多変量解析法の統計的性質に関する研究
− 主成分分析における軸の回転と部分空間法における影響診断に関する研究 −

M2006MM005 林 邦好
担当教員 田中 豊

1 はじめに

本研究は, 「多変量解析法の統計的性質に関する研究」という主題のもとで行われた. この研究は大きく 2 つの副主題から構成されている. 1 部は, 筆者が卒業研究以来取り組んできた, 「主成分分析における軸の回転」という題で議論が展開される. また, 2 部では, 「部分空間法における影響診断に関する研究」という題で議論が展開される. なお, 1 部の詳細については計算機統計学 19 巻 2 号 (2007 年 8 月 24 日 採択済) を参照されたい.

2 主成分分析における軸の回転

2.1 展開

探索的因子分析においては, 回転の不定性があるために軸の回転は積極的に利用され, 解釈のし易い軸が追究される. 一方, 主成分分析では主成分特有の性質が存在し, 解の回転によりこれらの性質が欠如するという観点等から, 軸の回転は一般的にはあまり利用されない. また, 少数個の主成分について議論する場合には, 固有ベクトルに基づいて適切に軸の解釈ができる場合が多い. しかしながら, 主成分分析においても多数の主成分を採用する場合には主成分の解釈が困難となり, 因子分析の場合のように回転を利用して解釈のし易い結果が得られれば好都合である. 気象学の分野で見受けられるように, ある特定の分野では主成分の回転が有効的に利用されている分野もあり, 主成分分析での回転を研究する価値は十分あると思われる. 主成分分析での回転は, 主に 2 つの代表的なアプローチが先行研究として存在する. 1 つは Jolliffe(1987, 1995) の立場であり, もう一つは Richman(1986) の review paper で紹介されている立場である. Jolliffe は固有ベクトルのノルムの取り方を 3 通りに表現し (基準化 1~3), 主成分係数行列 A を回転する立場により, 係数の直交性と主成分スコアの無相関性に着目し議論を展開している. 一方, Richman は PC モデルを導入し, 主成分負荷量行列 W の回転を試みている. 本研究では Richman と近い立場から, PCA を次元縮約として捉え, 縮約された低次元空間の中に, 解釈に便利な座標軸を導入するという考え方で主成分分析における回転問題を定式化し, 異なる基準化間の比較や A の回転と W の回転のどちらかを行うべきかについて, 実データ及び人工シミュレーションデータを用いて詳しく議論する. その結果, 得点の無相関性が保持できる回転では A を回転するよりも W を回転した方が回転後の軸に対比が生じることがなく, 解釈がし易いことが明らかになった.

2.1.1 Jolliffe の回転の考え方

中心化された $n \times p$ データ行列を X とし, その分散共分散行列を S とする. S の第 k 番目の固有値と, それに対応する固有ベクトルを λ_k, $\underline{v}_k$(ただし, ${\underline{v}_k}^T \underline{v}_k = 1$) とすると, 第 k 主成分スコア $\underline{z}_k$ は, $\underline{z}_k = X\underline{a}_k$ のように計算されるが, 係数ベクトル $\underline{a}_k$ の長さの決め方として以下の 3 通りの場合を考えている (Jolliffe, 1989,1995).

$$(1)\underline{a}_k = {\lambda_k}^{\frac{1}{2}}\underline{v}_k, (2)\underline{a}_k = \underline{v}_k, (3)\underline{a}_k = {\lambda_k}^{-\frac{1}{2}}\underline{v}_k$$

以後, これらを基準化 1〜3 と呼ぶことにする.

通常の PCA では, 上記の基準化 2 が採用される. 基準化 1 を考える理由として, $\underline{a}_k$ が第 k 番目の主成分スコアともとの変量 X との相関係数を与えることを挙げているが, この場合第 k 番目の主成分の分散は ${\lambda_k}^2$ となり, 通常の PCA の場合の λ_k に比べて ${\lambda_k}^2$ と誇張される. 個体間のもとの変数の空間内での距離と主成分の空間内での距離について, 基準化 2 や 3 の場合のように, X の空間でのユークリッド距離やマハラノビス距離が主成分の空間のユークリッド距離で近似されるという好ましい性質を持たない.

さらに, Jolliffe は回転を係数行列 $A = (\underline{a}_1, \cdots, \underline{a}_k)$(ただし,$k < p$) に対して適用して, $A \rightarrow A^* = AT = (\underline{a}_1{}^*, \cdots, \underline{a}_k{}^*)$ のように回転し, 回転後の主成分スコアを $\underline{z}_k{}^* = X\underline{a}_k{}^*$,$Z^* = XA^* = XAT$ により求め, これらの手順により求まる, 回転後の主成分係数及び主成分得点の直交性及び無相関性について検討している.

2.1.2 Richman の回転の考え方

Richman(1986) は, $n \times p$ データ行列 X を標準化されたデータとして, $X \cong ZW^T$ という形のモデル化を試みている. Z は PC スコアに相当し, W は因子分析における primary pattern 行列 (とくに直交の時は負荷行列) に相当する. ここで回転行列 T を導入して, $X = ZW^T = Z(T^T)^{-1}T^TW^T$ とおき, 回転後の W と Z を $W^* = WT$, $Z^* = Z(T^T)^{-1}$ と定義している. また, 心理学の領域 (例えば, Green(1978)) では, PCA は因子分析の 1 つの解法 (近似) と捉えられているので, 暗に Richman と同じモデルを想定して回転を行っている. Jolliffe(1995) は回転対象として主成分係数 A を考えるのに対して, Richman(1986) や Green(1978) は回転対象として主成分負荷量 W を考えているという違いがある.

2.1.3 我々の接近法 (林, 冨田, 田中)

我々の考え方は, PCA を次元縮約の方法として捉え, 縮約された低次元 (q 次元) 空間内の座標軸の回転を考えるというものである. その意味で Richman(1986) や Green(1978) に近い. X を $n \times p$ の中心化されたデータ行列とすれば, 特異値分解を用いることにより, 以下のように分解できる.

$$X = U_q L_q{}^{\frac{1}{2}} V_q{}^T + U_{p-q} L_{p-q}{}^{\frac{1}{2}} V_{p-q}{}^T \qquad (1)$$

式 (1) において, U と V はそれぞれノルムを n に基準化した XX^T の固有ベクトル及びノルムを 1 に基準化した X^TX の固有ベクトル, L_q と L_{p-q} はそれぞれ対角要素に固有値を持つ, 対角行列 $diag(\lambda_1, \cdots, \lambda_q)$, $diag(\lambda_{q+1}, \cdots, \lambda_p)$ である. 右辺第 1 項は, p 次元データ行列 X の最小 2 乗法の意味での q 次元近似を与えることが知られている (例えば, Johnson & Wichern(1992), pp.384-387). そこで, $Z = U_q(L_q{}^{\frac{1}{2}})^{1-\alpha}$,$W = V_q(L_q{}^{\frac{1}{2}})^{\alpha}$ とおき, Z を主成分スコア, W を主成分負荷行列と呼ぶことにする. 式 (1) の両辺に右から $W(W^TW)^{-1}$ を掛けると, $XW(W^TW)^{-1} = Z$ となり, 主成分係数行列 $A = (\underline{a}_1, \cdots, \underline{a}_q)$ との間に, $A = W(W^TW)^{-1} = V_q(L_q{}^{\frac{1}{2}})^{-\alpha}$ のような関係が導かれる. 逆に $W = A(A^TA)^{-1}$ の関係も容易に行列の変形から導くことができる. q 次元近似である $X \cong ZW^T$ の Z と W を q 次元空間 (通常 $q = 2$) に同時プロットしたグラフが, バイプロットとして知られている. この q 次元空間には, 通常主成分による座標軸が導入されるが, 我々はこの空間内でより解釈のし易い座標軸の回転を考え, 回転後の行列を $*$ 印をつけて表すことにする. W に対する回転 T を施し,

$$X \cong ZW^T = Z(T^{-1})^T T^T W^T = Z^* W^{*T} \qquad (2)$$

とする. 式 (2) より, $W \to W^* = WT$ と回転する時, スコアは $Z \to Z^* = Z(T^{-1})^T = XA(T^{-1})^T$, 特に直交回転の場合においては, $Z \to Z^* = Z(T^{-1})^T = XAT$ となり, 回転後の行列 A は $A^* = A(T^{-1})^T$, 直交回転なら $A^* = AT$ となることがわかる. 回転後の式 (1) は以下のように記述される.

$$\underset{n\times p}{\boldsymbol{X}} = Z^* W^{*T} + U_{p-q} L_{p-q}{}^{\frac{1}{2}} V_{p-q}{}^T \qquad (3)$$

式 (3) の右から $W^*(W^{*T}W^*)^{-1}$ を掛けて整理すると,

$$Z^* = XW^*(W^{*T}W^*)^{-1} = XA(T^{-1})^T = XA^* \qquad (4)$$

となり, 回転後においても $A^* = W^*(W^{*T}W^*)^{-1}$ の関係が成り立つ. この関係を $A^*(A^{*T}A^*)^{-1}$ に代入すると $A^*(A^{*T}A^*)^{-1} = W^*$ が成り立つことがわかる. 以上より, W, A, Z の回転について次の 1), 2), 3) の回転が対応することがわかる.

- 直交回転の時: $1)W \to W^* = WT$, $2)A \to A^* = AT$, $3)Z \to Z^* = ZT$
- 斜交回転の時: $1)W \to W^* = WT$, $2)A \to A^* = A(T^{-1})^T$, $3)Z \to Z^* = Z(T^{-1})^T$

回転は基準化の仕方によっては, A の単純構造を実現することと, W の単純構造を実現することとは同じではない. そこで我々は, 回転後の PCA の軸に対比が生じない場合を解釈のし易い軸と定義し, 回転後の A と W でどちらがより解釈のし易い軸 (つまり, 対比の軸が生じにくいか) を実データや人工データにより解析した. 本研究では回転行列 T として特に直交回転のみを考える.

2.2 A の回転と W の回転について

我々や Richman(1986) が回転対象としている A と Jolliffe(1989,1995) が回転対象としている W には, $A = W(W^TW)^{-1}$ という関係が 2.2.3 節でも言及したように存在する. このため, 直交回転の場合について, 2.2.1 節で触れた Jolliffe(1995) の 3 つの基準化に対応する A と W の形や回転前後の各性質について整理する.

- 基準化 1: $A = VL^{\frac{1}{2}} \Leftrightarrow W = VL^{-\frac{1}{2}}$
- 基準化 2: $A = V \Leftrightarrow W = V$
- 基準化 3: $A = VL^{-\frac{1}{2}} \Leftrightarrow W = VL^{\frac{1}{2}}$

表 1 主成分係数 A と主成分負荷行列 W の直交性

	回転前		回転後	
基準化	A^TA	W^TW	T^TA^TAT	T^TW^TWT
1	L	L^{-1}	T^TLT	$T^TL^{-1}T$
2	I	I	I	I
3	L^{-1}	L	$T^TL^{-1}T$	T^TLT

表 2 回転前後の主成分得点の無相関性

基準化	回転前	回転後
1	L^2	T^TL^2T
2	L	T^TLT
3	I	I

表 1 から回転後の直交性は基準化 2 のみにおいて成立し, 表 2 から回転後の無相関性は基準化 3 のみにおいて成立することがわかる. 基準化 3 では, A を回転した場合でも W を回転した場合でも, 回転後のスコアの無相関性は保持されているが, 生成される係数値は異なる. ゆえに, 人工データ等を用いて, 基準化 3 における両者の係数値符号のシミュレートすることは有意である.

2.3 人工データによる符号のシミュレーション

符号シミュレーションの結果, 回転後の A よりも回転後の W の方が対比の軸を生成しにくい傾向が読み取れた. シミュレーションの方法, 結果の詳細と現象の理論的説明は計算機統計学 19 巻 2 号を参照のこと.

3 部分空間法における影響診断に関する研究

3.1 展開

認知あるいは認識という活動は我々人間を含めた生物が最も得意とする活動に他ならない．しかしながら, 鳥脇 (2002) によれば, 1940 年代後半に電子計算機がこの世界に現れてから既に多くの時間が経過しているにも関わらず, この認知あるいは認識という活動を機械に行わせるパターン認識の問題全域を覆う基礎理論は未だ確立されていないという．本研究の 2 部ではこのようなパターン認識分野の識別手法である Watanabe(1967) が提案した部分空間法 (subspace method) の中の CLAFIC 法という手法に関心が向けられることになる．認識問題には (a) 音声認識, (b) 文字認識, (c) 物体認識, (d) 画像認識等のように種々の認識問題が存在するが, どのような識別器であれ構築した識別器の分析結果の安定性を把握することは, 技術者の倫理の観点からは不可欠である．統計学の分野では, この分析結果の安定性に関する研究というものが 1970 年代後半から主に回帰分析を中心に熟してきた．これらについては, Belsley, Kuh & Welsch(1980) や Cook & Weisberg(1982) や Atkinson(1985) 等の文献に要約されている．その後, 多変量解析の各手法に対して安定性を評価する手法が開発された．例えば, Campbell(1978) は, 判別分析に関する影響関数を導出した．また, Sibson(1979) は古典的な多次元尺度法 (MDS) における摂動解析を提案している．Tanaka(1984) や Tanaka & Tarumi(1986) は, 林の数量化法における感度分析を提案している．つまり, 統計学の分野では種々の解析手法の感度分析法の構築が行われてきている．このような視点から, もう一度パターン認識の分野に立ち返った時, 分析結果の安定性の議論は比較的数少ないと考えられる．本研究の 2 部では, 2 クラスの線形部分空間法に感度分析法を導入する．

3.2 部分空間法における感度分析法の提案

本予稿では紙面の都合上, 判別スコアの変化量の平均 (接近法 1) による経験影響関数を用いた単数観測値診断及び複数観測値診断, 標本影響関数を用いた単数観測値診断の解説にとどめる．

3.3 判別スコアの変化量の平均を用いた方法

前節で述べたように本節において, 提案手法の 1 つである判別スコアの変化量の平均 (接近法 1) による経験影響関数を用いた単数及び複数観測値診断を紹介する．

3.3.1 経験影響関数を用いた場合 (single version)

はじめに, 単数観測値診断の提案手法を紹介する．2 群の判別関数を以下のように定義する．

$$\begin{cases} z_{1i} = (\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i})^{\frac{1}{2}} - (\mathbf{x}_{1i}{}^T P_2 \mathbf{x}_{1i})^{\frac{1}{2}}, \ \ i = 1, \cdots, n_1 \\ z_{2j} = (\mathbf{x}_{2j}{}^T P_2 \mathbf{x}_{2j})^{\frac{1}{2}} - (\mathbf{x}_{2j}{}^T P_1 \mathbf{x}_{2j})^{\frac{1}{2}}, \ \ j = 1, \cdots, n_2 \end{cases} \tag{5}$$

(5) 式における P_1 は類相関行列 G_1 の subspace への射影子である．P_2 は類相関行列 G_2 の subspace への射影子である．z_{1i}, z_{2j} とも大きいほどよりよく判別することになる．さて, $k \in G_1$ に摂動を導入する．ここでの摂動とは影響関数タイプの摂動であり, $z_{1i} \to z_{1i} + \varepsilon \Delta_k z_{1i}$ を考えることに相当する．影響関数の定義の観点から, 摂動展開は ε の 1 次の係数までにとどめる．

$$z_{1i} + \varepsilon \Delta_k z_{1i} = ||(P_1 + \varepsilon P_{1k}{}^{(1)})\mathbf{x}_{1i}|| - ||P_2 \mathbf{x}_{1i}|| =$$

$$(\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i})^{\frac{1}{2}} \{1 + \varepsilon \frac{\mathbf{x}_{1i}{}^T (P_1{}^T P_{1k}{}^{(1)} + P_{1k}{}^{(1)T} P_1)\mathbf{x}_{1i}}{\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i}} +$$

$$\varepsilon^2 \frac{\mathbf{x}_{1i}{}^T P_{1k}{}^{(1)T} P_{1k}{}^{(1)} \mathbf{x}_{1i}}{\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i}} \}^{\frac{1}{2}} - (\mathbf{x}_{1i}{}^T P_2 \mathbf{x}_{1i})^{\frac{1}{2}} \tag{6}$$

(6) 式の前半を Taylor 展開により整理する．

$$z_{1i} + \varepsilon \Delta_k z_{1i} = (\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i})^{\frac{1}{2}} - (\mathbf{x}_{1i}{}^T P_2 \mathbf{x}_{1i})^{\frac{1}{2}} +$$

$$\frac{\varepsilon}{2} \cdot \frac{\mathbf{x}_{1i}{}^T (P_1{}^T P_{1k}{}^{(1)} + P_{1k}{}^{(1)T} P_1)\mathbf{x}_{1i}}{(\mathbf{x}_{1i}{}^T P_1 \mathbf{x}_{1i})^{\frac{1}{2}}} + O(\varepsilon^2) \tag{7}$$

同様にして $k \in G_1$ における $z_{2j} \to z_{2j} + \varepsilon \Delta_k z_{2j}$, $k' \in G_2$ における $z_{1i} \to z_{1i} + \varepsilon \Delta_{k'} z_{1i}$ 及び $k' \in G_2$ における $z_{2j} \to z_{2j} + \varepsilon \Delta_{k'} z_{2j}$ を導出する．これらを用いて k 番目 $(k = 1 \sim n_1 + n_2)$ に摂動を与えた時の判別スコアの変化量は以下のように定式化できる．

$k(\in G_1)$ の時

$$\varepsilon \cdot \Delta_k \bar{z}_{..} \equiv \varepsilon \cdot \frac{n_1 \cdot \Delta_k \bar{z}_{1\cdot} + n_2 \cdot \Delta_k \bar{z}_{2\cdot}}{(n_1 + n_2)} \tag{8}$$

$k'(\in G_2)$ の時

$$\varepsilon \cdot \Delta_{k'} \bar{z}_{..} \equiv \varepsilon \cdot \frac{n_1 \cdot \Delta_{k'} \bar{z}_{1\cdot} + n_2 \cdot \Delta_{k'} \bar{z}_{2\cdot}}{(n_1 + n_2)} \tag{9}$$

判別スコアの変化量の平均を用いた方法とした上記の議論では, ユークリッド距離を用いているが, ユークリッド 2 乗距離を用いることもできる．ユークリッド 2 乗距離を用いた方法ではそのまま, 複数観測値診断のへの拡張が容易にできる．2 群の分離の大きさは以下の式により与えることができる．

$$\begin{cases} z_i = \mathbf{x}_i{}^T Q \mathbf{x}_i, \ \ Q = \begin{cases} P_1 - P_2, \ i = 1 \sim n_1 \\ P_2 - P_1, \ i = n_1 + 1 \sim n_1 + n_2 \end{cases} \end{cases} \tag{10}$$

さて, ここで z_i に関して $k(k = 1 \sim n_1 + n_2)$ 番目の摂動を考える．

$$z_i \to \tilde{z}_i = \mathbf{x}_i{}^T (Q + \varepsilon Q_k{}^{(1)}) \mathbf{x}_i \tag{11}$$

ゆえに, 摂動後の全体の変化量は以下の通りとなる．

$$\tilde{Z} = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} \mathbf{x}_i{}^T Q \mathbf{x}_i + \varepsilon \cdot \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} \mathbf{x}_i{}^T Q_k{}^{(1)} \mathbf{x}_i \tag{12}$$

つまり, 判別スコアの変化量の平均を用いた提案手法では, (12) 式の第 2 項の視覚的なインデックスプロットにより影響のある観測値を見い出すことができる．

3.3.2 標本影響関数を用いた場合 (single version)

k 番目の個体を取り除いた判別スコアともともとの判別スコアから定義できる以下の標本影響関数を用いる.

$$\Delta_k z = -(n_1+n_2-1)\cdot\{\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}\mathbf{x}_i{}^T Q_{(-k)}\mathbf{x}_i -$$

$$\frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2}\mathbf{x}_i{}^T Q\mathbf{x}_i\} \tag{13}$$

3.3.3 経験影響関数を用いた場合 (multiple version)

複数の観測値診断は影響関数の加法性を用いて行われる (Tanaka, 1994). これは, 分布関数 F から $\tilde{F} = (1-\varepsilon)F + \varepsilon G$ への摂動により以下が成立するためである. ただし G は $G = k^{-1}\sum_{\mathbf{x}_i\in A}\delta\mathbf{x}_i$ である.

$$TIF(A;\theta) \equiv \theta_A{}^{(1)} = \lim_{\varepsilon\to 0}[\theta(\tilde{F}) - \theta(F)]/\varepsilon =$$

$$k^{-1}\cdot\sum_{\mathbf{x}_i\in A} TIF(\mathbf{x}_i;\theta) \tag{14}$$

式 (14) から以下の手順が踏襲される.

(1) 経験影響関数であるベクトル $\hat{\theta}_i^{(1)}$ を計算する.
(2) 個々に影響を与えているものや同じような影響のパターンをもつ集合を PCA により探索する.

つまり, 式 (12) の右辺第 2 項の部分が大きくなるような集合を見つけることが最大の関心事となる. そこで ${Q_k}^{(1)}$ が大きく, かつ類似の方向を向いた観測値の集合を探索することを行う. これには $vech[{Q_k}^{(1)}]$ の PCA を用いることができる.

4 実データへの応用

3.3 節で述べた提案手法をここでは実データに対して適用してみることにする. 用いるデータには, proben1 の公開データベースの中から, University of Massachusetts の視覚グループにより与えられている画像切り出しデータを採用する. このデータは 7 つの画像の切り出しデータであり, 各パターンベクトルは 19 個の特徴変量により定義されている. この 7 つの画像の内で, PATH と GRASS の 2 つの識別を行うことにする. 各画像の 30 例の訓練データを用いて提案した感度分析を実演する. 各群の射影行列の形成には各群第 1~4 基底を用いた (閾値を 90.0% に設定). 単数観測値診断, 経験影響関数の PCA の分析結果及び複数の観測値診断の結果を以下に示す. 図 1 の左側から単数観測値診断では観測値 40, 49, 58, 60 が識別器に大きな影響を与えていることが分かる. 表 3 の経験影響関数の PCA の分析結果から, 経験影響関数のデータが第 1~2 までの主成分によりほぼ説明できることが分かる. このことから, 各主成分スコアを布置すると図 2 となる. 図 2 から, 観測値 40 と 46 は同じ方向にあることが分かる. 複数の観測値診断により, 同じ影響を与える観測値の集落を見つけることができる.

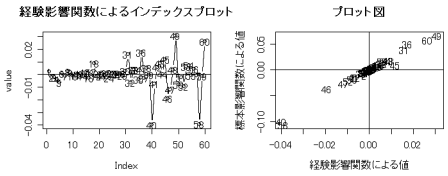

図 1 単数観測値診断

表 3 経験影響関数 $vech[{Q_k}^{(1)}]$ の PCA の分析結果

\	λ_1	λ_2	λ_3	$\cdots$	λ_{190}
固有値	1.202	0.915	0.080	$\cdots$	0.000
累積寄与率 (%)	51.4	90.5	94.0	$\cdots$	100.0

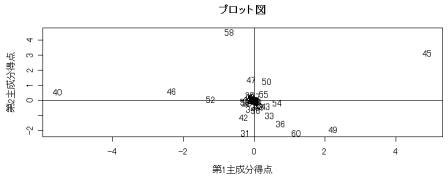

図 2 複数観測値診断

5 研究成果と今後の課題

本研究の 2 部では CLAFIC 法に感度分析法を導入することに成功した. 今後は種々の部分空間法の拡張手法に導入すること等がやるべき事として残っている.

参考文献

[1] http://www.cs.utoronto.ca/ ˜ delve/data/image-seg/desc.html

[2] Norm A. Campbell. (1978). The influence function as an aid in outlier detection in discriminant analysis, Appl.Statist, **27**, 3, 251-258.

[3] Watanabe, S., Lambert, P.F., Kulikowski, C.A., Buxton, J.L. and Walker, R. (1967). Evaluation and selection of variables in pattern recognition, in Computer and Information Sciences II. (Edited by J.T.Tou), Academic Press.

[4] Yutaka Tanaka. (1994). Recent advance in sensitivity analysis in multivariate statistical methods, J.Jpn.Soc.Comp.Statist., **7**, 1-25.

[5] Yutaka Tanaka. (1988). Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components, Commun. Statist. Theory Meth., **17**, 9, 3157-3175.

[6] 鳥脇純一郎. (2002). 認識工学 (パターン認識とその応用), コロナ社.

[7] 林邦好, 冨田誠, 田中豊. (2007). 主成分分析における軸の回転について, 計算機統計学, **19**, 2, 1-13.