統計数理研究所彙報
第29巻 第2号 (1982)

説明変数の配置構造の探索

統計数理研究所 川 合 伸 幸

(1981年6月 受付)

1. はじめに

データを眺め，それが含んでいる‘知識’を把えることは，データを分析する第一歩として非常に重要である．最初にデータを眺めることを怠り，無造作に手法を適用しその結果を信じ込むと思わぬ落とし穴に落ち込む危険が生ずる．いかなる分析手法を用いようと，データが含んでいる以上のものは引き出すことはできないからである．本稿では，回帰分析の一つの誤った適用を述べ，併せてその診断法について述べる．

2. 誤った適用

データに標準的な回帰モデル

$$\underset{n\times 1}{\boldsymbol{y}} = \beta_0 \underset{n\times 1}{\mathbf{1}_n} + \underset{n\times p}{X}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\varepsilon}$$

をあてはめ分析することを考える．ここで，$\mathbf{1}_n=(1,\cdots,1)'$, $\mathrm{E}[\varepsilon]=\mathbf{0}$, $\mathrm{E}[\varepsilon\varepsilon']=\sigma^2 I_n$ とする．

いま，説明のため $p=2$ とする．回帰のメカニズムの部分だけを見ると，$\beta_0+\beta_1X_1+\beta_2X_2$ となる．このとき，パラメータ $\beta_1(\beta_2)$ はそれぞれ，$X_2(X_1)$ を固定しておいて，$X_1(X_2)$ を1単位増やしたときの y の変化量と解釈される．したがって，β_1, β_2 の推定量 $\hat{\beta}_1, \hat{\beta}_2$ の大きさをもって変数 X_1, X_2 の y に及ぼす影響度と解釈することが考えられる．

ここで，この解釈を行なう際には，仮定した回帰モデルが有効である，すなわち，現象のメカニズムが，実験の範囲内では，$X_2(X_1)$ を固定しておいて $X_1(X_2)$ を1単位増やせば，$X_2(X_1)$ のレベルに関係なく y に $\beta_1(\beta_2)$ の変化を及ぼす，ということがわかっていなければならぬ．もし，その確信が得られないなら，モデル自体に対する知識も得られるように実験を考えねばならぬ．

しかし，説明変数 (X_1, X_2) の配置が図1のように直線上に乗っているとき，このデータからわかることは，この直線上で y がどうなるかということである．

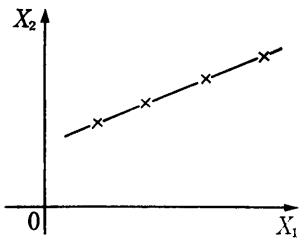

図 1

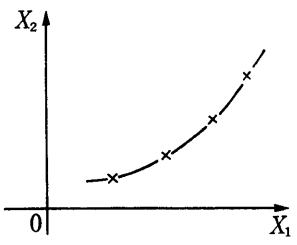

図 2

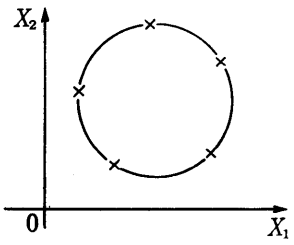

図 3

図2のように，直線でないがほぼ直線に近い曲線上に乗っている場合も同様である．ところで図3のように，配置が円周上にのみ乗っている場合はどうであろうか．この場合，円の内部

については実際には実験されておらず，この内部でもしモデルが平面でなかったとしても，事後的に残差分析によってチェックすることもできない．したがって，このときデータからわかることは，この円周上で y がどうなるかということで，知りうることは図1，図2の実験とほぼ同じであり，良い配置であるとはいえない．このとき，円周上の y の値だけから計算された $\hat{\beta}_1, \hat{\beta}_2$ をもって，$X_2(X_1)$ を固定しておいて $X_1(X_2)$ を1単位増やしたときの y の増分という解釈を円の内部全体に広げることは危険である．実験の範囲を広げようと円を大きくすることは，ますます，この種の危険を増大させることになる．

ところで，変数の値を制御できるような実験であれば，意識的にこれらの配置を避けることができる．しかし，データが受動的な観測によって得られる場合にも回帰モデルを適用して変数の y に及ぼす影響を調べたいときがある．そのとき，今のような配置が生じてないという保障はない．もし，そのような配置であれば，回帰係数の解釈は限定されることになる．次節では，これらの配置型検出のための診断法について述べる．

3. 診　断　法

今の例のように，2次元の場合ならば，説明変数の観測値 $\boldsymbol{X}_i=(X_{i1}, X_{i2})'$ $i=1, \cdots, n$ をプロットすることにより，直接，配置を見ることができる．しかし，多次元の場合には，プロットによる方法は不可能になり，量的な方法が必要になる．図1のように配置が平面上に乗っている場合には，

$$\boldsymbol{x}_i=(X_{i1}-\bar{X}_{\cdot 1}, \cdots, X_{ip}-\bar{X}_{\cdot p})', \quad i=1, \cdots, n$$

について，主成分分析を用いることによって，その形態を検出できる．ただし，$\bar{X}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

図2のように配置が平面に近い場合（多重共線関係にある場合）も同様 $\boldsymbol{x}_i$, $i=1, \cdots, n$ の主成分分析を行なうことによって，ある程度検出できる．（[1] の7章参照．主成分分析については，例えば奥野他 [2] を参照.）しかし，図3のような場合は，主成分分析だけからはわからない．この図3の例のような場合を量的に把える方法を以下に与える．

まず，原点を，説明変数ベクトルの重心，

$$\bar{\boldsymbol{X}}=(\bar{X}_{\cdot 1}, \cdots, \bar{X}_{\cdot p})', \quad \bar{X}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n} X_{ij}$$

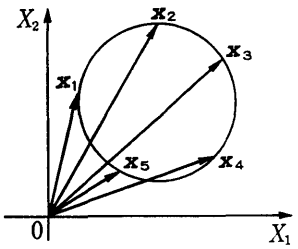

図4　図3の説明変数ベクトル

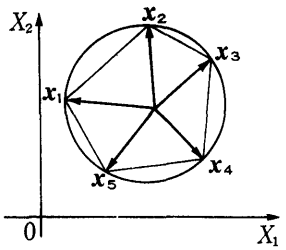

図5　重心から測られた説明変数ベクトルとそのベクトルのつくる凸多面体

に移す．$\bar{\boldsymbol{X}}$ から測られた $\boldsymbol{X}_i$ を $\boldsymbol{x}_i$ とする（図4，図5参照）．

そして，$\boldsymbol{x}_i$, $i=1, \cdots, n$ のつくる凸多面体に注目する（図5参照）．すると，図3の説明変数の配置は，図5で $\boldsymbol{x}_i$ がすべて凸多面体上の点であって，その内部に観測データがないという

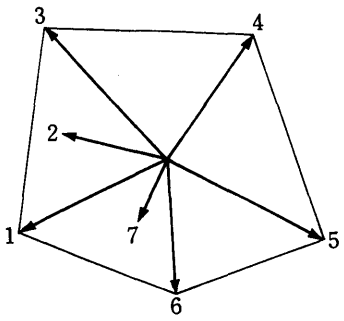

図 6　ベクトル 1, 3, 4, 5, 6 は柱ベクトル
ベクトル 2, 7 は凸多面体の内部にあるベクトル

状態として把えられる．

一般に，p 次元空間において，図 3 の例のような配置は，重心から測られた変数配置が，その配置のつくる凸多面体上に集中していてその内部に変数配置を含んでいない状態として把えられる．したがって．変数配置 $\{\boldsymbol{X}_i, i=1,\cdots,n\}$ が与えられたとき，重心から測られた変数配置 $\{\boldsymbol{x}_i, i=1,\cdots,n\}$ のつくる凸多面体を考え，$\{\boldsymbol{x}_i, i=1,\cdots,n\}$ をこの凸多面体を張るのに関与しているものと，凸多面体の内部にあって，凸多面体を張るのに関与していないものとに分ける（図 6 参照）．凸多面体を張るのに関与しているベクトルを，以下，柱ベクトルと呼ぶことにする．すると，$\{\boldsymbol{x}_i, i=1,\cdots,n\}$ のすべてが柱ベクトルであれば，配置が凸な物体の表面に集中しており内部にはつまっていないと判断でき，図 3 の配置形態を把えることができる．

具体的に柱ベクトルを求める手順は次節において与えられる．

4. 柱ベクトルを求める手順

手順中，補遺で述べられる凸多面体についての性質 1, 2 が，探索の依りどころとなる．アルゴリズムの基本方針としては，性質 1 に依りベクトル $\{\boldsymbol{x}_i, i=1,\cdots,n\}$ から組み合わせ的に $p+1$ 個のベクトル $(\boldsymbol{x}(0), \boldsymbol{x}(1), \cdots, \boldsymbol{x}(p))$ をとり出し，その中で他の p 個のベクトルによって

$$\boldsymbol{x}(i_0) = a_1\boldsymbol{x}(i_1) + \cdots + a_p\boldsymbol{x}(i_p) \tag{4.1}$$
$$(i_0, i_1, \cdots, i_p) \subset (1, 2, \cdots, p+1)$$
$$a_i \geq 0, \quad \sum_{i=1}^{p} a_i < 1$$

の形に表わされるベクトルを除いていく．その際，性質 2 に依り，まず明らかな柱ベクトルになるものをとり出し，それを基点として除去作業を行ない，ベクトル数 n を減少させておいてから，組み合わせ演算を開始する．

手順は以下のように与えられる．

[1st STEP]

$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ $i=1,\cdots,n$ において，第 k 成分中の最大・最小値をもつ MAX, MIN ベクトル $\boldsymbol{x}_{k\,\max}$, $\boldsymbol{x}_{k\,\min}$, $k=1,\cdots,p$ を探し出し，これらのベクトルを $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n_1)$ とする．$\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n_1)$ は性質 2 により，柱ベクトルである．但し，すべての i について $x_{ik}=0$ となる場合は除いて考える．この場合は第 k 成分を除いた他の成分について 1 つの次元のおちた関係が調べられる．

[2nd STEP]

[1st STEP] で見い出された MIN, MAX ベクトルのうちの p 個，$\boldsymbol{x}(i_1), \cdots, \boldsymbol{x}(i_p)$ によって，

$$\boldsymbol{x}_k = a_1\boldsymbol{x}(i_1) + \cdots + a_p\boldsymbol{x}(i_p), \quad a_i \geq 0, \quad \sum_{i=1}^{p} a_i < 1 \tag{4.2}$$
$$(i_1, \cdots, i_p) \subset (1, \cdots, n_1)$$

と表わされるベクトルを除いていく．

性質1によりこのようなベクトルは柱ベクトルではない.

[2nd STEP] で残ったベクトルを $\boldsymbol{x}(n_1+1), \cdots, \boldsymbol{x}(n_1+n_2)$ とする. これらは柱ベクトルの候補である.

$n_1<p$ のときは, 2nd STEP を省略して [1st STEP] での MIN, MAX ベクトル以外のすべてのベクトルを $\boldsymbol{x}(n_1+1), \cdots, \boldsymbol{x}(n_1+n_2)$ に格納して [3rd STEP] へとぶ.

[3rd STEP]

[1st STEP], [2nd STEP] で得られたベクトル $\boldsymbol{x}(1)\cdots, \boldsymbol{x}(n_1), \boldsymbol{x}(n_1+1), \cdots, \boldsymbol{x}(n_1+n_2)$ の中から任意に $(p+1)$ 個を取り出す. それを $\{\boldsymbol{x}(i_0), \boldsymbol{x}(i_1), \cdots, \boldsymbol{x}(i_p)\}$ とする. この組の中で, 他の p 個のベクトルによって, (4.1) の形に表わされるものがあれば除いていく.

$\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n_1), \boldsymbol{x}(n_1+1), \cdots, \boldsymbol{x}(n_1+n_2)$ についてのすべての $(p+1)$ 個の組み合わせについて [3rd STEP] の探索を行なった後, 残ったものが柱ベクトルである.

なお, この組み合わせ中, 2nd STEP でチェックした組については 3rd STEP での探索はとばされる.

係数 $a_1, \cdots, a_p$ を求めるには, 連立方程式

$$\boldsymbol{x}(i_0) = a_1\boldsymbol{x}(i_1)+\cdots+a_p\boldsymbol{x}(i_p) \tag{4.3}$$

を解けばよい.

$\boldsymbol{x}(i_1), \cdots, \boldsymbol{x}(i_p)$ が一次従属でなければ係数は一意に決まる. よって, (4.3) より例えば, (i_1) vs. (i_0) $(i_2)\cdots(i_p)$ の関係は, $a_1\neq 0$ ならば,

$$\boldsymbol{x}(i_1)=\frac{1}{-a_1}\boldsymbol{x}(i_0)+\frac{a_2}{-a_1}\boldsymbol{x}(i_2)+\cdots+\frac{a_p}{-a_1}\boldsymbol{x}(i_p) \tag{4.4}$$

として求まる. そこで, $a_1<0$ で残りの $a_i>0$ となっていて,

$$\frac{1}{-a_1}+\frac{a_2}{-a_1}+\cdots+\frac{a_p}{-a_1}<1$$

ならば, $\boldsymbol{x}(i_1)$ は $\boldsymbol{x}(i_0), \boldsymbol{x}(i_2), \cdots, \boldsymbol{x}(i_p)$ によってはさまれており柱ベクトルではないことがわかる. また, この例で $a_1=0$ であったとすると, これは $\boldsymbol{x}(i_0), \boldsymbol{x}(i_2), \cdots, \boldsymbol{x}(i_p)$ が一次従属であることを意味する. この場合, $\boldsymbol{x}(i_1)$ を除いた残りのベクトルについて1つ次元の落ちた超平面での (4.1) の関係が調べられる.

5. おわりに

本稿では, 回帰分析の際注意しなければならぬ, 多次元データの一つの配置パターンと, それを把えるための方法を与えた. もちろんこの方法で完全に把えることのできない配置型はいくつもある. しかし, モデルの検証という観点から見るならば, 重要な配置は完全には凸多面体の表面に乗っていないが, ほとんど表面に近いところに集中しているという場合であって, この型の配置を把えることが, 次の目標とされるべきである.

謝辞

研究の過程で, 岸野洋久氏には, 多くの貴重な意見をいただいた. また査読者の助言は本稿の改良に大きな力となった. ここに厚く感謝する.

参 考 文 献

[1] Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*, John Wiley & Sons. 佐和隆光，加納 悟（日本語訳）(1980)．回帰分析の実際，新曜社．
[2] 奥野忠一，久米 均，芳賀敏郎，吉澤 正 (1971)．多変量解析法，III 主成分分析，日科技連．

補 遺

第 4 節の手順中，凸多面体についての次の性質が利用される．

性質 1. $X=\{\boldsymbol{X}_i \in \boldsymbol{R}^p, i=1,\cdots,n\}$ $(n>p)$ とする．重心から測られた $\boldsymbol{X}_i$ を $\boldsymbol{x}_i$ とする．すなわち

$$\boldsymbol{x}_i = \boldsymbol{X}_i - \bar{\boldsymbol{X}}, \qquad \bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$$

そのとき

$$\hat{\boldsymbol{X}} = \{\boldsymbol{x}_i,\ i=1,\cdots,n\} \qquad (n>p)$$

とおくと，$\hat{\boldsymbol{X}}$ のつくる凸多面体の内部の点 $\boldsymbol{x}$ は，適当な $\hat{\boldsymbol{X}}$ の p 個のベクトル $\boldsymbol{x}_{i_1},\cdots,\boldsymbol{x}_{i_p}$ によって，

$$(*)\quad \boldsymbol{x} = \lambda_1\boldsymbol{x}_{i_1}+\cdots+\lambda_p\boldsymbol{x}_{i_p}, \qquad \lambda_i \geq 0,\ \sum_{i=1}^{p}\lambda_i < 1$$
$$(i_1,\cdots,i_p) \subset (1,\cdots,n)$$

と表わされる．

直観的証明. $\hat{\boldsymbol{X}}$ のつくる凸多面体の内点 $\boldsymbol{x}$ は適当に $\lambda(>1)$ 倍すれば凸多角形の境界と交わる．その点を $\boldsymbol{y}$ とする．しかるに $n>p$ のとき p 次元での凸多面体の境界は適当な X の点 $\boldsymbol{x}_{i_1},\cdots,\boldsymbol{x}_{i_p}$ を用いて，

$$\lambda_1\boldsymbol{x}_{i_1}+\cdots+\lambda_p\boldsymbol{x}_{i_p}, \qquad \lambda_i \geq 0,\ \sum_{i=1}^{p}\lambda_i = 1$$

と表わされる．

したがって，

$$\boldsymbol{x} = \frac{1}{\lambda}\boldsymbol{y} = \frac{1}{\lambda}(\lambda_1\boldsymbol{x}_{i_1}+\cdots+\lambda_p\boldsymbol{x}_{i_p})$$

改めて，$\lambda_i/\lambda=\lambda_i$ とおけば，

$$\boldsymbol{x} = \lambda_1\boldsymbol{x}_{i_1}+\cdots+\lambda_p\boldsymbol{x}_{i_p},\ \lambda_i \geq 0,\ \sum\lambda_i < 1$$

と表わされる．

性質 2. $\hat{\boldsymbol{X}}=\{\boldsymbol{x}_i=(x_{i1},\cdots,x_{ip})'\ i=1,\cdots,n\}$ の第 k 成分 x_{ik} の中で最大・最小な成分をもつベクトル $\boldsymbol{x}_{k\,\max}, \boldsymbol{x}_{k\,\min}$ は，他のどんな p 個のベクトルをとってきても (*) の形に表わせないことから柱ベクトルになっている．（但し，すべての i について $x_{ik}=0$ となる場合を除く．）

証明. 添字のわずらわしさを避けるため，$\boldsymbol{x}_{1\,\min}, \boldsymbol{x}_{1\,\max}$ のみについて証明する．他の k についても同様のやり方で証明できる．

補題 すべての $i=1,2,\cdots,n$ について，$x_{i1}=0$ ではないとする．そのとき，$x_{i1}, i=1,\cdots,n$ の中の最大値，最小値を $x_{1\,\max}$, $x_{1\,\min}$ とすると，$x_{1\,\max}>0$, $x_{1\,\min}<0$ である．

証明 $x_{1\,\min}<0$ について示す.

まず, $x_{1\,\min}>0$ とする. $\boldsymbol{x}_i$ は重心から測ったベクトルであるので, $\sum_{i=1}^{n}\boldsymbol{x}_i=\boldsymbol{0}$ ゆえ $\sum_{i=1}^{n}x_{i1}=0$ である. $x_{1\,\min}>0$ より, すべての i について $x_{i1}>0$, よって $\sum_{i=1}^{n}x_{i1}>0$ (矛盾). 次に $x_{i1}=0$ とすると, 矛盾がおこらないのはすべての i について $x_{i1}=0$ (元のデータでは $X_{i1}=$ 定数) となっているとき. この場合は仮定から除外される. すると, 残る可能な場合は $x_{1\,\min}<0$ の場合のみである. $x_{1\,\max}>0$ についても同様に証明できる.

$\boldsymbol{x}_{1\,\max}$ が柱ベクトルであることの証明.

仮に, $\boldsymbol{x}_{1\,\max}$ が他の p 個のベクトルによって (*) の形に表わされたとする. 特に第 1 成分 $x_{1\,\max}$ に注目すると,

$$x_{1\,\max}=a_1x_{i_1\,1}+\cdots+a_px_{i_p\,1},\quad a_i\geq 0,\quad \sum a_i<1$$

となっている. しかるに, $x_{1\,\max}$ は第 1 成分中の最大値であるので,

$$\begin{aligned}x_{1\,\max}&\leq a_1x_{1\,\max}+\cdots+a_px_{1\,\max}\\&=(a_1+\cdots+a_p)x_{1\,\max}<x_{1\,\max}\end{aligned}$$

これは矛盾. したがって $\boldsymbol{x}_{1\,\max}$ は (*) のようには表わしえない.

$\boldsymbol{x}_{1\,\min}$ が柱ベクトルであることの証明.

$\boldsymbol{x}_{1\,\min}$ が (*) のように表わされたとして, 第 1 成分に注目すれば,

$$x_{1\,\min}=a_1x_{i_1\,1}+\cdots+a_px_{i_p\,1},\quad a_i\geq 0,\quad \sum a_i<1$$

となっている. 補題より,

$$\begin{aligned}&x_{i_k\,1}\geq 0 \text{ に対しては, } -x_{i_k\,1}\leq 0<-x_{1\,\min}\\&x_{i_k\,1}<0 \text{ に対しては, } -x_{i_k\,1}\leq -x_{1\,\min}\end{aligned}$$

であるから

$$\begin{aligned}-x_{1\,\min}&=a_1(-x_{i_1\,1})+\cdots+a_p(-x_{i_p\,1})\\&\leq(a_1+\cdots+a_p)(-x_{1\,\min})\\&<-x_{1\,\min}\ (\text{矛盾}).\end{aligned}$$

A Search for Structure of the Explanatory Variables in the Linear Regression

Nobuyuki Kawai
(The Institute of Statistical Mathematics)

For regression analysis, it is desired to know the mechanism of the model.

But it is substantially impossible when the arrangement of the explanatory variables is ill. For instance, in the 2-dimensional space, the ill arrangement is the case that the observed explanatory variables are all concentrated on a covex object like a circle. There is no way to check the model over the wide interior range of the convex object.

In this paper, the method of detecting the arrangement of this type is proposed. The algorithm of the method is also given.