データ解析演習問題

2006.10.30

提出期限: 2006 年 11 月 2 日 (木) 12:00
提出場所: システム棟 5F レポート提出 BOX

学籍番号: ____________________

氏名: ____________________

A.

以下の表 A.1 のように, 3 変数 x, y, z に関する n 個のデータ与えられている. 3 変数 x, y, z

表 A.1: 変数 x, u, y のデータ

	x	y	z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	z_n

から合成変数

$$u = ax + by + cz \tag{A.1}$$

を作ることを考える. ここで, $a^2 + b^2 + c^2 = 1$ である.

変数 u の平均 $\bar{u}$ は, $\bar{x}$, $\bar{y}$, $\bar{z}$ を用いて,

$$\bar{u} = a\bar{x} + b\bar{y} + c\bar{z} \tag{A.2}$$

と表わすことができる. (なぜそうなるかを答えられるようにしておきなさい.) 変数 u の分散は,

$$s_u^2 = a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + 2ab s_{xy} + 2bc s_{yz} + 2ca s_{zx} \tag{A.3}$$

と表わすことができる. (なぜそうなるかを答えられるようにしておきなさい.)

主成分分析では, $a^2 + b^2 + c^2 = 1$ という条件のもとで, s_u^2 が最大になるよう a, b, c を決定する. このような, 条件付き最大化問題を解くには, ラグランジュの未定係数法を用いればよい. 即ち, 以下のラグランジュ関数 $L(a, b, c, \lambda)$

$$L(a, b, c, \lambda) = a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + 2ab s_{xy} + 2bc s_{yz} + 2ca s_{zx} - \lambda(a^2 + b^2 + c^2 - 1) \tag{A.4}$$

を考えて, L の極値を求めればよい.

A.1.　以下の空欄に最もよくあてはまる 数式または語句を記入せよ.

L の極値を与える a, b, c, λ を求めるには, L を a, b, c, λ で偏微分して, それぞれ, 0 と置いた連立方程式の解を求めればよい.

$$\frac{\partial L}{\partial a} = \boxed{\text{ア } 2as_x^2 + 2bs_{xy} + 2cs_{zx} - 2\lambda a} = 0, \tag{A.5}$$

$$\frac{\partial L}{\partial b} = \boxed{\text{イ } 2bs_y^2 + 2as_{xy} + 2cs_{yz} - 2\lambda b} = 0, \tag{A.6}$$

$$\frac{\partial L}{\partial c} = \boxed{\text{ウ } 2cs_z^2 + 2bs_{yz} + 2as_{zx} - 2\lambda c} = 0, \tag{A.7}$$

$$\frac{\partial L}{\partial \lambda} = \boxed{\text{エ } -(a^2 + b^2 + c^2 - 1)} = 0, \tag{A.8}$$

(A.5)〜(A.7) を a,b,c について整理すると,

$$S \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix} \tag{A.9}$$

が得られる. ここで,

$$S = \begin{bmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{bmatrix} \tag{A.10}$$

は, 変数 x, y, z の分散共分散行列である. したがって, λ は S の $\boxed{\text{オ 固有値}}$ であって, $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ は $\boxed{\text{カ } \lambda \text{ に属する固有ベクトル}}$ である.

A.2. **式 (A.9) と式 (A.8) を用いて $s_u^2 = \lambda$ を証明せよ．(どこで, どの式を用いたのかを明確に記入せよ.)**

$$
\begin{aligned}
s_u^2 &= a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + 2ab s_{xy} + 2bc s_{yz} + 2ca s_{zx} \quad (\text{式 (A.3) より.}) \\
&= [a\ b\ c] S \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\
&= [a\ b\ c] \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad (\text{式 (A.9) より.}) \\
&= \lambda (a^2 + b^2 + c^2) \\
&= \lambda \quad (\text{式 (A.8) より.})
\end{aligned}
$$

B.

表 B.1 に示すような 4 変数 x, y, z, w に関するデータが得られているとする．このデータに対して, 主成分分析のモデル

$$u = ax + by + cz + dw$$

を考える．以下の設問に答えよ．

表 B.1: x, y, z, w に関するデータ

No.	x	y	z	w	u
1	182.0	67.5	85.1	93.7	124.87
2	176.3	63.5	83.5	93.0	120.21
3	179.0	63.0	87.0	96.3	122.23
4	172.8	88.0	101.0	94.1	149.91
5	170.0	61.5	83.0	91.8	117.71
6	164.5	58.5	83.5	88.5	114.75
7	170.6	57.0	83.0	88.1	113.60
8	170.4	79.0	98.0	92.5	140.47
9	165.0	63.3	89.5	89.7	122.14
10	172.5	66.0	91.4	91.4	126.12

B.1. 表 B.1 のデータの分散共分散行列 S を計算せよ.

$$S = \begin{bmatrix} 28.12 & 7.45 & -2.05 & 10.05 \\ 7.45 & 83.00 & 51.95 & 10.90 \\ -2.05 & 51.95 & 38.02 & 4.98 \\ 10.05 & 10.90 & 4.98 & 6.01 \end{bmatrix}$$

B.2. S の最大の固有値と, それに属する固有ベクトルを求めよ.

最大の固有値 $= 118.78$. λ に属する固有ベクトル $\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0.068 \\ 0.832 \\ 0.540 \\ 0.110 \end{bmatrix}$.

B.3.

No. 1〜10 の各データに対する主成分得点 u を求めて, 表 B.1 の u の列に (小数点以下 2 桁まで) 記入せよ. (あるいは, Excel の表を印刷したものを, 添付せよ.) $\diamond$

C.

本講義「データ解析」についての感想, 要望, 質問等があれば記せ.