データ変換を用いた高次元2次判別方式について

筑波大学・数理物質系　矢田　和善
東京理科大学・情報科学科　石井　晶
筑波大学・数理物質系　青嶋　誠

高次元データに対する判別分析を考える．次元数が p の母集団が2個あると想定し，各母集団 π_i $(i = 1, 2)$ は共分散行列に p 次正定値対称行列 $\mathbf{\Sigma}_i$ をもつと仮定する．ここで，$\mathbf{\Sigma}_i$ の最大固有値を $\lambda_{i(\max)}$ とおく．

高次元データの2群判別は，2群の共分散行列が共通だと仮定すれば，Dudoit et al. (2002, JASA) や Bickel and Levina (2004, Bernoulli) による標本共分散行列の対角成分だけを使った判別方式がある．しかし，共分散行列の共通性を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を損なうことになる．共分散行列に共通性を仮定しない場合，Dudoit et al. (2002) による標本共分散行列の対角成分だけを使った判別方式，Aoshima and Yata (2014, AISM) のユークリッド距離に基づく判別方式（DBDA），Aoshima and Yata (2011, SA) による高次元データの幾何学的表現に基づく判別方式（GQDA）などがある．特に，Aoshima and Yata (2011,2014) では，Aoshima and Yata (2018) で提唱された，

$$\frac{\lambda_{i(\max)}^2}{\mathrm{tr}(\mathbf{\Sigma}_i^2)} \to 0 \ \text{ as } p \to \infty \text{ for } i = 1, 2$$

なる弱スパイク固有値 (NSSE) モデルのもと，DBDA と GQDA の高次元一致性と漸近正規性を示した．Aoshima and Yata (2018) では，

$$\liminf_{p\to\infty} \frac{\lambda_{i(\max)}^2}{\mathrm{tr}(\mathbf{\Sigma}_i^2)} > 0 \ \text{ for } i = 1, 2$$

なる強スパイク固有値 (SSE) モデルも提唱し，巨大なノイズを包含する SSE モデルのもとでは，統計的推測の精度保証が困難になることを示した．それに対して，Aoshima and Yata (2018, 2019)，Ishii et al. (2019, JJSD) では，SSE モデルから NSSE モデルへのデータ変換法を導入し，新たな高次元統計解析を展開している．特に，Aoshima and Yata (2019) では，高次元線形判別方式である DBDA にデータ変換を施し，SSE モデルのもと高次元一致性と漸近正規性を示した．

本講演では，GQDA 等の2次判別方式にデータ変換を施し，新たな高次元2次判別方式を提案する．数値実験と実データ解析も交えて，提案手法が SSE モデルのもと高精度に判別が可能であることを示す．

Aoshima, M. and Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models, *Statist. Sinica*, **28**, 43–62.

Aoshima, M. and Yata, K. (2019). Distance-based classifier by data transformation for high-dimension, strongly spiked eigenvalue models, *AISM*, **71**, 473–503.