1996年度日本オペレーションズ・リサーチ学会
春季研究発表会

不均一構造・不完全情報下での主成分分析

01300450　日本大学　*高橋磐郎　TAKAHASHI Iwaro
01011500　日本大学　大澤慶吉　OSAWA Keikichi
01404360　日本大学　西澤一友　NISHIZAWA Kazutomo
02991410　日本大学　王　克義　Keyi WANG

§１．はじめに

主成分分析[1]のデータは、一般にp種の項目 A_1、・・・、A_p に関するN個の個体を対象とする観測データから構成される。個体nのA_jに関する観測値をa_{nj}とするとき、n=1～N, j=1～p のすべての組み合わせに対してa_{nj}が揃っていない表１のような場合がしばしば起こる。

表１　不完全情報データ

n	A_1	A_2	A_3	A_4	p=4
1	a_{11}	a_{12}		a_{14}	
2	a_{21}	a_{22}	a_{23}	a_{24}	
3	a_{31}		a_{33}	a_{34}	
4		a_{42}	a_{43}	a_{44}	
5	a_{51}	a_{52}	a_{53}		
6	a_{61}		a_{63}	a_{64}	
7		a_{72}		a_{74}	
8	a_{81}	a_{82}	a_{83}	a_{84}	

N=8

データの欠けている部分を欠測部、データが存在する部分を実測部ということにする。

このような事が起こる原因として２通りの場合が考えられる。１つは欠測部の(n,j)の組合せは、構造上もともと起こり得ないものである場合、もう１つは構造的にはすべてのデータを採り得るが、情報が不完全のため、たまたまデータが欠損して、欠測部が生じた場合である。前者を不均一構造の場合、後者を不完全情報の場合と呼ぶことにしよう。両者は見かけ上は表１のように同じ姿をしているが、データ構造としては異なるものと考えねばならない。

前回[2]では（標題は「不完全情報」となっているが）じつは不均一構造の場合を取り扱ったことになるが、今回は主として不完全情報の場合について考察しよう。

§２．不完全情報の主成分分析のデータモデル

不完全情報の欠測部は観測すればデータを採れる可能性はあるが、たまたま欠落したものである。そこで対象個体n、項目jが欠測部である場合そのデータα_{nj}のデータ構造を、ここでは次のように仮定することにする。

(1)　$\alpha_{nj} = \beta p_n + \gamma q_j$

つまり(1)式によって欠測部のデータを予測しようということである。ここでp_nは個体nの個体平均、q_jは項目A_jの項目平均（→(3)）で、α、β、γは未知パラメータで、以下の方法で推定されるものである。

たとえば、個体を学生と考え、項目を授業科目と考え、選択科目方式の問題を不完全情報とみなすと、学生nがもし科目A_jを受けたとすると、どの程度の得点をとるかを予測する式が(1)である。

(1)は最も簡単なモデルであるが、問題によって各項目あるいは各個体の特殊性について何らかの情報があれば、それを考慮したモデルを作ることもできる。(1)を用いてα_{nj}を予測すれば、あとは完全情報の場合と同様に（第一）主成分のスコアを求め、それによって、各個体の総合評価を行うことができる。第一主成分が個体の総合評価として最も適切であることは[2]にも触れた。第二主成分以下については今回は特に考慮しない。

§３．不完全情報の主成分分析の原則

個体n、項目A_jに対し、欠測部では$\delta_{nj}=0$、$\varepsilon_{nj}=1$、実測部では$\delta_{nj}=1$、$\varepsilon_{nj}=0$とすると、n=1～N、j=1～pの全体に対して

(2) $\alpha_{nj}=\delta_{nj}a_{nj}+\varepsilon_{nj}(\alpha+\beta p_n+\gamma q_j)$

と書ける。ここで

(3) $\begin{cases} p_n = \sum_j \delta_{nj}a_{nj} / \sum_j \delta_{nj} & \cdots\text{個体平均} \\ q_j = \sum_n \delta_{nj}a_{nj} / \sum_n \delta_{nj} & \cdots\text{項目平均} \end{cases}$

個体nの第一主成分スコアをz_n、項目のウエイトを$u_1,\cdots,u_p$とすると

(4) $z_n = \sum_j \alpha_{nj}u_j$

となるがz_nの分散 $V_n = \Sigma(z_n-\bar{z})^2/N$, $\bar{z}=\Sigma z_n/N$ を $u_1^2+\cdots+u_p^2=1$ の下に最大にするように$u_1,\cdots,u_p,\alpha,\beta,\gamma$を決めようという考えが、ここで用いる原則である。この分散最大という原則は簡単ではあるが最も合理的な情報縮約を与えることは[2]にも述べたところである。

§４．解析および計算方式

上記の原則に基づく解の候補は、ラグランジュ関数

$L= Vz - \lambda(u_1^2+\cdots+u_p^2-1)$

の停留点の中に求められる。つまり

(5) $\partial L/\partial u_i = 0$ (i=1～p)

(6) $\partial L/\partial \alpha=0$, $\partial L/\partial \beta=0$, $\partial L/\partial \gamma=0$

の解となるが、このうち(5),(6)からそれぞれ次の式(7),(8)が得られる。

(7) $Ru = \lambda u$, $u^T=(u_1,\cdots,u_p)$

(8) $\begin{cases} (EE)\alpha+(PE)\beta+(QE)\gamma+(AE)=0 \\ (PE)\alpha+(PP)\beta+(QP)\gamma+(AP)=0 \\ (QE)\alpha+(QP)\beta+(QQ)\gamma+(AQ)=0 \end{cases}$

ここで行列R、スカラー(EE),(PE),・・等は次のように与えられる；

(9) $R=[r_{ij}]$, $r_{ij}=\sum_n \dot{\alpha}_{ni}\dot{\alpha}_{nj}/N$

(10) $(EE)=\sum_i\sum_j\sum_n \dot{\varepsilon}_{ni}\dot{\varepsilon}_{nj}u_iu_j$,

$(PE)=\sum_i\sum_j\sum_n \dot{p}_{ni}\dot{\varepsilon}_{nj}u_iu_j$,

$(QE)=\sum_i\sum_j\sum_n \dot{q}_{ni}\dot{\varepsilon}_{nj}u_iu_j$,

$(AE)=\sum_i\sum_j\sum_n \dot{a}_{ni}\dot{\varepsilon}_{nj}u_iu_j$, ・・・・

(11) $\dot{\alpha}_{nj}=\dot{a}_{nj}+\dot{\varepsilon}_{nj}\alpha+\dot{p}_{nj}\beta+\dot{q}_{nj}\gamma$

$\dot{a}_{nj}=a_{nj}-\bar{a}_j$ $(\bar{a}_j=\sum_n \delta_{nj}a_{nj}/N)$

$\dot{\varepsilon}_{nj}=\varepsilon_{nj}-\bar{\varepsilon}_j$ $(\bar{\varepsilon}_j=\sum_n \varepsilon_{nj}/N)$

$\dot{p}_{nj}=\varepsilon_{nj}p_n-\bar{p}_j$ $(\bar{p}_j=\sum_n \varepsilon_{nj}p_j/N)$

$\dot{q}_{nj}=\varepsilon_{nj}q_n-\bar{q}_j$ $(\bar{q}_j=\sum_n \varepsilon_{nj}q_j/N)$

(7)のRの中には、α、β、γが含まれているし、(8)の(EE),(PE),・・・等には$u_1,\cdots,u_p$が含まれているので、(7),(8)を連立させて解くことは殆ど不可能だから、次のような逐次近似計算を用いる；

$u_1,\cdots,u_p$の初期値として$u_1=\cdots=u_p=1/\sqrt{p}$を定め、(10)より(EE),(PE),・・・等を計算して(8)を解きα、β、γを求める。これを用いて(11),(9)よりRを求め(7)の主（最大固有値に対する）固有ベクトル$u_1,\cdots,u_p$を求める。これを新たな初期値として以上の手順を繰り返す。

§５．応用例および今後の課題

選択授業科目の問題が不完全情報に属するか不均一構造に属するかは微妙な問題であるが、ある大学のある学科についてこの方式を適用してみた。このような場合、第二主成分の問題をどのように扱うかまたどのように解釈すべきか等今後の問題として検討したい。

参考文献

[1]奥野忠一他：「多変量解析法」日科技連、１９７１

[2]高橋磐郎他：不完全情報下での主成分分析、１９９５年日本ＯＲ学会秋季研究発表会アブストラクト集、pp60-61(1995).