データ解析 2006 年度期末試験問題

静岡大学工学部システム工学科

安藤 和敏

2006 年 11 月 27 日

注意事項

- 学生証を, 写真のある面を上にして, 机の上に置いておくこと.
- (関数) 電卓のみ持ち込み可. 携帯電話の電卓機能の使用も不可.
- 試験の時間は 12:45-14:05 である.
- 問題用紙は持ち帰ってよい.
- 解答及び採点の結果は, Web ページ (`http://coconut.sys.eng.shizuoka.ac.jp/data/06/`) で公開する.
- 証明問題では, どこで, どの式を用いたのかを明確に記入せよ.
- 「小数点以下第 x 位まで記入せよ」という指示が問題文の中にあるときは, 小数点以下第 $x+1$ 位を四捨五入して小数点以下第 x 位まで記入せよ.

問題 1. (配点 26)

以下の表 1 のように, 3 変数 x, y, z に関する n 個の個体からなるデータが与えられている. 3 変数 x, y, z

表 1: 変数 x, y, z 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

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

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,

$$\bar{u} = \boxed{\text{ア}} \tag{2}$$

と表わすことができる. また, 変数 u の分散は

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

と表わすことができる.

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 が最大になるように 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 $L(a, b, c, \lambda)$

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

を考えて, L の極値を求めればよい.

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 と置いた以下の連立方程式の解を求めればよい.

$$\frac{\partial L}{\partial a} = \boxed{\text{イ}} = 0, \tag{5}$$

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

$$\frac{\partial L}{\partial c} = 2cs_z^2 + 2bs_{yz} + 2as_{xz} - 2\lambda c = 0, \tag{7}$$

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

(5)〜(7) を a,b,c について整理すると,

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

が得られる. ここで,

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

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 の固有値であって, $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ は λ に属する固有ベクトルである. 式 (9) と式 (8) から, $s_u^2 = \lambda$ であるから, 求める s_u^2 を最大にする $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ は［ウ］である.

［ウ］を改めて $\begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix}$ と書いて, $u_1 = a_1x + b_1y + c_1z$ を第 1 主成分と呼ぶ. 同様の考え方で, $\begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix}$ を［エ］とすると第 2 主成分 $u_2 = a_2x + b_2y + c_2z$ が得られる.

設問 (1) 空欄［ア］〜［エ］に最もよくあてはまる数式または語句を記入せよ.

設問 (2) 式 (2) を用いて, 式 (3) を証明せよ.

設問 (3) 式 (9) と式 (8) を用いて $s_u^2 = \lambda$ を証明せよ.

問題 2. (配点 21)

表 2 に示すような 3 変数 x, y, z 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

$$u = ax + by + cz$$

を考える. 表 2 のデータの分散共分散行列 S を計算すると,

$$S = \begin{bmatrix} 2673.82 & 1648.29 & -694.93 \\ 1648.29 & 2206.11 & -785.88 \\ -694.93 & -785.88 & 481.86 \end{bmatrix}$$

であった. さらに, S の固有値は, $174.97450, 804.40381, 4382.41169$ の 3 つであり, これらの固有値に属する固有ベクトルは, それぞれ,

$$\begin{bmatrix} -0.046 \\ -0.327 \\ -0.944 \end{bmatrix}, \begin{bmatrix} -0.689 \\ 0.694 \\ -0.207 \end{bmatrix}, \begin{bmatrix} 0.723 \\ 0.641 \\ -0.258 \end{bmatrix}$$

表 2: x, y, z に関するデータ

No.	x	y	z
1	401.2	58.4	19.7
2	436.0	47.5	-16.5
3	425.7	61.9	27.2
4	367.6	-20.9	44.9
5	431.4	66.1	-10.8
6	442.7	127.6	-18.9
7	485.2	99.0	-18.8
8	467.1	74.7	11.3
9	319.6	59.5	9.5
10	336.9	-37.0	31.3

であった.

以下の設問に答えよ.

設問 (1) 第 1 主成分 $u_1 = a_1x + b_1y + c_1z$ の係数 a_1, b_1, c_1 と第 2 主成分 $u_2 = a_2x + b_2y + c_2z$ の係数 a_2, b_2, c_2 を求めて, 小数点以下第 3 位まで記入せよ.

設問 (2) 第 1 主成分と第 2 主成分の寄与率, 及び, 第 2 主成分までの累積寄与率を求めて, 小数点以下第 3 位まで記入せよ.

設問 (3) No. 10 の個体に対する主成分得点 u_1 と u_2 を求めて, 小数点以下 1 桁まで記入せよ.

問題 3. (配点 22)

判別分析では, 表 3 のようなデータを扱う. n 個の個体は, グループ P とグループ Q の 2 つのグループに分割されており, 個体番号 1 から m はグループ P に, 個体番号 $m+1$ から n まではグループ Q に属している.

判別分析の目的は, これらの 2 つのグループがなるべく遠ざかって見えるように, 合成変数

$$z = ax + by + c \tag{11}$$

を求めることである. z を x と y の関数として見るときは, これを線形判別関数と呼ぶ.

$n_P = m, n_Q = n - m$ とする. すなわち, n_P と n_Q は, それぞれ, グループ P とグループ Q に属する個体の数である. $\bar{z}_P$ と $\bar{z}_Q$ を

$$\bar{z}_P = \frac{1}{n_P}\sum_{i=1}^{m} z_i, \quad \bar{z}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} z_j, \tag{12}$$

で定義する. すなわち, $\bar{z}_P$ はグループ P における z の平均であり, $\bar{z}_Q$ はグループ Q における z の平均である.

表 3: 判別分析のデータ

No.	x	y	P/Q
1	x_1	y_1	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m	y_m	P
$m+1$	x_{m+1}	y_{m+1}	Q
$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_j	y_j	Q
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	Q

z の変動 S_T

$$S_T = \sum_{k=1}^{n} (z_k - \bar{z})^2 \tag{13}$$

は,

$$S_T = S_B + S_W \tag{14}$$

と分解される. ここで, S_W と S_B は, それぞれ,

$$S_W = \sum_{i=1}^{m} (z_i - \bar{z}_P)^2 + \sum_{j=m+1}^{n} (z_j - \bar{z}_Q)^2 \tag{15}$$

$$S_B = n_P(\bar{z}_P - \bar{z})^2 + n_Q(\bar{z}_Q - \bar{z})^2 \tag{16}$$

と定義される.

S_B は 2 つのグループがどれだけ離れているかを表していると解釈できるため, S_B は [ア] と呼ばれる. また, S_W は, グループ P 内での z の変動とグループ Q 内での z の変動の和であるので, S_W は [イ] と呼ばれる. これらの観察から, 「2 つのグループがなるべく遠ざかって見える」ように, 合成変数 $z = ax + by + c$ を求めることは, 相関比

$$\eta^2 = \frac{S_B}{S_T} \tag{17}$$

を最大にするような a, b, c を求めることであることが分かる.

S_T と S_B は, それぞれ, a, b を用いて,

$$S_T = [\text{ウ}], \tag{18}$$

$$S_B = [\text{エ}] \tag{19}$$

というように書ける. ここで,

$$\bar{x}_P = \frac{1}{n_P}\sum_{i=1}^{m} x_i, \quad \bar{x}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} x_j, \tag{20}$$

$$\bar{y}_P = \frac{1}{n_P}\sum_{i=1}^{m} y_i, \quad \bar{y}_Q = \frac{1}{n_Q}\sum_{j=m+1}^{n} y_j \tag{21}$$

である.

したがって, 相関比 η^2 も

$$\eta^2 = \frac{\boxed{\text{エ}}}{\boxed{\text{ウ}}} \tag{22}$$

と書き直せる. 式 (22) で, $t = \frac{a}{b}$ と置いて, $\frac{d\eta^2}{dt} = 0$ を解けば, η^2 を最大にする $t = \frac{a}{b}$ が求まる.

z の分散 s_z^2 は 1 に等しいという仮定を置くと, b は t, s_x^2, s_{xy}, s_y^2 を用いて,

$$b = \frac{1}{\sqrt{\boxed{\text{オ}}}} \tag{23}$$

と表される. また, a は $a = bt$ によって求められる. こうして, a と b が決定される. さらに, 直線 $0 = ax + by + c$ が $(\bar{x}_P, \bar{y}_P)$ と $(\bar{x}_Q, \bar{y}_Q)$ の中点を通るように c を決めれば, c も決定される.

こうして求められた線形判別関数 $z = ax + by + c$ を用いて, 各個体が P と Q のいずれのグループに分類されるかを判定することができる.

設問 (1) 空欄 ア ～ オ に最もよくあてはまる数式または語句を記入せよ.

設問 (2) 式 (14) を証明せよ. このとき, 以下の 2 式を用いよ.

$$\sum_{i=1}^{m}(z_i - \bar{z}_P) = 0, \tag{24}$$

$$\sum_{j=m+1}^{n}(z_j - \bar{z}_Q) = 0. \tag{25}$$

問題 4. (配点 31)

表 4 に示すような 2 変数 x, y に関する判別分析のデータが得られているとする. このデータに対して, 線形判別関数

$$z = ax + by + c \tag{26}$$

を考える. 以下の設問に答えよ.

設問 (1) η^2 を最大にする $t = \frac{a}{b}$ は, -0.676 であることが分かっている. a, b, c を求めて, その結果を小数点以下第 3 位まで記入せよ.

表 4: x, y に関するデータ

No.	x	y	P/Q
1	43.0	16.9	P
2	56.0	21.6	P
3	38.0	12.2	P
4	21.0	16.0	P
5	25.0	10.5	P
6	50.0	15.5	Q
7	69.0	18.4	Q
8	93.0	26.4	Q
9	76.0	22.9	Q
10	88.0	18.6	Q
平均値	55.90	17.90	
分散	567.69	20.93	
共分散	86.93		

表 5: グループ P 関するデータ

No.	x	y	P/Q
1	43.0	16.9	P
2	56.0	21.6	P
3	38.0	12.2	P
4	21.0	16.0	P
5	25.0	10.5	P
平均値	33.60	15.44	
分散	159.44	15.06	
共分散	34.58		

表 6: グループ Q 関するデータ

No.	x	y	P/Q
6	50.0	15.5	Q
7	69.0	18.4	Q
8	93.0	26.4	Q
9	76.0	22.9	Q
10	88.0	18.6	Q
平均値	75.20	20.36	
分散	230.96	14.70	
共分散	44.33		

設問 (2) 設問 1 で求められた a, b, c に対して, S_B, S_W, 相関比 η^2 を求め, その結果を小数点以下第 3 位まで記入せよ.

設問 (3) No. 6 の個体に対する判別得点 z_6 を求めて, 小数点以下第 2 位まで記入せよ. 線形判別関数によって, No. 6 の個体はどちらのグループに属すると判定されるか?

設問 (4) No. 6 の個体 (x_6, y_6) の $(\bar{x}_P, \bar{y}_P)$ までのマハラノビス距離 D_P, 及び, $(\bar{x}_Q, \bar{y}_Q)$ までのマハラノビス距離 D_Q を求めて, 小数点以下第 2 位まで記入せよ. 計算のために, 以下の数値を用いよ.

$$\begin{bmatrix} 159.44 & 34.58 \\ 34.58 & 15.06 \end{bmatrix}^{-1} = \begin{bmatrix} 0.012 & -0.029 \\ -0.029 & 0.132 \end{bmatrix}, \quad \begin{bmatrix} 230.96 & 44.33 \\ 44.33 & 14.70 \end{bmatrix}^{-1} = \begin{bmatrix} 0.010 & -0.031 \\ -0.031 & 0.162 \end{bmatrix}.$$

マハラノビス距離によって, No. 6 の個体はどちらのグループに属すると判定されるか?

学籍番号		氏名	

問題 1(1) の解答欄 (配点 12)

ア. $\underline{a\bar{x} + b\bar{y} + c\bar{z}\ (3\text{ 点})}$

イ. $\underline{2as_x^2 + 2bs_{xy} + 2cs_{zx} - 2\lambda a\ (3\text{ 点})}$

ウ. $\underline{S\text{ の最大の固有値に属する固有ベクトル (3 点)}}$

エ. $\underline{S\text{ の 2 番目に大きい固有値に属する固有ベクトル (3 点)}}$

問題 1(2) の解答欄 (配点 7)

$$
\begin{aligned}
s_u^2 &= \frac{1}{n}\sum_{i=1}(u_i - \bar{u})^2 \\
&= \frac{1}{n}\sum_{i=1}(ax_i + by_i + cz_i - a\bar{x} - b\bar{y} - c\bar{z})^2 \quad (\text{式 (1) と式 (2) を用いた.}) \\
&= \frac{1}{n}\sum_{i=1}\{a(x_i - \bar{x}) + b(y_i - \bar{y}) + c(z_i - \bar{z})\}^2 \\
&= \frac{1}{n}\sum_{i=1}\{a^2(x_i - \bar{x})^2 + b^2(y_i - \bar{y})^2 + c^2(z_i - \bar{z})^2 \\
&\quad +2ab(x_i - \bar{x})(y_i - \bar{y}) + 2bc(y_i - \bar{y})(z_i - \bar{z}) + 2ca(z_i - \bar{z})(x_i - \bar{x})\} \\
&= a^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + b^2\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + c^2\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 \\
&\quad +2ab\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + 2bc\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z}) + 2ca\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})\} \\
&= a^2s_x^2 + b^2s_y^2 + c^2s_z^2 + 2abs_{xy} + 2bcs_{yz} + 2cas_{zx}.
\end{aligned}
$$

学籍番号		氏名	

問題 1(3) の解答欄 (配点 7)

$$
\begin{aligned}
s_u^2 &= a^2 s_x^2 + b^2 s_y^2 + c^2 s_z^2 + 2abs_{xy} + 2bcs_{yz} + 2cas_{zx} \\
&= [abc] S \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\
&= [abc] \lambda \begin{bmatrix} a \\ b \\ c \end{bmatrix} \quad \text{(式 (9) を用いた.)} \\
&= \lambda [abc] \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\
&= \lambda (a^2 + b^2 + c^2) \\
&= \lambda \quad \text{(式 (8) を用いた).}
\end{aligned}
$$

問題 2(1) の解答欄 (配点 3+3)

$a_1 =$ 0.723 (1 点)　　$b_1 =$ 0.641 (1 点)　　$c_1 =$ -0.258 (1 点)

$a_2 =$ 0.689 (1 点)　　$b_2 =$ -0.694 (1 点)　　$c_2 =$ 0.207 (1 点)

問題 2(2) の解答欄 (配点 3+3+3)

$C_1 =$ 0.817 (3 点)　　$C_2 =$ 0.150 (3 点)　　$C_{\text{all}} =$ 0.967 (3 点)

問題 2(3) の解答欄 (配点 3+3)

$u_1 =$ 211.8 (3 点)　　$u_2 =$ 264.3 (3 点)

学籍番号		氏名	

問題 3(1) の解答欄 (配点 13)

ア. 群間変動 (2 点)　　イ. 群内変動 (2 点)

ウ. $n(a^2s_x^2 + 2absx_y + s_y^2b^2)$ (3 点)

エ. $n_P\{a(\bar{x}_P - \bar{x}) + b(\bar{y}_P - \bar{y})\}^2 + n_Q\{a(\bar{x}_Q - \bar{x}) + b(\bar{y}_Q - \bar{y})\}^2$ (3 点)

オ. $t^2s_x^2 + 2ts_{xy} + s_y^2$ (3 点)

問題 3(2) の解答欄 (配点 9)

$$
\begin{aligned}
S_T &= \sum_{k=1}^{n}(z_k - \bar{z})^2 \\
&= \sum_{i=1}^{m}(z_i - \bar{z})^2 + \sum_{j=m+1}^{n}(z_j - \bar{z})^2 \\
&= \sum_{i=1}^{m}(z_i - \bar{z}_P + \bar{z}_P - \bar{z})^2 + \sum_{j=m+1}^{n}(z_j - \bar{z}_Q + \bar{z}_Q - \bar{z})^2 \\
&= \sum_{i=1}^{m}\{(z_i - \bar{z}_P)^2 + 2(z_i - \bar{z}_P)(\bar{z}_P - \bar{z}) + (\bar{z}_P - \bar{z})^2\} \\
&\quad + \sum_{j=m+1}^{n}\{(z_j - \bar{z}_Q)^2 + 2(z_j - \bar{z}_Q)(\bar{z}_Q - \bar{z}) + (\bar{z}_Q - \bar{z})^2\} \\
&= \sum_{i=1}^{m}(z_i - \bar{z}_P)^2 + 2\sum_{i=1}^{m}(z_i - \bar{z}_P)(\bar{z}_P - \bar{z}) + \sum_{i=1}^{m}(\bar{z}_P - \bar{z})^2 \\
&\quad + \sum_{j=m+1}^{n}(z_j - \bar{z}_Q)^2 + 2\sum_{j=m+1}^{n}(z_j - \bar{z}_Q)(\bar{z}_Q - \bar{z}) + \sum_{j=m+1}^{n}(\bar{z}_Q - \bar{z})^2 \\
&= \sum_{i=1}^{m}(z_i - \bar{z}_P)^2 + \sum_{j=m+1}^{n}(z_j - \bar{z}_Q)^2 + \sum_{i=1}^{m}(\bar{z}_P - \bar{z})^2 + \sum_{j=m+1}^{n}(\bar{z}_Q - \bar{z})^2 \quad \text{(式 (24) と式 (25) を用いた)} \\
&= \sum_{i=1}^{m}(z_i - \bar{z}_P)^2 + \sum_{j=m+1}^{n}(z_j - \bar{z}_Q)^2 + n_P(\bar{z}_P - \bar{z})^2 + n_Q(\bar{z}_Q - \bar{z})^2 \\
&= S_W + S_B \quad \text{(式 (15) と式 (16) より)}.
\end{aligned}
$$

学籍番号		氏名	

問題 4(1) の解答欄 (配点 9)

$a =$ -0.053　　$b =$ 0.078　　$c =$ 1.567 (1.487)

問題 4(2) の解答欄 (配点 9)

$S_B =$ 6.906　　$S_W =$ 3.094　　$\eta^2 =$ 0.691

問題 4(3) の解答欄 (配点 5)

$z_6 =$ 0.13 (0.05)　　判定 P

問題 4(4) の解答欄 (配点 8)

$D_P =$ 2.11 (3.17)　　$D_Q =$ 2.58

判定 P (Q)

注意: 出題ミスがありました. 表 5 の x の平均値 33.60 というのは間違いで, 36.60 というのが正しい. したがって, 問題 4(1),(2),(3) において, $x_P = 33.60$ で計算すると括弧内の答となり, $x_P = 36.60$ で計算すると括弧の外の答となる. どちらも正解とする.