HOKKAIDO UNIVERSITY

Title	一般化主成分分析法とその改良について
Author(s)	水田, 正弘
Citation	北海道大學工學部研究報告, 141, 209-215
Issue Date	1988-07-29
Doc URL	https://hdl.handle.net/2115/42119
Type	departmental bulletin paper
File Information	141_209-216.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第141号（昭和63年）

Bulletin of the Faculty of Engineering,
Hokkaido University. No. 141 (1988)

一般化主成分分析法とその改良について

水　田　正　弘

（昭和63年3月31日受理）

Generalized Principal Components Analysis and an Improvement in it

Masahiro MIZUTA

(Received March 31, 1988)

Abstract

Generalized Principal Components Analysis (GPCA) is explained briefly and an improvement in GPCA is proposed in this paper. GPCA is one of the methods that determine non-linear structures of the data. Most of the properties of 'non-linear' modes are not dependent on a coordinate system, but the results of GPCA are not always invariant under orthogonal transformations, parallel translations and similarities of the coordinate system. An improvement in GPCA is proposed in order to cope with this defect. The proposed method is invariant under these transformations of the coordinate system.

1. はじめに

最近，必ずしも線形ではない構造に対するデータ解析の方法が活発に研究されている。例えば，多変量データにおいて，変量が1個の目的変量と数個の説明変量に分かれているときに，その目的変量を説明変量の非線形な結合によって表す手法が数多く開発されている。しかし，変量を目的変量と説明変量に分けることが不可能な場合，すなわち，各変量が同等で等しく扱わなくてはならない場合における非線形な構造の解析方法はあまり研究されていない。Gnanadesikan[1]によって開発された一般化主成分分析法は，このようなデータに対する数少ない手法の一つである。一般化主成分分析法は通常の主成分分析法の拡張になっており，計算もかなり容易である。また一般化主成分分析法の解はデータに対する当てはめとなっているので幅広く適用できる可能性を持っている。そこで本報告では，はじめに一般化主成分分析法について簡単に説明する。

しかし，一般化主成分分析法には，利用する関数の選び方をはじめ解決しなくてはいけない問題が残されている。Mizuta[4]によって，座標系の直交変換に関して結果が変わらない一般化主成分分析法が数学的に特定されたが，座標系の平行移動および相似変換に関して結果が変わらない手法は発表されていない。そこで，本報告の後半では，これらの座標変換に関しても不変な手法を一般化主成分分析法の改良として報告する。

情報科学

2. 一般化主成分分析法と座標変換に関する不変性

はじめに一般化主成分分析法の紹介を行う。N個の個体に対して，それぞれp変量の観測値$\boldsymbol{x}=(x_1, x_2, \cdots, x_p)$が得られたとする。ここで$f_i(\boldsymbol{x})\ (i=1, 2, \cdots, k)$を$p$次元ベクトル$\boldsymbol{x}$について定義された$k$個の実数値関数とし，$\boldsymbol{F}(\boldsymbol{x})=(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_k(\boldsymbol{x}))'$と置く。この$\boldsymbol{F}(\boldsymbol{x})$によって決められる一般化主成分分析法とは，$f_i(\boldsymbol{x})\ (i=1, 2, \cdots, k)$の一次結合で表される変数（$\boldsymbol{x}$の関数）のうち分散が最大な変数を$z_1$とし，$z_1$と無相関でかつ分散が最大が変動を$z_2$，以下同様に変数$z_3, z_4, \cdots, z_k$を選ぶ方法である。ただし，

$$z_j=\sum_{i=1}^{k} l_{ij} f_i(x)=\boldsymbol{l}_j'\boldsymbol{F}(\boldsymbol{x}),$$

$$\boldsymbol{l}_j'=(l_{1j}, l_{2j}, \cdots, l_{kj}),$$

$$\boldsymbol{l}_j'\boldsymbol{l}_j=\sum_{i=1}^{k} l_{ij}^2=1, \quad (j=1, 2, \cdots, k) \tag{1}$$

とする。主成分分析法の場合と同様にベクトル$\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_k$は$(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_k(\boldsymbol{x}))$の分散共分散行列の固有ベクトルである。従って実際の計算は次の様に行う。

1) p変量データ$\{\boldsymbol{x}\}$を$\boldsymbol{F}(\boldsymbol{x})$によって$k$変量データ$\{\boldsymbol{F}(\boldsymbol{x})\}$に変換する。

2) データ$\{\boldsymbol{F}(\boldsymbol{x})\}$の分散共分散行列$\mathrm{Cov}(\boldsymbol{F}(\boldsymbol{x}))$を計算し，その固有値$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_k$および固有ベクトル$\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_k$を求める。ただし$\boldsymbol{l}_i$は固有値$\lambda_i$に対応する固有ベクトルとする。

3) 2)で求めた固有ベクトルによって変数（主成分）

$$z_1=\sum_{i=1}^{k} l_{i1} f_i(\boldsymbol{x})=\boldsymbol{l}_1'\boldsymbol{F}(\boldsymbol{x})$$

$$\cdots$$

$$z_k=\sum_{i=1}^{k} l_{ik} f_i(\boldsymbol{x})=\boldsymbol{l}_k'\boldsymbol{F}(\boldsymbol{x})$$

を得る。

最小固有値に対応するz_kは，$f_i(\boldsymbol{x})\ (i=1, 2, \cdots, k)$の一次結合（ただし係数の平方和は1）のうち分散が最も小さいものである。すなわちz_kはデータに対して適合する関数$z_k(\boldsymbol{x})$であると考えられる。特に，p次元空間における曲線$\{\boldsymbol{x} \in R^p;\ \bar{z}_k=z_k(\boldsymbol{x})\}$はデータに対する当てはめ（フィッティング）と見なすことができる。ただし，$\bar{z}_k$は$z_k(\boldsymbol{x})$の平均とする。

関数$\boldsymbol{F}(\boldsymbol{x})$をいろいろ選ぶことによって各種の一般化主成分分析法の手法が得られる。例えば，$f_i(\boldsymbol{x})=x_i\ (i=1, 2, \cdots, p)$と置くと通常の主成分分析法となる。またGnanadesikanが2次の主成分分析法と呼んでいるものは

$$f_i(\boldsymbol{x})=x_i \qquad (i=1, 2, \cdots, p)$$

$$f_i(\boldsymbol{x})=x_j x_m \qquad (i=p+1, \cdots, (p^2+3p)/2)$$

によって定義される。ただし，jとmは$p \geqq j \geqq m \geqq 1$なるすべての組み合わせをとる。2変量$(p=2)$のとき2次の主成分分析法は，

$$f_1(x, y)=x \quad f_2(x, y)=y \quad f_3(x, y)=x^2 \quad f_4(x, y)=xy \quad f_5(x, y)=y^2$$

で定義される。ここで$\boldsymbol{x}=(x, y)$とする。

形式的には$\boldsymbol{F}(\boldsymbol{x})$として任意の関数を選んでも一般化主成分分析法が定義できる。しかし，意味のある解析を行うために最小限必要な次の条件を仮定する。

条件1)　$f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$, …, $f_k(\boldsymbol{x})$は$\boldsymbol{x}$の関数として一次独立である。

条件2)　任意のp次直交行列Tに対して$\boldsymbol{F}(\mathrm{T}\boldsymbol{x})\equiv \mathrm{W}\boldsymbol{F}(\boldsymbol{x})$となる行列Wが存在する。

条件3)　$f_i(\boldsymbol{x})$は$\boldsymbol{x}$について連続関数である。

以下ではこれらの条件のもとで，座標系の直交変換に対して不変な一般化主成分分析法について述べる。

一般化主成分分析法の手法が『不変』であるとは，座標系を直交変換したとき得られる結果を元の座標系で考えると，元の座標系から直接得られた結果と常に一致していることを言う。そこで，このことを数学的に述べる。

座標系の任意な直交変換$\boldsymbol{x}^*=\mathrm{T}\boldsymbol{x}$について，$\boldsymbol{x}^*$を変量と見なしたときの結果を

$$z_j=\boldsymbol{l}_j^{*\prime}\boldsymbol{F}(\boldsymbol{x}^*)=\boldsymbol{l}_j^{*\prime}\boldsymbol{F}(\mathrm{T}\boldsymbol{x}) \qquad (j=1,\ 2,\ \cdots,\ k)$$

で表す。ここで$\boldsymbol{l}_j^*$は$\boldsymbol{F}(\boldsymbol{x}^*)$の分散共分散行列$\mathrm{Cov}(\boldsymbol{F}(\boldsymbol{x}^*))$の固有ベクトルである。手法が『不変』であるとは，任意の直交行列Tについて

$$\boldsymbol{l}_j'\boldsymbol{F}(\boldsymbol{x})\equiv\pm\boldsymbol{l}_j^{*\prime}\boldsymbol{F}(\mathrm{T}\boldsymbol{x}) \qquad (j=1,\ 2,\ \cdots,\ k)$$

がxの関数として成立することである。ただし，〝±〟の記号は単に固有ベクトルの向きを示しているだけで本質的ではない。

前述の条件1)，2)，3)のもとで座標系の直交変換に対して不変な手法は以下の関数により決定されるものに限り，逆に以下の関数により決定される手法は不変である[4]。

1)　つぎの二つからなる関数の組

$$f_{2i-1}(x,y)=g_i\left(x^M-\binom{M}{2}y^2x^{M-2}+\binom{M}{4}y^4x^{M-4}+\cdots\right)-h_i\left(My\,x^{M-1}-\binom{M}{3}y^3x^{M-3}+\binom{M}{5}y^5x^{M-5}-\cdots\right),$$

$$f_{2i}(x,y)=\pm\left\{g_i\left(My\,x^{M-1}-\binom{M}{3}y^3x^{M-3}+\binom{M}{5}y^5x^{M-5}-\cdots\right)+h_i\left(x^M-\binom{M}{2}y^2x^{M-2}+\binom{M}{4}y^4x^{M-4}-\cdots\right)\right\}$$

$$M=M_i \qquad (j=1,\ 2,\cdots,s)$$

ただし，g_i, h_iは$\sqrt{x^2+y^2}$についての任意な連続関数，M_iは任意な正の整数とする。

2)　いくつかの$\sqrt{x^2+y^2}$の連続関数。

補注：厳密に言えば不変な手法はベクトル値関数$\mathrm{P}\boldsymbol{F}(x,y)$による決まるものに限る。ここで$\boldsymbol{F}(x,y)$は上述の1)，2)からなるベクトル値関数，Pは直交行列とする。ただし，一般に$\boldsymbol{F}(\boldsymbol{x})$で決められる手法から得られる結果と$\mathrm{P}\boldsymbol{F}(\boldsymbol{x})$で決められる手法から得られる結果は一致する。

以上で2変量データについての不変な手法を決める関数の一般形を示したが，多項式の場合の関数を具体的な形で紹介する。各次数に応じて次の関数系がある。

2変量1次多項式：　x, y

2変量2次多項式：　x^2, $\sqrt{2}\,xy$, y^2

2変量3次多項式：x^3, $\sqrt{3}\,x^2y$, $\sqrt{3}\,xy^2$, y^3

2変量M次多項式：x^M, $\sqrt{\binom{M}{1}}\,x^{M-1}y$, $\sqrt{\binom{M}{2}}\,x^{M-2}y^2$, …, $\sqrt{\binom{M}{M-1}}\,xy^{M-1}$, y^M

これらは一般形から得られる関数系に適当な直交行列Pを掛けることにより求めることができる。

一般のp変量データにおいて不変な一般化主成分分析法の手法を定義する$\boldsymbol{F}(\boldsymbol{x})$は，任意の2変量についての直交変換に対して不変な一般化主成分分析法の手法を求めれば良い。例えば3変量(x,y,z)では，(x,y)，(y,z)，(z,x)についての直交変換に対して結果が変わらない一般化主成

分分析法の手法は3変量全体について見ても不変な手法である。

従って3変量以上では次のような関数系がある。

3変量1次多項式： $x,\ y,\ z$

3変量2次多項式： $x^2,\ y^2,\ z^2,\ \sqrt{2}xy,\ \sqrt{2}yz,\ \sqrt{2}zx$

3変量3次多項式： $x^3,\ y^3,\ z^3,\ \sqrt{3}x^2y,\ \sqrt{3}y^2z,\ \sqrt{3}z^2x,\ \sqrt{3}xy^2,\ \sqrt{3}yz^2,\ \sqrt{3}zx^2,\ \sqrt{6}xyz$

3変量 q 次多項式： $\sqrt{\dfrac{q!}{i!\,j!\,k!}}\,x^iy^jz^k$

$$(i+j+k=q\ ;\ 0\leqq i,j,k),$$

p 変量 q 次多項式： $\sqrt{\dfrac{q!}{\prod_{t=1}^{p}(i_t!)}}\prod_{t=1}^{p}(x_t)^{i_t}$

$$\sum_{t=1}^{p} i_t=q:\ 0\leqq i_t,$$

3. 一般化主成分分析法の改良

前章で紹介した，座標系の直交変換に関して不変な一般化主成分分析法も，座標系の平行移動や相似変換に関しては必ずしも不変ではない。前章と同様な方法で議論することにより，平行移動や相似変換に関して不変な一般化主成分分析法を求めることができる。しかし，これら全ての変換に対して不変な一般化主成分分析法は（通常の）主成分分析法に限定される。そこで，本章では一般化主成分分析法の改良として，これらの座標変換に関して不変な手法を提案する。

一般化主成分分析法を別の言い方をすると，（2変量2次では）

$$z=ax+by+cx^2+d\sqrt{2}xy+ey^2$$

と置いて，$a^2+b^2+c^2+d^2+e^2=1$の条件のもとで，z の分散を最大または最小にする a，b，c，d，e を求めることと言える。しかし，これでは関数族 (x,y) と $(x^2,\sqrt{2}xy,y^2)$ を同等に扱っているために，座標系の平行移動や相似変換に関して結果が不変ではない。そこで，条件 $a^2+b^2+c^2+d^2+e^2=1$ を $c^2+d^2+e^2=1$に代えることにより，関数族$(x^2,\ \sqrt{2}xy,\ y^2)$の係数のみに制約を加えることが考えられる。すなわち，第1主成分 z_1を，

$$\max_{c^2+d^2+e^2=1}\ \min_{a,\ b}\ (\mathrm{Var}\,z) \tag{2}$$

となる a，b，c，d，e による z として定義する。z_2は(2)式に z_1から得られた係数ベクトル（c, d, e）と直交するとの条件を加えたものである。さらに，z_3も同様に決定できる。z_3は

$$z_3=\min_{c^2+d^2+e^2=1}\ \min_{a,\ b}\ (\mathrm{Var}\,z)$$

と表現できる。以下，$\boldsymbol{l}^{(1)}=(a,b)'$，$\boldsymbol{l}^{(2)}=(c,d,e)'$，$(x,y,x^2,\sqrt{2}xy,y^2)$の分散共分散を

$$\Sigma=\begin{pmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{pmatrix}$$

とする。ただし，Σ_{11}は (x,y) の分散共分散行列，Σ_{22}は $(x^2,\sqrt{2}xy,y^2)$の分散共分散行列，Σ_{12}, Σ_{21}は (x,y) と $(x^2,\sqrt{2}xy,y^2)$ の共分散行列である。説明の都合上 Σ_{11}は正則行列とする（正則

行列でなければ，通常の主成分分析法を一度適用すべきである)。Var z はこれらの記号を使うと

$$\begin{aligned}\mathrm{Var}\, z &= (\boldsymbol{l}^{(1)\prime},\ \boldsymbol{l}^{(2)\prime})'\Sigma(\boldsymbol{l}^{(1)},\ \boldsymbol{l}^{(2)}) \\ &= \boldsymbol{l}^{(1)\prime}\Sigma_{11}\boldsymbol{l}^{(1)}+\boldsymbol{l}^{(1)\prime}\Sigma_{12}\boldsymbol{l}^{(2)}+\boldsymbol{l}^{(2)\prime}\Sigma_{21}\boldsymbol{l}^{(1)}+\boldsymbol{l}^{(2)\prime}\Sigma_{22}\boldsymbol{l}^{(2)}\end{aligned} \tag{3}$$

となる。

はじめに，$\boldsymbol{l}^{(2)}=(c, d, e)'$ を固定したときの，$\boldsymbol{l}^{(1)}=(a, b)'$ に関するVar zの最小値を考察する。

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{l}^{(1)}}\mathrm{Var}\, z &= 2\Sigma_{11}\,\boldsymbol{l}^{(1)}+2\Sigma_{12}\,\boldsymbol{l}^{(2)} \\ &= 0\end{aligned}$$

より，

$$\boldsymbol{l}^{(1)} = -\Sigma_{11}^{-1}\Sigma_{12}\,\boldsymbol{l}^{(2)} \tag{4}$$

でなくてはならない。(4)式を(3)式に代入すると

$$\mathrm{Var}\, z = \boldsymbol{l}^{(2)\prime}\,(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})\,\boldsymbol{l}^{(2)}$$

となる。したがって，$\boldsymbol{l}^{(2)\prime}\boldsymbol{l}^{(2)}=1$の条件下でのVar zの最大値，最小値は，それぞれ$(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$の最大固有値，最小固有値に対応し，そのときの $\boldsymbol{l}^{(2)}$ の値はそれぞれの固有ベクトルとなる。

以上のことから，a, b, c, d, e を求める計算方法は次の通りである。

1) $(x, y, x^2, \sqrt{2}xy, y^2)$の分散共分散 Σ を求める。

2) $\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$（3 × 3 行列）の固有値問題を解き，最大固有値に対応する固有ベクトルから順に $\boldsymbol{l}_i^{(2)}=(c_i, d_i, e_i)\,(i=1,\ 2,\ 3)$ と置く。

3) $\boldsymbol{l}^{(1)} = -\Sigma_{11}^{-1}\Sigma_{12}\,\boldsymbol{l}^{(2)}$

により $\boldsymbol{l}_i^{(1)}=(a_i, b_i)$ を計算する。

1)，2)，3)により得られた$(a_i, b_i, c_i, d_i, e_i)\,(i=1,\ 2,\ 3)$が求める値である，$i=1$のとき z の分散は最大，$i=3$のとき z の分散は最小となる。

一般的に表現すると，$\boldsymbol{F}(\boldsymbol{x})=(f_1(\boldsymbol{x}),\ f_2(\boldsymbol{x}),\ \cdots,\ f_k(\boldsymbol{x}))'$において，

$$\boldsymbol{F}(\boldsymbol{x}) = (\boldsymbol{F}_1(\boldsymbol{x})',\ \boldsymbol{F}_2(\boldsymbol{x})')'$$

$$\boldsymbol{F}_1(\boldsymbol{x}) = (f_1(\boldsymbol{x}),\ f_2(\boldsymbol{x}),\ \cdots,\ f_{r-1}(\boldsymbol{x}))'$$

$$\boldsymbol{F}_2(\boldsymbol{x}) = (f_r(\boldsymbol{x}),\ f_{r+1}(\boldsymbol{x}),\ \cdots,\ f_k(\boldsymbol{x}))'$$

（r は $r>1$なる整定数）

$$\boldsymbol{l} = (\boldsymbol{l}^{(1)\prime},\ \boldsymbol{l}^{(2)\prime})'$$

$$\boldsymbol{l}^{(1)} = (l_1,\ l_2,\ \cdots,\ l_{r-1})'$$

$$\boldsymbol{l}^{(2)} = (l_r,\ l_{r+1},\ \cdots,\ l_k)'$$

$$\begin{aligned}z &= \boldsymbol{l}'\boldsymbol{F}(\boldsymbol{x}) \\ &= \boldsymbol{l}^{(1)\prime}\boldsymbol{F}_1(\boldsymbol{x})+\boldsymbol{l}^{(2)\prime}\boldsymbol{F}_2(\boldsymbol{x})\end{aligned}$$

とおき，

$$\max_{\boldsymbol{l}^{(2)\prime}\boldsymbol{l}^{(2)}=1}\quad \min_{\boldsymbol{l}^{(1)}}\ (\mathrm{Var}\, z)$$

などを $\boldsymbol{l}$ について解くことになる。この場合も 2 変量 2 次のときと全く同じ計算方法で求めることができる。

これらのときも $\boldsymbol{F}(\boldsymbol{x})$ として 2 章で紹介した関数族を利用する限り，この新たな手法は座標系の直交変換に関して不変となる。さらに，多項式の関数系を使い最高次の係数のみに制約式を科すと座標系の平行移動や相似変換に関して不変となる。

4. 数 値 例

次に数値例を紹介する。サンプル数47の2変量データ（平均0；各変量の分散1）に，次の3種類の手法を適用する。

手法1) $\boldsymbol{F}(x, y) = (x, y, x^2, xy, y^2)'$による一般化主成分分析法（図1）

手法2) $\boldsymbol{F}(x, y) = (x, y, x^2, \sqrt{2}xy, y^2)'$による一般化主成分分析法（図2）

手法3) $\boldsymbol{F}(x, y) = (x, y, x^2, \sqrt{2}xy, y^2)'$による新たな手法（図3）

それぞれ，座標系を45°回転させた結果，x軸，y軸ともに＋1.0ずつ平行移動させた結果も描いている。

この結果からも分かるように，手法1)（Gnanadesikanの2次の主成分分析法）は回転・直交変換にも平行移動にも不変ではない。手法2)は回転・直交変換には不変であるが，平行移動には不変ではない。それに対して手法3)はこれらの変換に対して不変である。

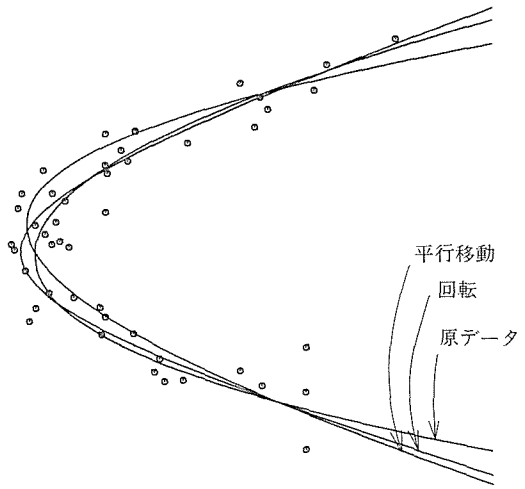

図1 2次の主成分分析法（手法1）

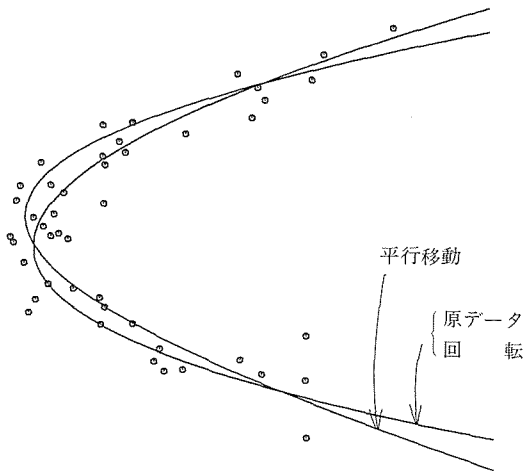

図2 直交変換に不変な一般化主成分分析法（手法2）

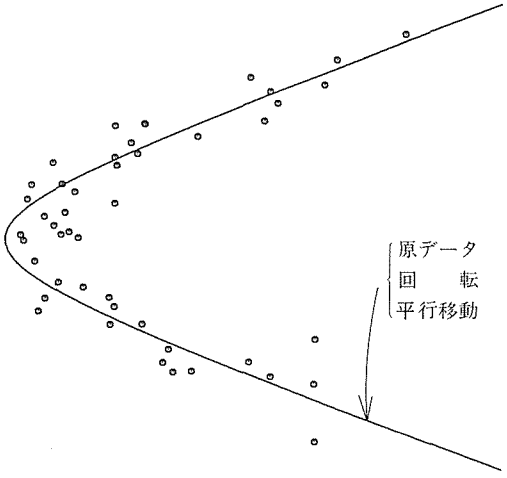

図3　一般化主成分分析法の改良（手法3）

5. あ と が き

一般化主成分分析法はデータ解析における基本的な課題である『関数の当てはめ』に対する方法の一つと見なせる。しかし，『関数の当てはめ』として利用するためには座標系の変換に対してなるべく影響を受けないことが好ましい。すでに，座標系の直交変換に関して不変な手法は数学的に特定されているが，その他の変換についての不変性については発表されていない。しかし，本報告のように一般化主成分分析法を少し改良することによって，不変性の問題は解決できることが示された。

一般化主成分分析法における今後の問題としては，第1主成分の意味づけ，固有値の解釈，有効性の評価などが残されている。

参 考 文 献

1) Gnanadesikan, R.: Methods for Statistical Data Analysis of Multivariate Observations, (1977) John Wiley & Sons.（丘本　正，磯貝恭史訳，統計的多変量データ解析，日科技連）.
2) Gnanadesikan, R. & Wilk, M. B.: In Multivariate Analysis (P. R. Krishnaiah, ed), (1969) Academic Press, p. 593-638.
3) Kendall, M.: Multivariate Analysis, 2nd ed. (1980) Griffin, London.
4) Mizuta, M.: J. Japan Statist. Soc., 14, (1983) p. 1-9.