1995年度日本オペレーションズ・リサーチ学会
秋季研究発表会

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]の原データは、一般にp種類の項目A_1、・・・、A_pに関するN個の対象の観測値から構成されている。対象nのA_iに関する観測値をa_{ni}とするとき、$n=1\sim N$, $i=1\sim p$のすべての組み合わせに対してa_{ni}が揃っていない、たとえば表1のような、場合がしばしば起こる。このような場合を不完全情報と名付ける。また、データが欠如している部分を欠測部という。これに対してデータが完全に揃っている場合を完全情報という。

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}	

$N=8$

たとえば、A_1、・・・、A_pをある大学の学科目とし、対象を学生とするとき、a_{ni}は学生nの科目A_iの得点となるが、学生によって選択する科目が異なるので不完全情報となる。このような場合主成分分析をどのように行えばよいかがこの研究の目的である。

主成分分析の主な目的は、多数の（p種の）観測可能な項目のもつ情報を、できるだけ少数の変量、いわゆる主成分、に集約することにある。そのためたとえば第一主成分は、その分散が最大になるように、各項目のウエイトを決めるという原則で求められる。しかし、この第一主成分は、同時に、それと各項目の観測値の相関係数の二乗和が最大となるという性質をもっている。またこの第一主成分を説明変数とし、各項目の観測値を被説明変数とした場合の誤差の二乗和が最小になるという性質も持っている。以上から少なくとも第一主成分は与えられたデータの情報を最も合理的に集約したものと考えられる。

そこで、たとえば、上の例で学生の成績評価をする場合、単なる形式的な平均点などよりは、第一主成分によるのが、より合理的と考えられる。また入試などで選択科目がある場合も不完全情報と考えられるが、その場合の評価もここで提案する方法が妥当と思われる。その他企業分析などでも、各項目ごとにデータが揃わない場合も実際には多く見られるので、不完全情報の主成分分析の方法は幅広い応用があるものと信ずる。

§2．第一主成分の解析

δ_{ni}を欠測部に対しては0、そうでなければ1としておく。項目A_iのデータ全体が欠測しているとか、対象nの全項目に対するデータが欠測していることはないものとする。

第一主成分の、項目A_jに対するウエイトをu_jとするとき、対象nのスコアz_nを

$$(1)\quad z_n = \sum_{j=1}^{p} \delta_{nj} a_{nj} u_j / U_n,$$

$$U_n = \left(\sum_{j=1}^{p} \delta_{nj} u_j^2\right)^{1/2},\ n=1\sim N$$

で定義する。ここで、U_nは対象nの有するデータに欠測がある場合それを補正する係数と考えればよい。たとえば、

表1では、z_1は

$$z_1 = \frac{a_{11}u_1}{\sqrt{u_1^2+u_2^2+u_4^2}} + \frac{a_{12}u_2}{\sqrt{u_1^2+u_2^2+u_4^2}} + \frac{a_{14}u_4}{\sqrt{u_1^2+u_2^2+u_4^2}}$$

となる。二乗和の平方根をU_nとしたのは、u_jの中には負となるものもあるからである。

第一主成分のウエイト$u_1,\cdots,u_p$を決める原則は、従来の主成分分析の場合と同様、z_nの分散V_z

(2) $V_z = \sum_{n=1}^{N}(z_n - \bar{z})^2 / N$

を、条件$u_1^2+u_2^2+\cdots+u_p^2 = 1$の下に最大にすることとする。このような$u_1,\cdots,u_p$は、条件式に対するラグランジュ乗数を$2\lambda/N$とすると、

(3) $\sum_n (z_n - \bar{z})\partial z_n / \partial u_i = \lambda u_i$, $(i=1\sim p)$

の解として求まる。

ここで

(4) $\frac{\partial z_n}{\partial u_i} = \frac{\delta_{ni}a_{ni}}{U_n} - \frac{\delta_{ni}u_i\sum_j \delta_{nj}a_{nj}u_j}{U_n^3}$

となるが、この右辺の第2項は不完全情報のための補正項と考えられるものである。また

(5) $z_n-\bar{z} = \sum_j(\delta_{nj}a_{nj}/U_n - \bar{a}_j)u_j$

$\bar{a}_j = (\sum_n \delta_{nj}a_{nj}/U_n)/N$ $(j=1\sim p)$

(4),(5)を(3)に代入して整理すると、$u_1,\cdots,u_p$を決める式は次の固有値問題となる。

(6)

$(M_{11}-M_1)u_1+M_{12}u_2+\cdots+M_{1p}u_p=\lambda u_1$

$M_{21}u_1+(M_{22}-M_2)u_2+\cdots+M_{2p}u_p=\lambda u_2$

$\cdots\cdots$

$M_{p1}u_1+M_{p2}u_2+\cdots+(M_{pp}-M_p)u_p=\lambda u_p$

つまり(6)の最大固有値に対する固有ベクトル（主固有ベクトル）が求める$u_1,\cdots,u_p$となる。ここで

(7)

$M_{ij}=\sum_n \delta_{ni}a_{ni}\delta_{nj}a_{nj}/U_n^2 - N\bar{a}_i\bar{a}_j$

$M_i=\sum_n(z_n-\bar{z})\sum_j \delta_{ni}\delta_{nj}a_{ni}u_j/U_n^3$

$(i,j=1\sim p)$

§3. 計算手順と第二以下の主成分

(6)は形式的には対称行列の固有値問題であるが、その主固有ベクトルを求めるにはべき乗法等を用いればよいが、M_{ij}, M_i自身が$u_1,\cdots,u_p$の関数となっている。そこでM_{ij}, M_iの適当な初期値を与え(6)の主固有ベクトルを求め、その解をM_{ij}, M_iに代入した新たな(6)の主固有ベクトルを求めるという逐次近似法を用いればよい。

M_{ij}, M_iの初期値としては；欠測部に対するデータa_{nj}としては、項目A_jの平均 $\sum_n \delta_{nj}a_{nj}/\sum_n \delta_{nj}$

を用い、$u_1=\cdots=u_p=1/\sqrt{p}$、$U_n=1$ $(n=1\sim N)$として(7)より計算したものを用いる。

第二主成分

(8) $y_n = \sum \delta_{nj}a_{nj}v_j/V_n$,

$V_n = \sqrt{\sum_j \delta_{nj}v_j^2}$

のウエイト$v_1,\cdots,v_p$を決める原則は、zとyとの共分散が0、

(9) $V_{z,y} = (z_n-\bar{z})(y_n-\bar{y}) = 0$

という条件と、$v_1^2+\cdots+v_p^2=1$という条件の下でy_nの分散V_y

(10) $V_y = \sum_{n=1}^{N}(y_n-\bar{y})^2/N$

を最大にすることである。

この解は、(9)に対するラグランジュ乗数が近似的に0になるから、(6)の行列$M=[M_{ij}-M_i]$の第二固有ベクトル$v_1,\cdots,v_p$となる。したがって

$M_1 = M - \lambda uu^T$

の主固有ベクトルをべき乗法によって求めればよい。この場合も、M_{ij}, M_i, λ, uは第一主成分の解法で得られたものを用いればよい。

参考文献

[1]奥野忠一他：「多変量解析法」日科技連、１９７１