高次元小標本におけるナイーブ正準相関係数の漸近分布

同志社大学・文化情報学部　　玉谷　充
島根大学・総合理工学研究科　内藤　貫太

1. はじめに　Fan and Fan (2008, *The Annals of Statistics*, **36**, 2605-2637) は，高次元小標本における 2 群判別手法として naive Bayes を導入し，その誤判別確率の漸近上界を与えた．Tamatani, Koch and Naito (2012, *Journal of Multivariate Analysis*, **111**, 350-367) では，ナイーブ正準相関の観点から naive Bayes を導出し，ナイーブ正準相関係数の一致性について議論した．本講演では，誤判別確率の漸近上界がナイーブ正準相関係数に依存することから，その漸近分布を導出することを考える．

2. ナイーブ正準相関係数 λ　以下では d を次元とし，第 ℓ 群 ($\ell = 1, 2$) の母平均ベクトルを $\boldsymbol{\mu}_\ell$，母分散共分散行列 $\Sigma = [\sigma_{ij}]_{1 \le i,j \le d}$ は各群で等しいと仮定する．ナイーブ正準相関では，以下の量が重要となる：

$$C = D^{-1/2} E[(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{Y}^T] E[\boldsymbol{Y}\boldsymbol{Y}^T]^{-1/2}.$$

ここで，$\boldsymbol{X}$ は観測ベクトル，π_ℓ を第 ℓ 群の事前確率，$\boldsymbol{e}_\ell$ を第 ℓ 成分が 1 でその他の成分が 0 であるようなベクトルとするとき，$\boldsymbol{Y}$ は $P(\boldsymbol{Y} = \boldsymbol{e}_\ell) = \pi_\ell$ を満たす K 次元確率ベクトル，$\boldsymbol{\mu} = \sum_{i=1}^K \pi_i \boldsymbol{\mu}_i$，そして $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{dd})$ である．特に 2 群においては，ナイーブ正準相関は $\lambda = \pi_1 \pi_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ と陽に書けることに注意する．

3. ナイーブ正準相関係数 $\hat{\lambda}_{n,d}$ の漸近分布　n_ℓ を第 ℓ 群のデータ数とし，$n = n_1 + n_2$ とする．このとき，ナイーブ正準相関係数 $\hat{\lambda}_{n,d}$ の推定量は

$$\hat{\lambda}_{n,d} = \frac{n_1 n_2}{n^2} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^T \widehat{D}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$$

として与えることができる．ただし，$\widehat{\boldsymbol{\mu}}_\ell$ 及び $\widehat{D} = \mathrm{diag}(\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{dd})$ については，$\boldsymbol{\mu}_\ell$, D に対する最尤推定量である．このとき，$\hat{\lambda}_{n,d}$ の $n \ll d$ での漸近分布を求めるためにいくつかの統計量に分解し，漸近同等性のアプローチによって

$$\begin{aligned}
&\sqrt{\frac{n}{\lambda}} \left\{ \hat{\lambda}_{n,d} - \frac{n_1 n_2}{n^2} \frac{n-2}{n-4} \boldsymbol{\alpha}^T D^{-1} \boldsymbol{\alpha} - \frac{n_1 n_2}{n^2} \frac{nd}{n_1 n_2} \frac{n-2}{n-4} \right\} \\
&= \sqrt{\frac{n}{\lambda}} \frac{n_1 n_2}{n^2} \left\{ \left(\lambda^*_{n,d} - \boldsymbol{\alpha}^T D^{-1} \boldsymbol{\alpha} - \frac{nd}{n_1 n_2} \right) + \left(W_{n,d} - \frac{2}{n-4} \boldsymbol{\alpha}^T D^{-1} \boldsymbol{\alpha} \right) \right\} + O_P\left(\frac{1}{\sqrt{n}} \right)
\end{aligned}$$

と与えることができる．ただし，$\boldsymbol{\alpha} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ であり，

$$\lambda^*_{n,d} = (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^T D^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2), \quad W_{n,d} = \boldsymbol{\alpha}^T D^{-1} \left(\frac{n-2}{n-4} D - \widehat{D} \right) D^{-1} \boldsymbol{\alpha}$$

である．本講演では, Srivastava (2011, *Journal of Multivariate Analysis*, **102**, 1090-1103) を参考に，$\hat{\lambda}_{n,d}$ の漸近分布をマルチンゲール差分列に対する中心極限定理によって導出すると共に，シミュレーション結果について報告する．