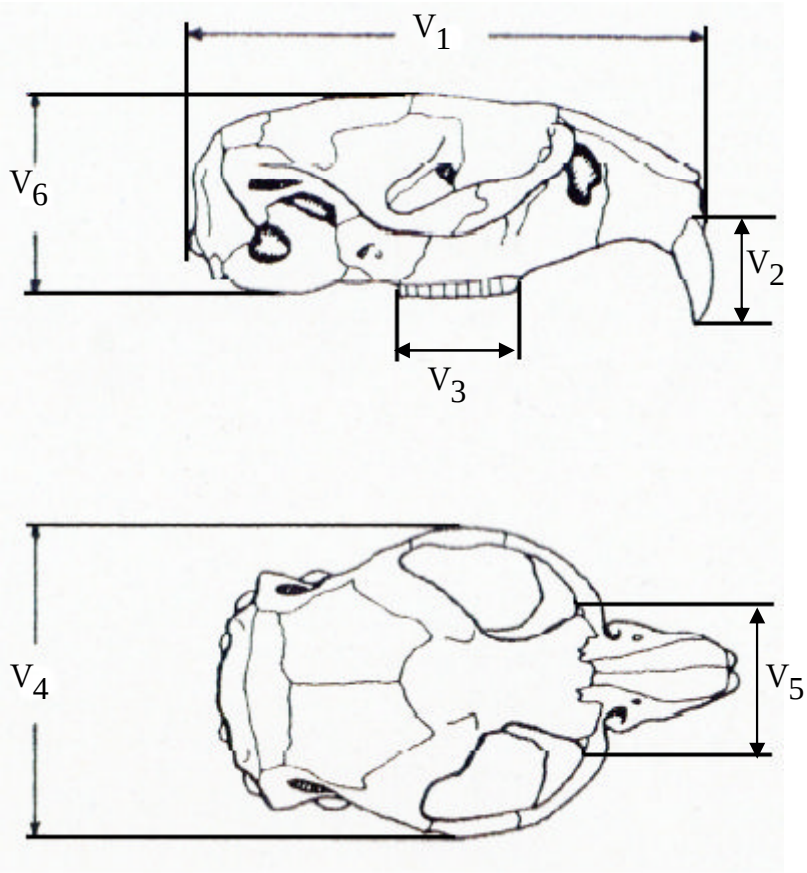
2母集団における固有ベクトルの検定

Shin-ichi Tsukada　塚田真一(明星大学)
Takakazu Sugiyama　杉山髙一(中央大学)

科研費 基盤研究B 代表者 杉山 高一
「高次元データの推測理論の開発と応用」
2009年12月5日～6日

Introduction

Airoldi and Hoffmann [1] は2種類のハタネズミの頭蓋骨を計測した。



Microtus californicus
($N_1 = 40$)

Microtus ochrogaster
($N_2 = 44$)

平均と標準偏差

Microtus Californicus						
	V_1	V_2	V_3	V_4	V_5	V_6
Mean	280.83	50.32	72.37	160.24	34.56	109.12
S.D.	12.22	4.52	3.24	6.48	1.73	4.09

Microtus Ochrogaster						
	V_1	V_2	V_3	V_4	V_5	V_6
Mean	266.64	44.76	64.33	150.31	39.87	104.27
S.D.	9.11	3.09	2.42	6.33	1.36	2.92

主成分分析の結果

Microtus Californicus		Variation	Coefficients					
Component No.	Latent roots	explained	V_1	V_2	V_3	V_4	V_5	V_6
1	198.7	.822	.850	.208	.191	.394	.010	.206
2	20.69	.086	.519	−.264	−.232	−.674	−.133	−.367

Microtus Ochrogaster		Variation	Coefficients					
Component No.	Latent roots	explained	V_1	V_2	V_3	V_4	V_5	V_6
1	117.3	.788	.819	.186	.151	.479	.009	.206
2	17.61	.118	.453	.206	.070	−.862	.052	−.039

Kzanowski(1979)

大きい固有値に対するいくつかの固有ベクトルで張られる部分空間の近さを与えた。

Flury(1988)

Common Principal Components を提案し、正規性の下での尤度比検定を提案している。

Flury(1988)

Common Principal Components

$$\Sigma_g = \Gamma \Lambda_g \Gamma'$$

Partial Common Principal components

$$\Sigma_g = \Gamma_g \Lambda_g \Gamma_g, \quad \Gamma_g = (\Gamma_1 : \Gamma_{2g})'$$

$\Lambda_g = \mathrm{diag}(\lambda_1^{(g)}, \ldots, \lambda_p^{(g)})$　Γ は直交行列

（ただし、$\lambda_j^{(g)}$は大きさの順に並んでいない。）

仮説検定

ここで考える帰無仮説は

$$H_0 : \boldsymbol{\eta}_j^{(1)} = \boldsymbol{\eta}_j^{(2)},$$
$$H_1 : \boldsymbol{\eta}_j^{(1)} \neq \boldsymbol{\eta}_j^{(2)}$$

共分散行列の固有値はすべて異なり、母集団分布の4次モーメントまでが存在すると仮定する。

検定統計量

標本固有ベクトルは漸近的に、下記の正規分布に従う。

$$\sqrt{n}(\boldsymbol{h}_j - \boldsymbol{\eta}_j) \overset{a}{\sim} N(\mathbf{0}, V_{Gj}),$$

$$\begin{aligned} V_{Gj} &= \{\boldsymbol{\eta}_j' \otimes \Gamma(\lambda_j I - \Lambda)^+ \Gamma'\} M_4(x) \{\boldsymbol{\eta}_j \otimes \Gamma(\lambda_j I - \Lambda)^+ \Gamma'\}, \\ M_4(x) &= E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})' \otimes (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})']. \end{aligned}$$

検定統計量

$n = n_1 + n_2,\ r_k = n_k/n$ として、帰無仮説の下では

$$\sqrt{nr_1r_2}\left(\boldsymbol{h}_j^{(1)} - \boldsymbol{h}_j^{(2)}\right) \overset{a}{\sim} N(\mathbf{0}, r_2V_{Gj}^{(1)} + r_1V_{Gj}^{(2)}).$$

これより

$$C_G = nr_1r_2\left(\boldsymbol{h}_j^{(1)} - \boldsymbol{h}_j^{(2)}\right)'\left(r_2\hat{V}_{Gj}^{(1)} + r_1\hat{V}_{Gj}^{(2)}\right)^{-1}\left(\boldsymbol{h}_j^{(1)} - \boldsymbol{h}_j^{(2)}\right)$$

は漸近的に自由度 $(p-1)$ のカイ2乗分布に従う。

Criteria(3)

特に正規母集団や楕円母集団を仮定すると

$$V_N = \sum_{i \neq j}^{p} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i', V_E = (\kappa + 1) \sum_{i \neq j}^{p} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i',$$

κは kurtosis parameter.

$$C_N = n r_1 r_2 \left(\boldsymbol{h}_\alpha^{(1)} - \boldsymbol{h}_\alpha^{(2)} \right)' \left(r_2 \hat{V}_{Nj}^{(1)} + r_1 \hat{V}_{Nj}^{(2)} \right)^{-1} \left(\boldsymbol{h}_\alpha^{(1)} - \boldsymbol{h}_\alpha^{(2)} \right)$$

$$C_E = n r_1 r_2 \left(\boldsymbol{h}_\alpha^{(1)} - \boldsymbol{h}_\alpha^{(2)} \right)' \left(r_2 \hat{V}_{Ej}^{(1)} + r_1 \hat{V}_{Ej}^{(2)} \right)^{-1} \left(\boldsymbol{h}_\alpha^{(1)} - \boldsymbol{h}_\alpha^{(2)} \right)$$

Simulation

これらの検定統計量のカイ2乗分布への収束を調べるため、帰無仮説を下記のように設定し

$$
\begin{aligned}
H_0 &: \boldsymbol{\eta}_1^{(1)} = \boldsymbol{\eta}_1^{(2)}, \\
H_1 &: \boldsymbol{\eta}_1^{(1)} \neq \boldsymbol{\eta}_1^{(2)}
\end{aligned}
$$

標本数を100, 200, 500, 1000 とした。シミュレーション回数は100万回で 、有意水準を5%とした。

Simulation(2)

母集団は次のように設定した。

【 Case1 】(第1主成分の寄与率70%、第2主成分の寄与率20%)

$$\Lambda_{71} = \mathrm{diag}(160, 50, 8, 6, 4, 2, 1)$$

【 Case2 】(第1主成分の寄与率80%、第2主成分の寄与率10%)

$$\Lambda_{72} = \mathrm{diag}(240, 30, 16, 8, 4, 2, 1)$$

Simulation(3)

<u>共分散行列</u>

$$\Sigma_{71} = \Lambda_{7i}, \qquad \Sigma_{72} = \Gamma_p \Lambda_{7i} \Gamma_p' \ (i = 1, 2)$$

$$\Gamma_7 = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.981 & -0.086 & -0.086 & -0.086 & -0.087 & -0.087 \\ 0.000 & 0.056 & 0.984 & -0.084 & -0.085 & -0.086 & -0.087 \\ 0.000 & 0.069 & 0.058 & 0.985 & -0.083 & -0.085 & -0.086 \\ 0.000 & 0.083 & 0.070 & 0.058 & 0.985 & -0.084 & -0.086 \\ 0.000 & 0.098 & 0.084 & 0.070 & 0.058 & 0.984 & -0.086 \\ 0.000 & 0.114 & 0.098 & 0.083 & 0.069 & 0.056 & 0.981 \end{pmatrix}$$

Normal population

$$\Pi_1 \quad : \quad N_7(\mathbf{0}, \Sigma_{71}), \quad \Pi_2 \quad : \quad N_7(\mathbf{0}, \Sigma_{72})$$

Contaminated Normal population

$$\begin{aligned} \Pi_1 \quad &: \quad 0.1 N_7(\mathbf{0}, \Sigma_{71}) + 0.9 N_7(\mathbf{0}, 4^2 \Sigma_{71}), \\ \Pi_2 \quad &: \quad 0.1 N_7(\mathbf{0}, \Sigma_{72}) + 0.9 N_p(\mathbf{0}, 4^2 \Sigma_{72}) \end{aligned}$$

Skew Normal population

$$\begin{aligned} \Pi_1 \quad &: \quad SN_7(\mathbf{0}, \Omega_{71}, \boldsymbol{\alpha}_{71}), \\ \Pi_2 \quad &: \quad SN_7(\mathbf{0}, \Omega_{72}, \boldsymbol{\alpha}_{72}) \end{aligned}$$

$$
\begin{aligned}
\boldsymbol{\alpha}_{71} &= (-0.694, -2.069, -5.664, -3.705, 5.984, 81.21, 439.4)'/1000, \\
\boldsymbol{\alpha}_{72} &= (-0.752, 1.741, -0.406, -0.912, 2.090, 39.14, 209.8)'/1000,
\end{aligned}
$$

$$
\Omega_{71} = \begin{pmatrix}
25600.0 & 2000.0 & 320.0 & 240.0 & 160.0 & 80.00 & 40.00 \\
2000.0 & 2500.0 & 100.0 & 75.00 & 50.00 & 25.00 & 12.50 \\
320.0 & 100.0 & 64.00 & 12.00 & 8.00 & 4.00 & 2.00 \\
240.0 & 75.00 & 12.00 & 36.00 & 6.00 & 3.00 & 1.50 \\
160.0 & 50.00 & 8.00 & 6.00 & 16.00 & 2.00 & 1.00 \\
80.00 & 25.00 & 4.00 & 3.00 & 2.00 & 4.00 & 0.50 \\
40.00 & 12.50 & 2.00 & 1.50 & 1.00 & 0.50 & 1.00
\end{pmatrix},
$$

$$
\Omega_{72} = \begin{pmatrix}
25600.0 & 1929.7 & 326.9 & 251.2 & 174.6 & 97.83 & 59.68 \\
1929.7 & 2327.4 & 43.54 & 25.95 & 9.454 & -4.404 & -9.263 \\
326.9 & 43.54 & 66.81 & 13.31 & 8.778 & 4.375 & 2.311 \\
251.2 & 25.95 & 13.31 & 39.43 & 7.569 & 3.997 & 2.278 \\
174.6 & 9.454 & 8.778 & 7.569 & 19.06 & 3.315 & 2.006 \\
97.83 & -4.404 & 4.375 & 3.997 & 3.315 & 5.982 & 1.481 \\
59.68 & -9.263 & 2.311 & 2.278 & 2.006 & 1.481 & 2.226
\end{pmatrix}.
$$

表1 : Case 1の場合

Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.0951	.1090	.2006	.0966
200	.0803	.0866	.1683	.0898
500	.0721	.0741	.1500	.0851
1000	.0696	.0705	.1448	.0840
Contaminated Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.1380	.1139	.2070	.0960
200	.1181	.0881	.1708	.0888
500	.1074	.0747	.1508	.0849
1000	.1042	.0707	.1454	.0843
Skew Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.0863	.0987	.0809	.0791
200	.0728	.0775	.0625	.0728
500	.0666	.0675	.0582	.0703
1000	.0660	.0657	.0592	.0710

表2 : Case 2の場合

Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.0758	.0885	.1696	.0809
200	.0645	.0701	.1413	.0737
500	.0584	.0602	.1258	.0698
1000	.0563	.0571	.1209	.0683

Contaminated Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.1138	.0921	.1750	.0804
200	.0981	.0713	.1434	.0732
500	.0895	.0606	.1264	.0692
1000	.0867	.0573	.1214	.0685

Skew Normal population				
N_1, N_2	C_N	C_E	C_G	C_F
100	.0759	.0875	.0676	.0728
200	.0647	.0692	.0541	.0662
500	.0597	.0605	.0528	.0637
1000	.0592	.0589	.0537	.0641

Normal population

- 次元数が小さい場合には、C_Fの精度もよくなる。
- 標本数が大きくなると、C_E の精度も C_N の精度に近くなる。

Contaminated Normal population

- C_E の精度がよいが、次元数が小さい場合には、 C_F の精度もよくなる。

Skew Normal population

- 標本数が大きくなると C_G の精度がよくなる。

応用例

先程のハタネズミのデータに対して、有意水準5%で検定を行う。

主成分分析の結果

Microtus Californicus		Variation	Coefficients					
Component No.	Latent roots	explained	V_1	V_2	V_3	V_4	V_5	V_6
1	198.7	.822	.850	.208	.191	.394	.010	.206
2	20.69	.086	.519	−.264	−.232	−.674	−.133	−.367

Microtus Ochrogaster		Variation	Coefficients					
Component No.	Latent roots	explained	V_1	V_2	V_3	V_4	V_5	V_6
1	117.3	.788	.819	.186	.151	.479	.009	.206
2	17.61	.118	.453	.206	.070	−.862	.052	−.039

第１固有ベクトルについて仮説検定を行う。

$$H_0 : \boldsymbol{\eta}_1^{(1)} = \boldsymbol{\eta}_1^{(2)},$$
$$H_1 : \boldsymbol{\eta}_1^{(1)} \neq \boldsymbol{\eta}_1^{(2)}$$

検定統計量の値は

$$C_N = 4.490, C_E = 5.069, C_G = 8.017 \text{ and } C_F = 3.377.$$

棄却点は $\chi_5^2 = 11.070$ であるので、
帰無仮説は棄却さない。

第２固有ベクトルについて仮説検定を行う。

$$H_0 : \boldsymbol{\eta}_2^{(1)} = \boldsymbol{\eta}_2^{(2)},$$
$$H_1 : \boldsymbol{\eta}_2^{(1)} \neq \boldsymbol{\eta}_2^{(2)}$$

検定統計量の値は

$$C_N = 28.05, C_E = 31.60, C_G = 12.88 \text{ and } C_F = 17.73.$$

棄却点は $\chi_5^2 = 11.070$ であるので、
帰無仮説は棄却される。