高次元の場合での判別問題における誤判別確率の推定について

広島大学大学院　理学研究科　中川 智之

p 次元母集団 Π_k からのトレーニングデータ $\boldsymbol{X}_{k1}, ..., \boldsymbol{X}_{kN_j}$ から誤判別確率の推定を行う. 誤判別確率の推定方法はクロスバリデーション (CV) や漸近展開を用いた近似式など数多く提案されている. 漸近展開はより高次の近似まで可能であり, 計算負荷も少ない. しかし, 分布や判別手法に依るのでそれぞれで導出する必要があるので使うのが難しい. 一方, CV は分布や判別手法に依らず推定ができる手法であるので, 数多くのの場合で使われている. また, CV は大標本漸近理論:$N_j \to \infty$, p:固定の下では, $O(N_j^{-2})$ の漸近バイアスしか持たないことが分かっている. しかし, CV は計算負荷が大きく, p が大きい高次元データに対して, CV による推定の精度が良くない場合がある.

本報告では, 2 つの正規母集団 $\Pi_j : N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ の場合での線形判別関数:

$$L(\boldsymbol{X}) = L(\boldsymbol{X}; \bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2, \boldsymbol{S}) = \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right)' \boldsymbol{S}^{-1} \left\{\boldsymbol{X} - \left(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2\right)/2\right\},$$

による判別問題を考える. このとき, Π_k のトレーニングデータから i 番目の個体を抜いて作る推定量を $\bar{\boldsymbol{X}}_k^{(i)}$, $\boldsymbol{S}_k^{(i)}$ と表すとすると CV による推定は次のようになる.

$$\hat{P}_L^{CV}(2|1) = \frac{1}{N_1} \sum_{i=1}^{N_1} 1(L(\boldsymbol{X}_{1i}; \bar{\boldsymbol{X}}_1^{(i)}, \bar{\boldsymbol{X}}_2, \boldsymbol{S}^{(i)}) \leq c),$$

$$\hat{P}_L^{CV}(1|2) = \frac{1}{N_2} \sum_{i=1}^{N_1} 1(L(\boldsymbol{X}_{2i}; \bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2^{(i)}, \boldsymbol{S}^{(i)}) > c),$$

ただし, $1(\cdot)$:定義関数, c:カットオフポイントである.

高次元大標本漸近理論 $N_j \to \infty$, $p \to \infty$, $N = N_1 + N_2 > p$, の下で, CV による推定量の漸近バイアスや漸近分散, MSE の評価を行う. また, 次のような Π_k から 2 個抜いた CV

$$\hat{P}_L^{CV_2}(2|1) = \frac{1}{2} \sum_{l=1}^{2} \frac{2!(N_1-2)!}{N_1!} \sum_{i,j} 1(L(\boldsymbol{X}_{1l}; \bar{\boldsymbol{X}}_1^{(i,j)}, \bar{\boldsymbol{X}}_2, \boldsymbol{S}^{(i,j)}) \leq c),$$

$$\hat{P}_L^{CV_2}(2|1) = \frac{1}{2} \sum_{l=1}^{2} \frac{2!(N_2-2)!}{N_2!} \sum_{i,j} 1(L(\boldsymbol{X}_{2l}; \bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2^{(i,j)}, \boldsymbol{S}^{(i,j)}) > c),$$

を用いることで, CV のバイアスを補正した推定方法の提案を行う 最後に, 有限の場合での CV と漸近展開, 新しく提案した推定方法の推定の精度を数値実験で比較する.

参考文献

[1] Fujikoshi, Y. , Seo, T. (1998), Asymptotic approximations of EPMC's of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large, *Random Oper. Stochastic Equations*, **6**, 269-280.

[2] Lachenbruch, P. A. and Mickey, M. R. (1968), Estimation of Error Rates in Discriminant analysis. *Technometrics*, **10**, 1-11.

[3] Stone, M. (1974), Cross-Validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* , **B36**, 111-147.

[4] Tonda, T. and Wakaki, H. (2003), EPMC Estimation in Discriminant Analysis When the Dimension and Sample Sizes are Large. *Hiroshima Statistical Research Group Technical Report*, TR, 03-08.

[5] Yanagihara, H. and Fujisawa, H. (2012), Iterative bias correction of the cross-validation criterion. *Scand. J. Stat.* , Vol. 39, 116-130.