多変量高次元データ解析の理論と応用

研究代表者　杉山高一　研究員

多変量高次元相関行列の各固有値の同等性検定
～パーミュテーションテストを用いて～

多変量高次元ベクトル変量 $\boldsymbol{X}^{(i)}, \quad i=1,2$

平均ベクトル $\mu^{(i)}$

相関行列 $P^{(i)}$

$$H_0 \ : \ \xi_j^{(1)} = \xi_j^{(2)}$$
$$H_1 \ : \ \xi_j^{(1)} \neq \xi_j^{(2)} \quad \text{for } j=1,\ldots,p$$

のような仮説を考える. ただし, $\xi_j^{(i)}$ を 第 i 母集団の j 番目固有値とする

N_i 個の無作為標本 $\boldsymbol{x}_1^{(i)},\cdots,\boldsymbol{x}_{N_1}^{(i)}$ より共分散行列 $S^{(i)}=(s_{jj}^{(i)})$ を作成

固有ベクトル $r_1^{(i)},\ldots,r_p^{(i)}, \quad (r_1^{(i)} \geq \cdots \geq r_p^{(i)})$

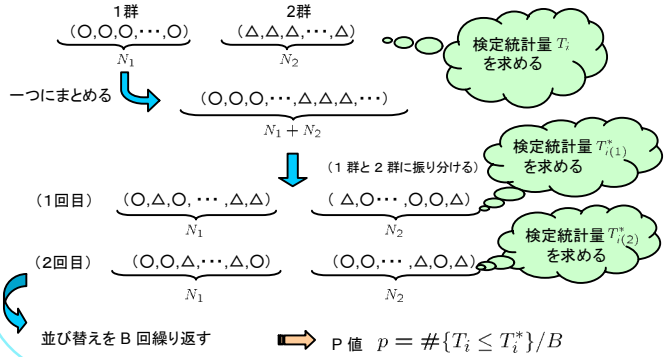
固有値 $\boldsymbol{b}_1^{(i)},\ldots,\boldsymbol{b}_p^{(i)}$

主成分スコア

$$z_{j\alpha}^{(i)} = \boldsymbol{b}_j^{\prime(i)}\left(\boldsymbol{x}_\alpha^{(i)} - \bar{\boldsymbol{x}}^{(i)}\right)/\sqrt{s_{jj}^{(i)}} \qquad i=1,2, \ \alpha=1,\ldots,N_i$$

検定統計量

$$T_1 = \sqrt{N}\log(r_j^{(2)}/r_j^{(1)}) \quad T_2 = \sqrt{N}(r_j^{(2)} - r_j^{(1)})$$
$$T_3 = \left|\frac{A_N - \mathrm{E}[A_N]}{\sqrt{\mathrm{var}[A_N]}}\right| \left(A_N = \frac{N_1(N+1)}{2} - \sum_{i=1}^{N}\left|i - \frac{N+1}{2}\right|V_i\right), \ T_4 = \left|\frac{\bar{M} - \mathrm{E}[\bar{M}]}{\sqrt{\mathrm{var}[\bar{M}]}}\right| \left(\bar{M} = \sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 V_i\right)$$



高次元漸近理論の開発
～大標本漸近理論から高次元漸近理論へ～

$$\lim_{n\to\infty} \quad \Rightarrow \quad \lim_{n\to\infty, p\to\infty, n/p\to c}$$

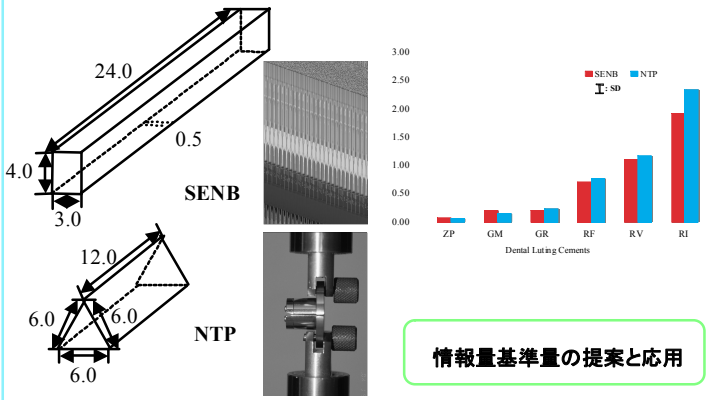
・主成分寄与率
・小さい主成分分散の同等性検定
・正準相関係数
} 等の高次元漸近分布を導出

$$\sigma_{m,n}^{-1}\{(\sum_{j=q+1}^{p}\log\ell_j - m\log\frac{1}{m}\sum_{j=q+1}^{p}\ell_j) - \mu_{m,n}\} \xrightarrow{d} N(0,1)$$

ℓ_j ： 標本共分散行列の j 番目に大きい固有値

$$\mu_{m,n} = m\log m - m\psi(\frac{mn}{2}) + \psi_m(\frac{n}{2})$$
$$\sigma_{m,n}^2 = \psi_m'(\frac{n}{2}) - m^2\psi'(\frac{mn}{2}), \ \psi_m a = \sum_{j=1}^{m}\psi(a - \frac{1}{2}(j-1))$$

情報量基準による関数関係モデルの選択問題
歯科用合着用セメントの2種類の強度測定法
SENB; NTP の同等性研究



情報量基準量の提案と応用

M1 : $\mu_{2j} = c\mu_{1j}, (c \neq 0) \qquad (j=1,\cdots,m)$

M2 : $\mu_{2j} = \alpha + \beta\mu_{1j}, (\alpha,\beta \neq 0) \qquad (j=1,\cdots,m)$

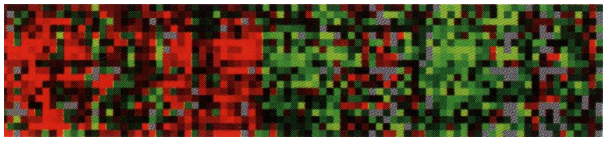
M3 : $\mu_{2j} \neq \mu_{1j} \qquad (j=1,\cdots,m)$ $\qquad (m \geq 2)$

M1 のもとでの推定方程式

$$\sum_{j=1}^{m}\frac{(\hat{c}\bar{x}_{1j\cdot} - \bar{x}_{2j\cdot})(\bar{x}_{1j\cdot} + \hat{c}\bar{x}_{2j\cdot})}{(1+\hat{c}^2)(s_{1j}^2 + s_{2j}^2) + (\hat{c}\bar{x}_{1j\cdot} - \bar{x}_{2j\cdot})^2} = 0$$

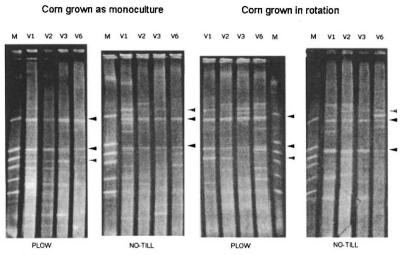
高次元多変量2値データの解析
～どの遺伝子で判別できるか～

フィンガープリントデータ



Terry, Speed: *Interdisciplinary Statistics Statistical Analysis of Gene Expression Microarray Data.* Chapman & Hall/CRC(2003)

4種類のとうもろこし：84個の遺伝子（変数）



どの遺伝子で判別できる？

変数選択法の開発

Wilbur, J.D., Ghosh,J.K., Nakatsu,C.H., Brouder, S.M., and Doerge, R.W. : Variable selection in high-dimensional multivariate binary data with application to the analysis of microbial community DNA fingerprints. ***Biometrics*, 58**, 378-386. (2002)

$$AIC = -n\log\frac{|W_{22\cdot1}|}{|T_{22\cdot1}|} + np(1+\log 2\pi)$$
$$+2\{qk + p - k + \frac{1}{2}p(p+1)\}$$