で n 個の p 次元データを示す

したがい，標本平均ベクトル ($\bar{x}$)，分散，共分散行列 ($\boldsymbol{S}$)，相関行列 ($\boldsymbol{R}$) は(1), (2), (3)のごとく定義できよう。

$$\bar{\boldsymbol{x}}'=(\bar{x}_1, \bar{x}_2, \cdots\cdots \bar{x}_p)=\frac{1}{n}\underset{(1\times n)(n\times p)}{\cdot \boldsymbol{1}\cdot \boldsymbol{X}'} \tag{1}$$

$$\boldsymbol{S}=(s_{ij})=\frac{1}{n-1}(\boldsymbol{X}-\bar{\boldsymbol{X}})(\boldsymbol{X}-\bar{\boldsymbol{X}})' \tag{2}$$

$$\boldsymbol{R}=\boldsymbol{D}_{1/\sqrt{s_{ii}}}\cdot\boldsymbol{S}\cdot\boldsymbol{D}_{1/\sqrt{s_{ii}}} \tag{3}$$

ただし，$\boldsymbol{D}_{1/\sqrt{s_{ii}}}$ は i 対角要素が $1/\sqrt{s_{ii}}$ $(i=1, 2, \cdots\cdots p)$ に等しい対角行列である。

このとき，主成分を見つけることは代数的には標本分散行列 $\boldsymbol{S}$ の固有値と固有ベクトルを見つける問題に帰着する。つまりベクトル $\boldsymbol{a}'=(a_1, a_2, \cdots\cdots a_p)$ と $\boldsymbol{x}$ との線形結合 $\boldsymbol{a}'\cdot\boldsymbol{x}$ を求め

$$\boldsymbol{a}'\cdot\boldsymbol{x} \tag{4}$$

式(4)を $\boldsymbol{a}'\cdot\boldsymbol{a}=1$ の条件下で第1主成分が最大の分散を持つように $\boldsymbol{a}$ を求める。このためには(4)式と条件 $\boldsymbol{a}'\boldsymbol{a}=1$ より，ラグランジュ乗法により

$$\phi = \boldsymbol{\alpha}' \cdot \boldsymbol{x} - \boldsymbol{\lambda}(\boldsymbol{\alpha}' \cdot \boldsymbol{\alpha} - 1) \qquad \text{subject to} \qquad \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1 \tag{5}$$

と書き改め，

$$\partial\phi / \partial\boldsymbol{\alpha} = 0 \tag{6}$$

とすれば，分散$V(\boldsymbol{\alpha}' \cdot \boldsymbol{x}) = \boldsymbol{\alpha}' \cdot \boldsymbol{S} \cdot \boldsymbol{\alpha}$ より

$$\partial\phi / \partial\boldsymbol{\alpha} = 2 \cdot \boldsymbol{S} \cdot \boldsymbol{\alpha} - 2 \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\alpha} = 0$$

$$(\boldsymbol{S} - \boldsymbol{\lambda} \cdot \boldsymbol{I}) \cdot \boldsymbol{\alpha} = 0 \tag{7}$$

が得られる。したがい，(7)式が$\boldsymbol{\alpha} = 0$以外の解を持つのは

行列式 $|\boldsymbol{S} - \boldsymbol{\lambda} \cdot \boldsymbol{I}|$ が

$$|\boldsymbol{S} - \boldsymbol{\lambda} \cdot \boldsymbol{I}| = 0 \tag{8}$$

となることである。(8)を解くには分散行列$\boldsymbol{S}$の固有値を求めればよいので$\boldsymbol{\alpha}' \cdot \boldsymbol{\alpha} = 1$より

$$\boldsymbol{\alpha}'(\boldsymbol{S} - \boldsymbol{\lambda} \cdot \boldsymbol{I})\boldsymbol{\alpha} = \boldsymbol{\alpha}'\boldsymbol{S}\boldsymbol{\alpha} - \boldsymbol{\lambda} = 0$$

$$\therefore \ \boldsymbol{\alpha}'\boldsymbol{S}\boldsymbol{\alpha} = \boldsymbol{\lambda}$$

でかつ，Max $(\boldsymbol{\alpha}'\boldsymbol{S}\boldsymbol{\alpha})$ とするには Max $\{\lambda_i\ ;\ i = 1, 2, \cdots\cdots p\}$ となる $\lambda = \lambda_i$ を選択し，それに対応する固有ベクトルを求めれば良い。

しかしこの手順にも欠点がある。それは変量 x_i の測定単位が全て同じの場合には都合が良いが測定単位を変更する場合には，分散，共分散行列の両側に対角行列 $(D_{1/\sqrt{s_{ii}}})$ を掛けねばならない ((3)式)。このような場合には変量の尺度不変性は維持されず固有値，固有ベクトルに尺度交換がどのようなことを及ぼすかは複雑となる（固有値，固有ベクトルが全く異ってしまう場合がある。[*1]）。したがい，主成分分析を行う場合に通常相関行列 $\boldsymbol{R}$ からスタートする方がよりスムーズとされている。

しかし，相関行列から出発するには変量相互間は互いに線型関係を持っていることが前提とされており，その有意性の検定を行わなければならないので変量相互間に非線型の関係がみられる場合にはどのようにすべきなのかという問題に行きつく[*2]。一般的には変量相互間に線型でない関係がみられるときには通常相関係数は使用せず相関比により2つの関係の度合（$\eta_{x_i x_j}$）を測ることができるが $\eta_{x_i x_j}$ の行列を作っても相関行列と異なり非対称行列[*3]になるためにこのような接近法は考えられない。

他方R. ニヤナデシカンは非線型の主成分分析を（*4）の中で展開している。その骨旨を（*4）に従って述べれば次のようになろう。

例として $p = 2$ の場合について説明している。$y = f(x_1, x_2)$ の関数型 f を最初に特定化する。

---

*1. 日科技連奥野忠一他，多変量解析法 pp. 181-182
*2. 非線型の成分分析の示唆については，竹内啓「計量経済学の研究」東洋経済新報社，p. 160 に記載されている。
*3. 非対称行列の対称化の方法は存在しないであろう。
*4. R. ニヤナデシカン，丘本他訳，統計的多変量データの解析　日科技連

2次多項式としてfを特定化する。つまり

$$y=a_1x_1+a_2x_2+a_3x_1\cdot x_2+a_4\cdot x_1^2+a_5\cdot x_2^2 \tag{9}$$

とyを置いて，yの分散を最大になるように係数を定める。

ここで，$x_1\cdot x_2$をx_3, x_1^2をx_4, x_2^2をx_5と表わせば

式(9)は

$$y^*=a_1x_1+a_2x_2+a_3x_3+a_4x_4+a_5x_5 \tag{10}$$

となるので$\boldsymbol{a}^{*\prime}=(a_1, a_2, a_3, a_4, a_5)$で係数ベクトルを表わすならば問題は$\boldsymbol{a}'^*\boldsymbol{a}=1$のもとで$\boldsymbol{a}^{*\prime}\cdot\boldsymbol{y}^*$の分散を最大にするように$\boldsymbol{a}^*$を決定することになる。また標本平均ベクトルと分散，共分散行列は次のように定義できる

$$\bar{\boldsymbol{x}}^{*\prime}=(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \bar{x}_5) \tag{11}$$

$$\boldsymbol{S}^*=\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i^*-\bar{\boldsymbol{x}}^*)(\boldsymbol{x}_i^*-\bar{\boldsymbol{x}}^*)' \tag{12}$$

そして，通常の主成分分析（線型主成分分析）を実行する。また$n\times p$コのデータ，マトリックスが与えられたときに2次の主成分分析を実行するにはpコの変数に$p+[p(p-1)/2]$コの変数を付け加えて

$$\left[2p+\frac{p(p-1)}{2}\right]\times\left[2p+\frac{p(p-1)}{2}\right]$$

の分散行列を使い，通常の主成分分析を行えば良いとしている。

次にこの手法による例*4を1つ上げその後の問題に触れよう。

（例）　表1-1はx_1のデータを与えたときの，放物線$x_2=2+4x_1+4x_1^2$の軌跡をえがいている。このときの2次の主成分分析を実行した結果，得られた固有値と固有ベクトルが表1-1と表1-2に載っているものである。

次に，式(9)のバリエーションとしてデータに統計的な誤差を含めた場合の固有値と固有ベクトルが表2に載っている。この場合の統計的誤差とは各観測値の組に正規乱数$\left(\text{平均ゼロ，分散}\frac{1}{6}\right)$を

表 1-1　2次の主成分分析の例に対する固有値解析

		固　有　値		
2,163.634	219.915	2.258	2.223	0.000009
		固 有 ベ ク ト ル		
−.002513	.246077	.508757	−.442445	.696310
.169321	.011882	.758811	.604229	−.164076
−.094253	.932909	−.212548	.274991	.0000004
.044843	−.243106	.319056	.593499	.696311
.980015	.099425	−.135640	−.106239	.0000003

*4　統計的多変量データ解析，ニヤナデシカン，丘本他訳，日科技連 pp. 49-50

表 1-2　データおよび最大，最小固有値から得られた座標

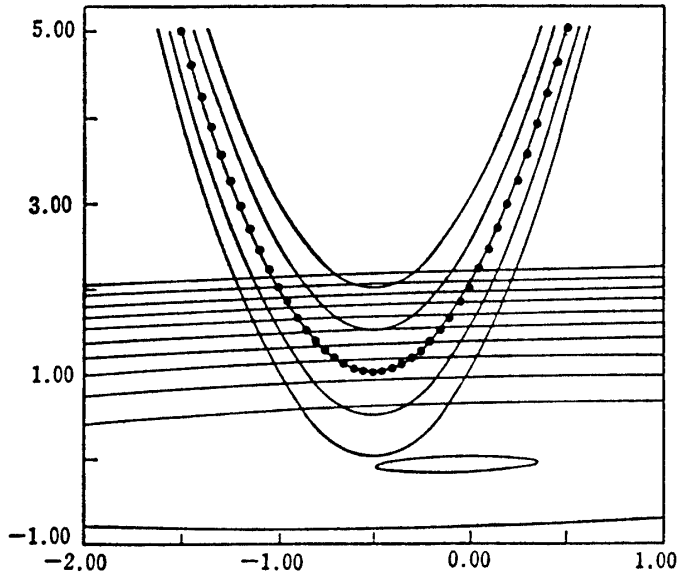
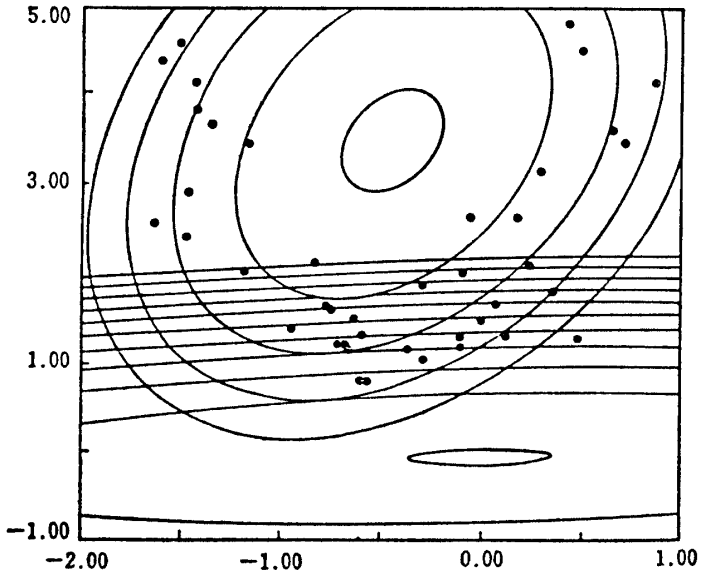


表 2-1　2次の主成分分析の例に対する固有値解析

		固　有　値		
1969.563	285.938	5.717	2.034	1.081
		固 有 ベ ク ト ル		
−.010915	.260105	−.269383	.610264	.698023
.172863	.012363	.189273	.761768	−.594853
−.106665	.930317	.301394	−.146763	−.103707
.062983	−.230187	.892766	.062053	.377047
.977064	.117118	−.061142	−.147978	.077413

表 2-2　データおよび最大，最小固有値から得られた座標



加えたものである。この２つの結果をみると，誤差が，固有値，固有ベクトルに影響を与えていることもわかる。

しかし，ここまでのところで変量が誤差を含む場合のモデルについて疑問点が生じよう。

先づ第１にノイズが入った場合について式(9)をそのまま適用しても良いのであろうか。通常，変量に誤差が含まれる場合，変量 x_1, x_2 は

$$x_1=\hat{x}_1+\varepsilon_1 \tag{13}$$

$$x_2=\hat{x}_2+\varepsilon_2$$

ただし，$\hat{x}_1, \hat{x}_2$ は真の値，$\varepsilon_1, \varepsilon_2$ は互いに $\varepsilon_1 \sim N\left(0, \frac{1}{6}\right)$，$\varepsilon_2 \sim N\left(0, \frac{1}{6}\right)$ とする。

と表わすことができるので，(9)式は

$$y=a_1\hat{x}_1+a_2\hat{x}_2+a_3\hat{x}_1\hat{x}_2+a_4\hat{x}_1^2+a_5\hat{x}_2+\zeta_i \tag{14}$$

ただし，$\zeta_i=a_1\varepsilon_1+a_2\varepsilon_2+a_3\varepsilon_1\varepsilon_2+a_4\varepsilon_1^2+a_5\varepsilon_2^2+a_3\hat{x}_1\varepsilon_2+a_3\hat{x}_2\varepsilon_1+2a_4\hat{x}_1\varepsilon_1+2a_5\hat{x}_2\varepsilon_2$

と書き改めることができよう。したがい，データに統計的誤差が含まれるときには，誤差 ζ_i をうまく考慮して後に(14)式を処理すべきではなかろうか。

第２は $y=\sum_i a_i f_i(x_1x_2\cdots\cdots x_p)$ の関数型の特定化であろう。f の役割は p 次元空間的に散布している n 個の点をいかにうまくその曲面上にのせるかにあるからである。