高次元小標本における第１主成分の漸近的性質と平均ベクトルの検定

筑波大学・数理物質科学　石井　晶
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

情報化の進展に伴い，高次元データの統計的な解析が益々重要になってきている．データの次元数 p が標本数 n よりも遥かに大きな高次元小標本においては，従来の多変量解析の理論は崩壊する．Aoshima and Yata (2011) は，二標本問題の検出力を保証する検定方式や正判別確率を保証する判別方式など 8 つの統計的推測について，高次元データの母集団の差異を幾何学的表現で捉える先駆的な理論と方法論を与えた．

本講演では，高次元小標本の枠組みの中でも，次元数 $p \to \infty$ だが n は固定のもと，平均ベクトルの検定について議論する．平均に p 次のベクトル $\boldsymbol{\mu}$，共分散行列に p 次の非負定値対称行列 $\boldsymbol{\Sigma}$ をもつ母集団を考える．n 個の p 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出して，$p \times n$ データ行列 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ を定義する．ただし，$p > n$ である．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p (\geq 0)$ とし，$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ とおく．対応する直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_p]$ で $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$ と固有値分解する．標本共分散行列を $\boldsymbol{S} = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^T$ とおく．ここで，$\overline{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, ..., \bar{\boldsymbol{x}}]$，$\bar{\boldsymbol{x}} = \sum_{j=1}^{n} \boldsymbol{x}_j / n$ である．双対な標本共分散行列 $\boldsymbol{S}_D = (n-1)^{-1}(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})$ について，固有値 $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{n-1} (\geq 0)$ と対応する固有ベクトル $\hat{\boldsymbol{u}}_j$, $j = 1, ..., n-1$ によって，$\boldsymbol{S}_D = \sum_{j=1}^{n-1} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$ と固有値分解する．

Yata and Aoshima (2012) のノイズ掃き出し法を用いると，最大固有値の推定量は $\tilde{\lambda}_1 = \hat{\lambda}_1 - (n-2)^{-1}\{\mathrm{tr}(\boldsymbol{S}_D) - \hat{\lambda}_1\}$ で与えられる．Ishii et al. (2015) は，以下の 2 つの定理を与えた．

定理 1　$\boldsymbol{\Sigma}$ の最大固有値 λ_1 について，適当な正則条件のもと，$p \to \infty$，n 固定で以下が成り立つ．

$$(n-1)\frac{\tilde{\lambda}_1}{\lambda_1} \Rightarrow \chi^2_{n-1}$$

ここで，$\Rightarrow$ は分布収束，χ^2_{n-1} は自由度 $n-1$ の χ^2 分布に従う確率変数を表す．

平均ベクトルの検定 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs. $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ について，検定統計量を次のように定義する．

$$F_0 = \frac{n||\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0||^2 - \mathrm{tr}(\boldsymbol{S}_D)}{\tilde{\lambda}_1} + 1$$

定理 2　適当な正則条件のもと，$p \to \infty$，n 固定で以下が成り立つ．

$$F_0 \Rightarrow F_{1,n-1} \text{ under } H_0$$

ここで，$F_{1,n-1}$ は自由度 $(1, n-1)$ の F 分布に従う確率変数を表す．

定理 2 に基づいて，$\alpha \in (0, 1/2)$ に対して，検定方式を次のように与える．

$$F_0 \geq F_{1,n-1}(\alpha) \text{ ならば } H_0 \text{ を棄却}$$

ここで，$F_{1,n-1}(\alpha)$ は自由度 $(1, n-1)$ の F 分布における上側 α 点を表す．

当日は，二標本における平均ベクトルの検定についても述べ，数値実験や遺伝子発現データを用いた解析例も紹介する．

[1] Aoshima, M. and Yata, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Anal. (Editor's special invited paper)* 30, 356-399.

[2] Ishii, A., Yata, K. and Aoshima, M. (2015). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. Revised in *J. Stat. Plan. Inference*.

[3] Yata, K. and Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *J. Multivariate Anal.* 105, 193-215.