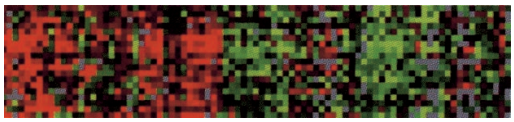
多変量高次元データ解析の理論と応用

研究代表者　杉山高一　研究員

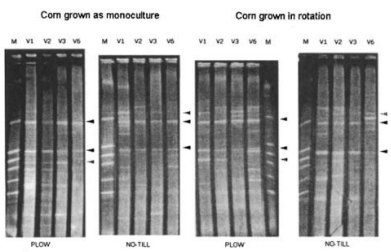
高次元における独立性の検定と頑健性
～各遺伝子は独立か～

フィンガープリントデータ



Terry, Speed: *Interdisciplinary Statistics Statistical Analysiss 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)



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)

DNA(デオキシリボ核酸)とは、
A(アデニン)、T(チミン)、G(グアニン)、C(シトシン)
の4種の塩基の配列

各DNAの独立性の検定

$$t_{n,p} = \frac{\sum_{i=2}^{p}\sum_{j=1}^{i-1} r_{ij}^2 - \frac{1}{2n}p(p-1)}{\sigma_{t_{n,p}}}$$

$t_{n,p}$の漸近正規の精度を解明

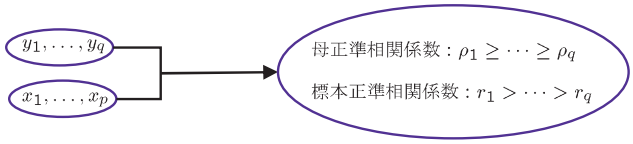
高次元漸近理論の開発
～大標本漸近理論から高次元漸近理論へ～

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

正準判別分析における
判別関数の重要度基準(固有値)

等の高次元漸近分布を導出

高次元Fisher-z変換の発見



A1 : q; fixed, $p \to \infty$, $n \to \infty$, $m = n - p \to \infty$, $c = p/n \to c_0 \in (0,1)$.

・高次元Fisher-z変換に基づく信頼区間

高次元枠組A1のもとでのρ_αの信頼区間は次のようになる.

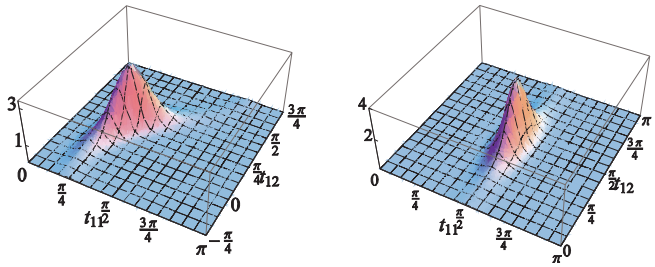
$$\sqrt{\frac{2-c}{2(1-c)}\left\{\tanh^2(u_1) - \frac{c}{2-c}\right\}} \leq \rho_\alpha \leq \sqrt{\frac{2-c}{2(1-c)}\left\{\tanh^2(u_2) - \frac{c}{2-c}\right\}}.$$

とくに, $c = 0$のとき, Fisher-z変換に基づく信頼区間と一致する

最大固有値に対応する固有ベクトルの挙動

3変量正規母集団からえられた標本共分散行列の最大固有値に対応する固有ベクトルの挙動を考察する. Sugiyama (1966) では固有ベクトルの極座標の同時確率密度関数を陽な形として与えている. 近年の計算機技術の発展によりその関数の計算が可能になった. 本研究では確率密度関数を数値計算可能な形で表現しなおし、グラフを描いた.

$$\boldsymbol{h}_1 = (\sin t_{11} \sin t_{12},\ \sin t_{11} \cos t_{12},\ \cos t_{11})'$$



t_{11}, t_{12}の同時密度関数のグラフ

CHUO UNIVERSITY