High-dimensional inference on covariance structures via the extended cross-data-matrix methodology

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

ゲノム科学，情報工学，金融工学などの現代科学の 1 つの特徴は，データがもつ次元数の膨大さにある．例えば，次世代シークエンサによるゲノム配列データなど，次元数が数百万を超えるデータも解析の対象になる．それゆえ，膨大なデータを処理するためには，低い計算コストで高精度な解析結果を出力できるようなアルゴリズムが求められる．

最近, Yata and Aoshima (2013, JMVA) において, Yata and Aoshima (2010, JMVA) で提案したクロスデータ行列法 (CDM) を漸近最適な組み合わせに基づいて拡張した「拡張クロスデータ行列法 (ECDM)」を提案し，新しい推定量・検定統計量を構築した．本講演では，ECDM による共分散構造に関する推定量・検定統計量を考え，高次元のもと一致性と漸近正規性を有するための条件を与える．さらに，実際のマイクロアレイデータ解析に応用する．

2. 共分散構造に関する検定

いま，母集団に p 次元の分布を考え，n 個のデータ $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出したとする．ただし，p_1 次元のベクトル $\boldsymbol{x}_{2j}$ と p_2 $(= p - p_1)$ 次元のベクトル $\boldsymbol{x}_{2j}$ を用いて，$\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^T, \boldsymbol{x}_{2j}^T)^T$, $j = 1, ..., n$ と表記する．ここで，$\boldsymbol{\Sigma}_i = \mathrm{Var}(\boldsymbol{x}_{ij})$，$\boldsymbol{\Sigma}_* = \mathrm{Cov}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j})$ とおく．このとき，適当な $\boldsymbol{\Sigma}_*$ の候補 $\boldsymbol{\Sigma}_0$ を用いた

$$H_0 : \boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\Sigma}_* \neq \boldsymbol{\Sigma}_0$$

なる検定問題を高次元の枠組みで考える．例えば，$\boldsymbol{\Sigma}_0 = \boldsymbol{O}$ のとき，上記は無相関性の検定となる．$\Delta = \mathrm{tr}(\boldsymbol{\Sigma}_* \boldsymbol{\Sigma}_*^T)$ $(= ||\boldsymbol{\Sigma}_*||_F^2)$ とおき，ECDM を用いて構築した Δ の不偏推定量を $\widehat{T}_n$ とおく．そのとき，Yata and Aoshima (2015) は，それぞれ適当な正則条件のもとで，

$$\widehat{T}_n / \Delta = 1 + o_P(1), \quad p \to \infty, n \to \infty$$

なる一致性と，

$$\frac{\widehat{T}_n - \Delta}{\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)/n}} \Rightarrow N(0,1), \quad p \to \infty, n \to \infty$$

なる漸近正規性が成り立つことを示した．詳細は Yata and Aoshima (2015) を参照のこと．

当日は, $\widehat{T}_n$ による推定・検定が高次元小標本 $(n/p \to 0)$ の枠組みでも有効であることを，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，高速かつ高精度に共分散構造の推定・検定ができることを示す．

[1] Yata, K. and Aoshima, M. (2015). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology, submitted (arXiv:1503.06492).