

変数選択を用いた高次元2次判別方式について

筑波大学・数理物質系 矢田 和善
筑波大学・数理物質系 青嶋 誠

1. はじめに 高次元データにおける判別分析法を考える．2群の共分散行列が等しいと仮定すれば，Dudoit et al. (2002, JASA) 等による共分散行列の対角成分だけを使った判別方式や Fan and Fan (2008, AS) 等による変数選択に基づく判別方式などがある．しかし，2群の共分散行列が等しいと仮定する問題設定の単純化は，高次元データが本来もつ2群の差異に関する情報に目を瞑ることになる．2群の共分散行列が等しいことを仮定しない2次判別分析の場合，Dudoit et al. (2002) による共分散行列の対角成分だけを使った判別方式 (Diagonal Quadratic Discriminant Analysis; DQDA), Aoshima and Yata (2014, AISM) 等のユークリッド距離に基づく判別方式，Aoshima and Yata (2011, SQA) による高次元小標本の幾何学的表現に基づく判別方式などがある．

本講演では，DQDA に着目し，DQDA と変数選択法を組み合わせることで，平均の差異だけでなく，共分散行列の対角成分の差異の情報までも含めた2次判別方式を与える．

2. 変数選択を用いた高次元2次判別方式 母集団が2個あると想定し，各母集団 (π_i) は平均に p 次のベクトル $\mu_i = (\mu_{i(1)}, \dots, \mu_{i(p)})^T$ ，共分散行列に p 次の正定値対称行列 Σ_i をもつと仮定する．さらに， Σ_i の j 番目の対角成分を $\sigma_{i(j)}$ とおく．各母集団 π_i から， n_i (≥ 2) 個のトレーニングデータ $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$ を無作為に抽出する．ただし， $\mathbf{x}_{ij} = (x_{i1j}, \dots, x_{ipj})^T$ とし， $n_{\min} = \min_{i=1,2} n_i$ とする．判別対象のデータを $\mathbf{x}_0$ とする．そのとき，次のDQDAタイプの判別関数を考える．

$$W(\mathbf{x}_0) = (\mathbf{x}_0 - \bar{\mathbf{x}}_1)^T \mathbf{S}_{1(d)}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1) - (\mathbf{x}_0 - \bar{\mathbf{x}}_2)^T \mathbf{S}_{2(d)}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_2) - \log \left\{ \frac{\det(\mathbf{S}_{2(d)})}{\det(\mathbf{S}_{1(d)})} \right\} - p/n_1 + p/n_2 \quad (1)$$

ここで， $-p/n_1 + p/n_2$ はバイアス補正項であり， $s_{it} = \sum_{j=1}^{n_i} (x_{itj} - \bar{x}_{it})^2 / (n_i - 1)$ ， $\bar{x}_{it} = \sum_{j=1}^{n_i} x_{itj} / n_i$ ， $\mathbf{S}_{i(d)} = \text{diag}(s_{i1}, \dots, s_{ip})$ である．判別方式は $W(\mathbf{x}_0)$ が負ならば $\mathbf{x}_0 \in \pi_1$ ，非負ならば $\mathbf{x}_0 \in \pi_2$ とする．ここで， $\Delta_i = \sum_{j=1}^p \{ (\mu_{1(j)} - \mu_{2(j)})^2 / \sigma_{i'(j)} + \sigma_{i(j)} / \sigma_{i'(j)} - 1 + \log(\sigma_{i'(j)} / \sigma_{i(j)}) \}$ ($i \neq i'$) とおき， $\Delta_{\min} = \min_{i=1,2} \Delta_i$ とおく． $\mathbf{x}_0 \in \pi_i$ を誤判別する確率を $e(i)$ と表記し，そのとき，Aoshima and Yata (2018, MCAP) は次の定理を与えた．

定理. 適当な正則条件と $p^2 \log p / (n_{\min} \Delta_{\min}^2) = o(1)$ のもとで， $p \rightarrow \infty$ のとき判別方式 (1) について各 $\mathbf{x}_0 \in \pi_i$ で $e(i) \rightarrow 0$ が成り立つ．

すなわち，DQDA による2群の差異 $\Delta_{\min}$ が “ $p^2 \log p / (n_{\min} \Delta_{\min}^2) = o(1)$ ” を満たすほど十分大きければ誤判別率に関する一貫性をもつ．しかしながら， $\Delta_{\min}$ が十分大きくない場合，高次元のもと “ $e(i) \rightarrow 1$ ” なる強不一致性をもつ判別方式 (1) の事例が Aoshima and Yata (2018) によって報告されている．そこで，平均の差異だけでなく共分散行列の対角成分の差異まで考慮した変数選択法を提案し，DQDA に適用することで新たなDQDAの判別方式を与える．

当日は， $\Delta_{\min}$ が小さい場合においても新しいDQDAが有効であることを，理論的かつ数値的に示す．さらに，実際のマイクロアレイデータ解析を交えて，高速かつ高精度に判別が可能であることを示す．