5D-6

データに対して最小２乗距離を有する２次曲線について

水田正弘

(北海道大学　工学部)

１．目的

多次元データに対して、曲線や曲面を『当てはめる』ことによってデータの持つ構造を解析・把握することが可能になることがある。Gnanadesikan[1]は、データの非線形特異性を調べる方法として一般化主成分分析法を提案した。これはデータの布置に適した非線形な座標系を決定する方法であり、その座標系のいくつかはデータに対する非線形関数の当てはめと見なすことができる。一般化主成分分析法によってデータに曲線等を当てはめることが可能になるが、この曲線等は必ずしも垂直な距離の２乗和を最小にしていない。そこで、今回の報告では、２次元データに対して垂直な距離の２乗和が最小となる２次曲線を求めるアルゴリズムを提案する。

２．問題とその定式化

はじめに、記号の定義と問題の定式化をする。

n個の２次元データを$(\alpha_i, \beta_i)$$(i=1,2,\cdots,n)$で表し、２次曲線を $ax+by+cx^2+dxy+ey^2+g=0$ で表す。 データ点(α_i, β_i)と２次曲線 $ax+by+cx^2+dxy+ey^2+g=0$ との２乗距離r_iを

$$r_i = \min\{(x-\alpha_i)^2+(y-\beta_i)^2 \; ; \; ax+by+cx^2+dxy+ey^2+g=0\}$$

で表す。 r_iを実現する２次曲線上の点を(x_i, y_i)とする。すなわち、

$$r_i = (x_i-\alpha_i)^2+(y_i-\beta_i)^2$$
$$ax_i+by_i+cx_i^2+dx_iy_i+ey_i^2+g=0$$

となる。 だたし、(x_i, y_i)は一般には一意に定まらないこともあるが、そのときには任意の一点を選ぶことにする。点(α_i, β_i)から点(x_i, y_i)への線分は、２次曲線の点(x_i, y_i)における法線となる。 すべてのデータ点についてのr_iの総和Rを$R=\Sigma r_i$によって定義する。

本報告で解く問題は、Rを最小にする２次曲線 $ax+by+cx^2+dxy+ey^2+g=0$ を求めることである。

３．最小２乗距離を有する２次曲線

２乗距離の総和が最小となる２次曲線を求める方法を２段階に分けて説明する。はじめに、データ点と２次曲線との２乗距離r_i の計算法を述べる。これにより、２乗距離の総和Rを、a,b,c,d,e,g を変数とする６変数関数と見なせる。そこで、次に多変数関数Rの偏微分を求めRの最小値を求める方法を述べる。

３．１　データ点と２次曲線との２乗距離の計算法

はじめに、 a,b,c,d,e,g が与えられたときのr_iの計算方法を述べる。 あるデータ点(α_i, β_i)に対して２次曲線上の点を(x,y)とおき、変数として扱う。 r_iを求めるには、$ax+by+cx^2+dxy+ey^2+g=0$ を制約式として、$(x-\alpha_i)^2+(y-\beta_i)^2$ の最小値問題をラグランジュの未定乗数法で解けばよい。

$$f_i(x,y,\lambda) = (x-\alpha_i)^2+(y-\beta_i)^2+\lambda(ax+by+cx^2+dxy+ey^2+g)$$

とし、f_iをx,y,λで各々微分したものを０とおき、

$$2(x-\alpha_i)+\lambda(a+2cx+dy)=0 \tag{1}$$
$$2(y-\beta_i)+\lambda(b+2ey+dx)=0 \tag{2}$$
$$ax+by+cx^2+dxy+ey^2+g=0 \tag{3}$$

を得る。(1),(2)をｘ，ｙの連立方程式と見なして解くと、

$$x=(-2ae\lambda^2-2a\lambda+bd\lambda^2-2d\beta\lambda+4e\lambda\alpha+4\alpha)/(4ce\lambda^2+4c\lambda-d^2\lambda^2+4e\lambda+4)$$
$$y=(ad\lambda^2-2bc\lambda^2-2b\lambda+4c\beta\lambda-2d\lambda\alpha+4\beta)/(4ce\lambda^2+4c\lambda-d^2\lambda^2+4e\lambda+4)$$

となる。すなわち、ｘとｙはλの関数となる。これらを(3)に代入することによりλの４次方程式

$$\begin{aligned}&(-4a^2ce^2+a^2d^2e+4abcde-abd^3-4b^2c^2e+b^2cd^2+16c^2ge^2\\&-8cd^2ge+d^4g)\lambda^4 + 8(-a^2ce-a^2e^2+abcd+abde-b^2c^2\\&-b^2ce+4c^2ge-cd^2g+4cge^2-d^2ge)\lambda^3 + 4(-a^2c-4a^2e\\&+3abd-2acd\beta_i+ad^2\alpha_i-2ade\beta_i+4ae^2\alpha_i-4b^2c-b^2e\\&+4bc^2\beta_i-2bcd\alpha_i+bd^2\beta_i-2bde\alpha_i+4c^2g+4c^2e\beta_i^2\\&-cd^2\beta_i^2-4cde\beta_i\alpha_i+16cge+4ce^2\alpha_i^2+d^3\beta_i\alpha_i-2d^2g\\&-d^2e\alpha_i^2+4ge^2)\lambda^2 + 8(-a^2-2ad\beta_i+4ae\alpha_i-b^2+4bc\beta_i\\&-2bd\alpha_i+4cg+4ce\beta_i^2+4ce\alpha_i^2-d^2\beta_i^2-d^2\alpha_i^2+4ge)\lambda\\&+16(a\alpha_i+b\beta_i+c\alpha_i^2+d\beta_i\alpha_i+g+e\beta_i^2)=0\end{aligned} \tag{4}$$

が得られ、フェラリ法またはニュートン・ラプソン法によって解くことができる。 たかだか、 ４個の得られた実数の解λについてｘとｙの組を求め、 そのなかで$(x-\alpha_i)^2+(y-\beta_i)^2$ の最小値が求めるr_iである。また、最小値をとるｘとｙがx_i，y_iである。

Curve of Second Class with Least-Squares Distance on 2 Dimensional Data

Masahiro MIZUTA

HOKKAIDO university.

3.2 2乗距離の総和の最小化

前節より、距離の2乗和R は a,b,c,d,e,g の多変数関数と見なすことができる。Rの最小値を数値的に解くために、Rの a,b,c,d,e,gに関する偏微分を求める。ここでは、aに関する偏微分の計算法のみを示すが、他の変数についても全く同様に計算できる。x_i, y_iも a,b,c,d,e,g の関数なので、

$$R=\sum_{i=1}^{n}\{(x_i-\alpha_i)^2+(y_i-\beta_i)^2\} \qquad (5)$$

より、

$$\frac{\partial R}{\partial a}=2\sum_{i=1}^{n}\{(x_i-\alpha_i)x_i'+(y_i-\beta_i)y_i'\} \qquad (6)$$

となる。ただし、$x_i'=\partial x_i/\partial a$, $y_i'=\partial y_i/\partial a$ とする。ここで、

$$ax_i+by_i+cx_i^2+dx_iy_i+ey_i^2+g=0$$

の両辺をaで微分すると、

$$(a+2cx_i+dy_i)x_i'+(b+dx_i+2ey_i)y_i'+x_i=0 \qquad (7)$$

が得られる。また、点(x_i, y_i)から点(α_i, β_i)への線分は、2次曲線の点(x_i, y_i)における法線なので、

$$(y_i-\beta_i)(a+2cx_i+dy_i)=(x_i-\alpha_i)(b+2ey_i+dx_i) \qquad (8)$$

が成り立つ。両辺をaで微分して整理すると、（以下、x_i, y_i, α_i, β_iの添字は省略して、単にx, y, α, βと書く）

$$(b-2cy+2c\beta+2dx-d\alpha+2ey)x' +(-a-2cx-2dy+d\beta+2ex-2e\alpha)y'-(y-\beta)=0 \qquad (9)$$

となる。そこで、(7),(9)をx_i'とy_i'の一次連立方程式として解くと、

$x_i'=(-ax+by-b\beta-2cx^2-dxy+2x^2e-2x\alpha e+2y^2e-2y\beta e)/q$

$y_i'=(-ay+a\beta-bx-2dx^2+dx\alpha-dy^2+dy\beta-2xye)/q$

（ただし、$q=(a^2+4acx+3ady-ad\beta-2axe+2a\alpha e+b^2-2bcy+2bc\beta+3bdx-bd\alpha+4bye+4c^2x^2+4cdxy-4cx^2e+4cx\alpha e-4cy^2e+4cy\beta e+2d^2x^2-d^2x\alpha+2d^2y^2-d^2y\beta+4dxye+4y^2e^2)$）

が得られる。これらを(6)の各項に代入することにより、

$$\{(x_i-\alpha_i)\partial x_i/\partial a+(y_i-\beta_i)\partial y_i/\partial a\}$$
$$=(-((ax-by+b\beta+2cx^2+dxy-2x^2e+2x\alpha e-2y^2e+2y\beta e)(x-\alpha)+(ay-a\beta+bx+2dx^2-dx\alpha+dy^2-dy\beta+2xye)(y-\beta)))/q$$

となり、∂R/∂aが求まる。一般化主成分分析法の解を初期値として、これらによる偏微分を利用して(ダビドン法などにより)Rの最小値を求めることができる。

4. 数値例

次に数値例を示す。サンプル数47の2次元データ（平均 0；各成分の分散1）に、一般化主成分分析および本報告の方法を適用する。ここで一般化主成分分析で利用する関数系は x, y, x^2, xy, y^2 とする。

一般化主成分分析法ではR＝1.831077となった。本報告の方法では、一般化主成分分析の解を初期値とした場合も乱数を初期値とした場合も共に同じ2次曲線が得られR＝1.622973となった。図1に本報告の方法による2次曲線を示す。

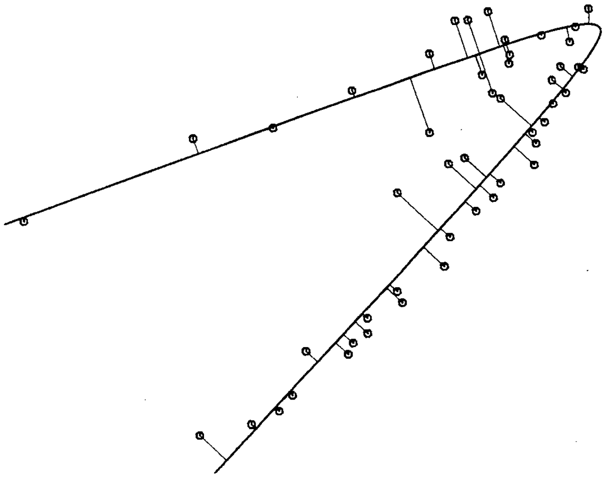

図1．本報告の方法による2次曲線

5. 結論

以上により、2次元データに対して垂直な距離の2乗和が最小となる2次曲線を求めるアルゴリズムを提案した。このアルゴリズムは多次元データに関する2次超曲面に対しても拡張可能である。例えば、3次元データでは(4)が6次方程式となり、(8)が2本の式になる。

本研究における式の展開・整理のために北海道大学大型計算機センターの数式処理システム Reduce を利用している。さらにFortranプログラムを作成するためにも Reduce の出力を使った。

本研究に関して、統計数理研究所の共同研究(63-共研-8)の協力を得ました。

参考文献

[1]Gnanadesikan,R.(1977). Methods for Statistical Data Analysis of Multivariate Observations, John Wiley & Sons.（和訳　丘本　正、磯貝恭史　(1979). 統計的多変量データ解析、日科技連）

[2]Mizuta,M.(1983). Generalized Principal Components Analysis Invariant under Rotations of a Coordinate System, J. Japan Statist. Soc.,14, pp. 1-9.