高次元空間における特徴ベクトルの一致推定について

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. ゲノム科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．こういった高次元データの第 1 の特徴は，次元数が標本数を遥かに超えることである．第 2 の特徴は，高次元データは豊富な情報を有するものの，それが巨大なノイズに埋もれて見つけ難いことである．これらの理由から，通常の多変量解析法では高次元データの推測に精度を保証することができず，間違った解析結果を導くことさえある．Yata and Aoshima (2009,CSTM) は，高次元データにおける PCA の性質を研究し，PCA が一致性をもつための標本数 n の d に関するオーダー条件を導き，高次元小標本において PCA が不適解を起こすことを示した．この問題を解決する策として，Yata and Aoshima (2012,JMA) は，高次元データ空間の幾何学的表現を研究し，それに基づいて “ノイズ掃き出し法” とよばれる方法論を考案した．最近，Aoshima and Yata (2017,SS) は，ノイズ掃き出し法による新たな高次元平均ベクトルの検定法を考案した．

本講演では，高次元空間における特徴ベクトルの一致推定について論じる．具体的には，高次元固有ベクトルや高次元平均ベクトルの一致推定を考える．

2. 高次元固有ベクトルの一致性. 共分散行列に d 次の半正定値行列 $\boldsymbol{\Sigma}$ をもつ母集団を考える．母集団から n (≥ 3) 個の d 次データベクトル $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ を無作為に抽出する．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_d]$ で $\boldsymbol{\Sigma}$ を $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ と分解する．標本共分散行列 $\boldsymbol{S}$ のスペクトル分解を $\boldsymbol{S} = \sum_{i=1}^{d} \hat{\lambda}_i \hat{\boldsymbol{h}}_i \hat{\boldsymbol{h}}_i^T$ とする．そのとき，Yata and Aoshima (2009) は，適当な正則条件のもと，

$$\hat{\lambda}_j/\lambda_j = 1 + \delta_j + o_p(1) \ \text{ and } \ \boldsymbol{h}_j^T \hat{\boldsymbol{h}}_j = (1+\delta_j)^{-1/2} + o_p(1) \quad \text{as } d, n \to \infty \tag{1}$$

が成り立つことを示した．ただし，$\delta_j = \lambda_j^{-1} \sum_{i=m+1}^{d} \lambda_i/(n-1)$，$m$ $(\geq j)$ はある定数である．一方で，Yata and Aoshima (2012) は，高次元データ空間の幾何学的表現を研究し，それに基づいて “ノイズ掃き出し法” とよばれる方法論を考案し，$\tilde{\lambda}_j = \hat{\lambda}_j - (\mathrm{tr}(\boldsymbol{S}) - \sum_{i=1}^{j} \hat{\lambda}_i)/(n-1-j)$ $(j = 1, ..., n-2)$ なる固有値の推定量を提案した．さらに，$\boldsymbol{\Sigma}$ の固有ベクトル $\boldsymbol{h}_j$ を $\tilde{\boldsymbol{h}}_j = (\hat{\lambda}_j/\tilde{\lambda}_j)^{1/2} \hat{\boldsymbol{h}}_j$ で推定する．そのとき，Yata and Aoshima (2012) は，適当な正則条件のもと，

$$\tilde{\lambda}_j/\lambda_j = 1 + o_p(1) \ \text{ and } \ \boldsymbol{h}_j^T \tilde{\boldsymbol{h}}_j = 1 + o_p(1) \quad \text{as } d, n \to \infty \tag{2}$$

が成り立つことを示した．それゆえ，$\boldsymbol{h}_j$ の内積に関する一致性をもつ．

ここで，$||\tilde{\boldsymbol{h}}_j||^2 = \hat{\lambda}_j/\tilde{\lambda}_j$ であることに注意し，(1) と (2) より，適当な正則条件のもと，

$$||\tilde{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = 2\{1 - (1+\delta_j)^{-1/2}\} + o_p(1) \text{ and } ||\tilde{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = \delta_j + o_p(1) \quad \text{as } d, n \to \infty$$

となる．すなわち，$\liminf_{d,n\to\infty} \delta_j > 0$ のとき，$\hat{\boldsymbol{h}}_j$ と $\tilde{\boldsymbol{h}}_j$ はノルムに関する一致性をもたない．

本講演では，閾値を用いて $\tilde{\boldsymbol{h}}_j$ を補正することで，$\delta_j \to \infty$ のもとでも，

$$||\acute{\boldsymbol{h}}_j - \boldsymbol{h}_j||^2 = o_p(1)$$

が成り立つような新たな固有ベクトルの推定量 $\acute{\boldsymbol{h}}_j$ を提案し，理論的かつ数値的に既存の推定量と比較する．さらに，上記のアイデアを高次元平均ベクトルの推測にも応用し，新たな高次元平均ベクトルの一致推定量も考案する．