北海道大学工学部研究報告
第165号（平成5年）

Bulletin of the Faculty of Engineering
Hokkaido University No.165(1993)

射影追跡におけるデータの摂動の布置に与える影響の評価について

今井　英幸　　佐藤　順一　　伊達　惇

（平成5年3月30日受理）

Evaluation of Effect on Configurations by Perturbation of The Data Points.

Hideyuki IMAI, Jun-ichi SATO and Tsutomu DA-TE

(Received March 30, 1993)

abstract

Projection pursuit (PP) is one of the multivariate methods to find the most 'interesting' low-dimensional projections of a high dimensional data set. PP defines the measure of interestingness as nonnormality of projected distribution.

In PP algorithm, nonnormality is translated into the computable expression called "projection index". So the projection index must have a large value when the projected distribution is nonnormal and small value otherwise. PP procedure searches projective plane by numerically maximizing the projection index. Some projection indexes are suggested by Friedman, Hall, and so on.

The Multivariate data may include outliers, and they give bad effect on the result of analysis. In this paper, we try to evaluate how the configuration is affected by perturbation of the data points to find such data.

1. はじめに

射影追跡は，多次元データのもつ構造の特徴を，直線や平面などの低次元空間に射影することによって探索するデータ解析の手法である。射影追跡では「射影指標」と呼ばれる射影されたデータの興味深さを示す指標を最大にする空間を探し出し，それを端末などに表示する事によりデータのもつ構造を探索することから，計算機の計算能力やグラフィック機能を活用した手法であるといえる。

本論文では平面へ射影する場合に，いくつかのデータの摂動が射影追跡で探し出される射影平面上の布置に与える影響を感度分析的手法によって評価することを試みる。

2. 射影追跡法

データの持つ特徴を捉えるために行なわれる方法のなかで最も一般的なのはデータが一変量であればヒストグラム，二変量であれば散布図を書いてみることである。しかしこれらの方法が可能である

情報工学科　情報数理工学講座

のは変量数が少ない場合に限られており，三変量の散布図を書くこともコンピュータグラフィックを用いれば不可能ではないが実際には二変量までが限界といえる。

データの変量が少なければ各変量ごとのヒストグラムや二変量づつの組合せの散布図を描いてみることもできるが，この方法は変量数が増えると困難になり，さらにデータの特徴が変量ごとのヒストグラムや散布図に表れるとは限らないという欠点がある。

変量の多いデータを解析するために多く用いられる手法に主成分分析がある。通常，いくつかの主成分に関するヒストグラムや散布図を描いて構造を調べるが，寄与率の高い主成分にデータの特徴が表れるとはいえないため，すべての主成分によるヒストグラムや散布図を描く必要があり，この方法も変量が多くなると困難であることに変わりはない。

射影追跡は多変量データの持つ構造を最も特徴的に表すような低次元空間(普通一次元または二次元)をコンピュータによって自動的に見つけ出しそれを端末に表示することにより構造を探索しようとする手法である。このことから射影追跡はコンピュータの計算能力やグラフィック機能を活用した手法であるといえる。

3. 射影指標

以下では平面(二次元空間)への射影について述べる。これらは容易に一般の低次元空間への射影に拡張できる。

n個のp次元ベクトル $x_i, i=1,\ldots,n$ をデータとすると，互いに直交する単位ベクトル a,b で張られる平面へ射影されたデータの座標は

$$(a'x_i, b'x_i), i=1,\ldots,n$$

である。a,b として$(0,\ldots,1,\ldots,0)'$, $(0,\ldots,1,\ldots,0)'$をとれば各変量ごとの散布図，第i主成分ベクトルと第j主成分ベクトルを用いれば第i主成分と第j主成分による散布図ができる。

射影追跡では射影されたデータの持つ興味深さ，面白さ(interestigness)が最も大きくなるような方向を求め表示する。そのためには興味深さ，面白さを計算可能な関数として定式化する必要がある。この関数を射影指標 (projection index)と呼ぶ。射影指標はデータが与えられた場合，方向ベクトルの関数になるので $I(a,b)$ と書くことにする。多変量解析の手法の多くは射影指標を適当に選ぶことにより射影追跡の一つであるとみなすことができる。

例えば，データの分散共分散行列をΣとして射影指標を

$$I(a,b)=a'\Sigma a+b'\Sigma b$$

とすれば，$I(a,b)$ は a,b がそれぞれΣの固有値が大きいものから二つに対応する固有ベクトルのとき最大となる。したがって主成分分析は射影追跡に含まれることがわかる。

主成分分析ではデータのちらばり(分散)に注目して，それを指標とした。このようにどのようなデータの構造に着目するかによって用いる射影指標が決定される。興味深い構造の例として全体がいくつかのクラスターから成り立っている場合や，変量間に何らかの関係がある場合などが考えられるが，それらすべてを含む形での指標の定式化は困難である。そこで射影追跡では

「正規分布が最もつまらない構造である」

と定めた。つまりデータの構造は正規分布から離れるほど興味深い，と考える。その理由として平均と分散が一定の分布のなかでエントロピーが最大のものが正規分布である，等があげられている(Huber[3])。

4. 射影指標がもつべき性質

射影指標 $I(a, b)$ は次のような性質を持つことが望ましい（岩崎・福永[5]）。

1． $I(a, b)$ は射影されたデータが正規分布のとき最小値をとり，正規分布から離れるほど大きな値をとる。

2． データのアファイン変換によって指標の値は変わらない。

3． 計算量が少なく，a, bに関して連続偏微分可能である。

1は射影指標が非正規性の尺度であることに対応している。2は射影されたデータの構造は変量の尺度を変えたり，散布図を回転することによって変わることはないという考えに基づいている。3はコンピュータで計算するために必要な性質である。性質2によりデータ $x_i, i=1, \ldots, n$ は平均0，分散共分散行列I_pに標準化されているものとしてよい。このような性質を満たす指標として今までに以下のようなものが提案されている。これらの指標は 岩崎・福永[5]で導入された多項式指標の例である。

モーメント指標

$$I_1(a, b)=\sum_{k=0}^{m}\sum_{j=0}^{m-k}\{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{j!k!}}H_j(a'x_i)H_k(b'x_i)\}^2$$

ただし$H_j(x)$は j 次のHermit多項式である。この指標はモーメントの線形和になる。

Friedman による指標

$$I_2(a, b)=\sum_{k=0}^{m}\sum_{j=0}^{m-k}\{\frac{1}{n}\sum_{i=1}^{n}\frac{2}{\sqrt{(2j+1)(2k+1)}}P_j(2\Phi(a'x_i)-1)P_k(2\Phi(b'x_i)-1)\}^2$$

ただし $P_j(x)$ は j 次のLegendre多項式, $\Phi(x)$ は標準正規分布関数である。この指標は正規性の検定の一つであるNeyman の smooth test に用いられる統計量の定数倍になっている（柴田[9]）。

Hall による指標

$$I_3(a, b)=\sum_{k=0}^{m}\sum_{j=0}^{m-k}\{\frac{1}{n}\sum_{i=1}^{n}2\sqrt{\frac{\pi}{j!k!}}H_k(\sqrt{2}a'x_i)\phi(a'x_i)H_j(\sqrt{2}b'x_i)\phi(b'x_i)\}^2$$

ただし$\phi(x)$ は標準正規密度関数である。

上の三つの指標の中で用いられているmはデータの密度関数への近似の程度を決める定数で大きいほど近似が正確になるが，計算量は多くなる。岩崎・福永[5]は$m=4$で十分によい近似が得られることを数値例により示した。実際の計算では適当な初期値を与えて非線形最適化の手法を用いて最適な方向ベクトル a, b を求める。

5. 感度分析

感度分析はいくつかのデータを微小に動かしたときの結果の変動をみる手法である。得られたデータの任意の部分集合Sに対して

$$\begin{cases}(1-\varepsilon)x_i, & x_i\in S\\ x_i, & x_i\notin S\end{cases}$$

という摂動を与えたとき射影方向 a, b がどのように影響を受けるのかを調べる。以下，Sは固定しておく。摂動εを与えたデータに対する射影指標は方向 a, b と重みεの関数と見なせるので，これを $I(a, b, \varepsilon)$ と書くと求める射影方向は

$$\left(\frac{\partial I(a,b,\varepsilon)}{\partial a'},\ \frac{\partial I(a,b,\varepsilon)}{\partial b'}\right)=0$$

の解として得られる。摂動を与えないデータ（もとのデータ）に対する射影方向は，$\varepsilon=0$ としたときの解であり，これを a_0, b_0 とすると摂動による方向ベクトルの変動はεのオーダで次のように評価できる（今井・佐藤[4]）。

定理 1

$$L_1(a,b,\varepsilon)=\left(\frac{\partial I(a,b,\varepsilon)}{\partial a_1},\ldots,\frac{\partial I(a,b,\varepsilon)}{\partial a_p}\ \frac{\partial I(a,b,\varepsilon)}{\partial b_1},\ldots,\frac{\partial I(a,b,\varepsilon)}{\partial b_p}\right)$$

$$L_2(a,b,\varepsilon)=\begin{pmatrix}\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_1\partial a_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_1\partial a_p}&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_1\partial b_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_1\partial b_p}\\ \vdots&&\vdots&\vdots&&\vdots\\ \frac{\partial^2 I(a,b,\varepsilon)}{\partial a_p\partial a_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_p\partial a_p}&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_p\partial b_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial a_p\partial b_p}\\ \frac{\partial^2 I(a,b,\varepsilon)}{\partial b_1\partial a_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_1\partial a_p}&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_1\partial b_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_1\partial b_p}\\ \vdots&&\vdots&\vdots&&\vdots\\ \frac{\partial^2 I(a,b,\varepsilon)}{\partial b_p\partial a_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_p\partial a_p}&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_p\partial b_1}&\cdots&\frac{\partial^2 I(a,b,\varepsilon)}{\partial b_p\partial b_p}\end{pmatrix}$$

とする。$L_1(a,b,\varepsilon)$ が点 $(a_0, b, 0)$ の近傍で連続微分可能で $det(L_2(a_0, b_0, 0))\neq 0$ であれば点 $(a_0, b_0, 0)$ の近傍で (a, b) はεの関数として表すことができる。これを $(a(\varepsilon), b(\varepsilon))$ とすると

$$\begin{pmatrix}a(\varepsilon)\\ b(\varepsilon)\end{pmatrix}=\begin{pmatrix}a_0\\ b_0\end{pmatrix}-L_2^{-1}(a_0,b_0,0)\frac{\partial L_1}{\partial\varepsilon}(a_0,b_0,0)\varepsilon+o(\varepsilon)$$

ここで，$o(\varepsilon)$ はすべての要素が $o(\varepsilon)$ である$2p$次元ベクトルを表すものとする。

一次元への射影の場合，方向ベクトルの変化の大きさによって摂動の影響を評価できるが，平面への射影では方向ベクトルの変化ではなく，布置の変化の大きさを評価する必要がある。そのために，Sibson[10]による次の定理を用いる。

定理 2

$x_i, y_i, i=1,\ldots,n$ をそれぞれ平均 0 の p 次元ベクトルとする。$X=[x_1\cdots x_n]$, $Y=[y_1\cdots y_n]$, Q を直交行列とするとき

$$G(X,Y)=\min_{Q\in O_p}\operatorname{tr}(X-QY)'(X-QY)=\operatorname{tr}XX'+\operatorname{tr}YY'-2\operatorname{tr}(YX'XY')^{\frac{1}{2}}.$$

この定理は適当な直交行列Qを用いてYをXにもっとも近くなるように変換してからデータ点間の距離の二乗和を求めたものでありXとYの回転による差を考慮したものであるといえる。

Xを摂動を与えないデータの布置Yを摂動を与えたデータの布置とすると定理1から

$$Y=X+\varepsilon Z+o(\varepsilon),\quad Z=-L_2^{-1}(a_0,b_0,0)\frac{\partial L_1}{\partial \varepsilon}(a_0,b_0,0)X$$

であるから，定理2を適用することによりつぎの定理を得る。

定理3

$$G(X,X+\varepsilon Z+o(\varepsilon))=\varepsilon^2(\mathrm{tr}ZZ'-\sum_{i=1}^{2}\frac{b_i-(\frac{a_{ii}}{2\lambda_i})^2}{\lambda_i})+o(\varepsilon^2).$$

ここで

λ_iはXX'の第i固有値，$\Lambda=\mathrm{diag}\ [\lambda_1,\lambda_2]$，$XX'=P\Lambda P'$，$PP'=I_p$，$D=PZX'P'$，

$D\Lambda+\Lambda D=(a_{ij})$，　$DD'=(d_{ij})$，$b_i=d_{ii}+\sum_{j\neq i}\frac{(a_{ij})^2}{\lambda_i^2-\lambda_j^2}$．

証明は Appendix.

定理3によりデータに摂動を与えた場合の布置に対する影響を評価できる。

6. 数値実験

ここでは，射影指標として 4. で述べたFriedmanの指標を用いて，次のような人工データによる数値実験を行なった。平均が$\mu_1=(0,0,0,0)'$，$\mu_2=(4,0,0,0)'$，$\mu_3=(2,3,0,0)'$である3個の正規母集団からそれぞれ10個，$\mu_4=(0,-3,3,0)'$である正規母集団から1個，正規乱数を発生させて31個の標本を作った。ただしそれぞれの母集団の分散共分散行列はすべて I_4 である。図1はこのデータに対して射影追跡を行なった結果を図示したものである。1から10までは平均μ_1，11から20までは平均μ_2，21から30までは平均μ_3，31は平均μ_4の母集団からのデータ点である。求まった平面の方向ベクトルは $a=(0.507,0.811,-0.308,0.028)'$，$b=(-0.864,0.501,-0.106,0.012)'$となった。この図からデータが3個のクラスタと1個の外れ値からできていることがわかる。

このデータに対してある1個のデータに微小な変化を与えた場合について感度分析を行ないそれをデータの布置の変化で評価してグラフで表したものが図2である。

この結果から次のようなことが言える。

- 31番目のデータが分析結果に大きな影響を及ぼしている。
- 布置の変化が小さいデータはクラスタの中心に近いものが多い。
- クラスタから外れて見えるデータでも結果に与える影響が少ないものもある。

つまりこのような個々のデータに対する評価により，一見外れ値には見えないが結果に大きな影響を与えているデータ(2,3など)，あるいは外れているように見えるが，結果にさほど影響を与えていないデータ(5,10など)も見つけることができる。

7. まとめ

多変量データを解析する際，疑似外れ値や目では判別できない外れ値が存在することがあり，それが結果に悪影響を及ぼすことがある。今回の研究では，射影追跡を用いた場合のそのような外れ値の影響を考察するために，感度分析を用いて個々のデータの摂動がデータの布置にどれだけの変化を与えるかを調べた。また数値実験により実際に外れ値を探索する際に有効であることを確認した。

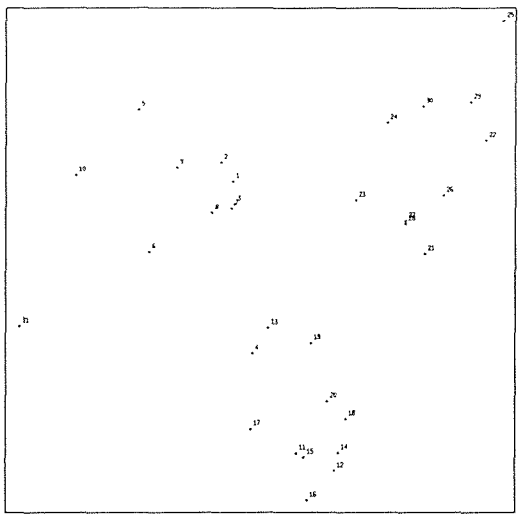

図1　最適射影平面上の散布図

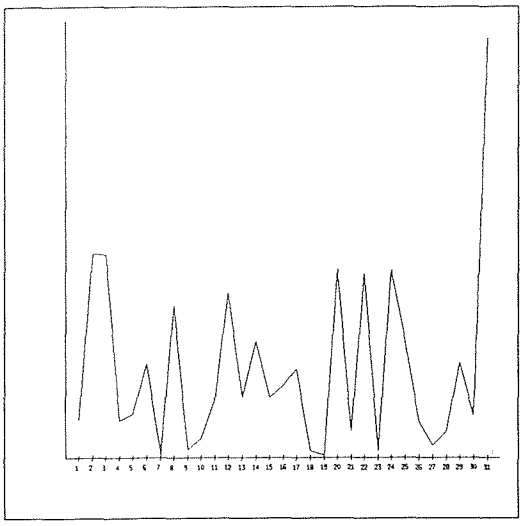

図2　感度分析の結果

Appendix

XX'を$p\times p$行列として証明する．定理2は$p=2$の場合である。

$$\begin{aligned} G(X, X+\varepsilon Z+o(\varepsilon)) &= \mathrm{tr}XX'+\mathrm{tr}(X+\varepsilon Z+o(\varepsilon))(X+\varepsilon Z+o(\varepsilon))' \\ &\quad -2\mathrm{tr}\{(X+\varepsilon Z+o(\varepsilon))X'X(X+\varepsilon Z+o(\varepsilon))'\}^{\frac{1}{2}} \\ &= 2\mathrm{tr}XX'+2\varepsilon\mathrm{tr}XZ'+\varepsilon^2\mathrm{tr}ZZ' \\ &\quad -2\mathrm{tr}\{XX'XX'+\varepsilon(ZX'XX'+XX'XZ')+\varepsilon^2ZX'XZ'+o(\varepsilon^2)\}^{\frac{1}{2}} \end{aligned}$$

ここで，λ_iをXX′の第i固有値，$\Lambda=\mathrm{diag}[\lambda_1,\ldots,\lambda_p]$, $XX'=Q'\Lambda Q$, $QQ'=I$, $D=QZX'Q'$とすると，

$$\begin{aligned} &XX'XX'+\varepsilon(ZX'XX'+XX'XZ')+\varepsilon^2ZX'XZ'+o(\varepsilon^2) \\ &\quad =Q'\{\Lambda^2+\varepsilon(D\Lambda+\Lambda D)+\varepsilon^2DD'+o(\varepsilon^2)\}Q \end{aligned}$$

であるから，$\Lambda^2+\varepsilon(D\Lambda+\Lambda D)+\varepsilon^2DD'+o(\varepsilon^2)$ の固有値をl_i^2とすると，Shiotani, Hayakawa and Fujikoshi[11]の定理4.6.1より

$$l_i^2=\lambda_i^2+\varepsilon a_{ii}+\varepsilon^2\left(b_{ii}+\sum_{j\neq i}\frac{(a_{ij})^2}{\lambda_j-\lambda_i}\right)+o(\varepsilon^2).$$

ここで$(a_{ij})=D\Lambda+\Lambda D$, $(b_{ij})=DD'$.

したがって$c_i=b_{ii}+\sum_{j\neq i}\frac{a_{ij}^2}{\lambda_j-\lambda_i}$とおけば，

$$l_i=\lambda_i+\varepsilon\frac{a_{ii}}{2\lambda_i}+\varepsilon^2\frac{c_i-\left(\frac{a_{ii}}{2\lambda_i}\right)^2}{2\lambda_i}+o(\varepsilon^2).$$

$\mathrm{tr}XX'=\Sigma_i\lambda_i$, $\mathrm{tr}XZ'=\mathrm{tr}D=\Sigma_i\frac{a_{ii}}{2\lambda_i}$より，

$$G(X, X+\varepsilon Z+o(\varepsilon))=\varepsilon^2\left(\mathrm{tr}ZZ'-\sum_{i=1}^{p}\frac{c_i-\left(\frac{a_{ii}}{2\lambda_i}\right)^2}{\lambda_i}\right)+o(\varepsilon^2).$$

参考文献

1) Friedman, J. H. (1987). Exploratory projection pursuit. J. A. S. A., 82, 249-266.

2) Hall, P. (1989). On Polynomial-based projection indices for exploratory projection pursuit. Ann. Statist., 17, 589-605.

3) Huber, P. J. (1985). Projection pursuit (with discussion). Ann. Statist., 13, 435-535.

4) 今井英幸・佐藤義治(1992). 射影追跡法における感度分析. 計算機統計学, 5(2), 101-106.

5) 岩崎学・福永真美(1989). 多項式指標による射影追跡. 応用統計学, 18, 103-128.

6) 岩崎学(1991). 射影追跡：その考え方と実際. 計算機統計学, 4(2), 41-56.

7) Jones, M. C. and Sibson, R. (1987). What is projection pursuit ? (with discussion). J. R. S. S., A150, 1-36.

8) 佐藤順一・今井英幸・伊達惇(1992). 射影追跡法におけるデータの摂動の布置に対する影響について. 電気関係学会北海道支部連合大会講演論文集

9) 柴田義貞(1981). 正規分布：特性と応用. 東京大学出版会.

10) Sibson, R. (1978). Studies in robustness of multidimensional scaling:Procrustes statistics. J. R. S. S., B40, 234-238.

11) Siotani, M., Hayakawa, T. and Fujikoshi, Y. (1985). Modern Multivariate Statistical Analysis:A Graduate Course and Handbook. American Sciences Press.