高次元小標本における固有値の推定とその応用

矢田和善（筑波大学・数理物質科学研究科・院）
青嶋　誠（筑波大学・数理物質科学研究科）

1　はじめに

マイクロアレイデータやMRIデータに見られるように，情報化の進展に伴い，データの次元数 d が標本数 n よりも遥かに大きな高次元小標本 (HDLSS) データが，解析対象になる場面が増えてきている．このようなHDLSSデータに対して従来の統計手法を用いると，次元の呪いによって解析が上手くいかない．低次元空間への次元縮約の最も一般的な手法の一つに，主成分分析 (PCA) がある．しかしながら，HDLSSデータに対して従来型のPCAの使用には限界があることが，Muller et al. (2008), Jung and Marron (2008) 等によって報告されている．本発表では，HDLSSデータに対して，従来型の固有値や固有ベクトルの推定が有効になるための，標本数 n の d に関するオーダーを導いた．さらに，Yata and Aoshima (2008) のアプローチを用いて新しい固有値の推定法が提案され，これが，HDLSSデータに対してPCAの適用範囲を広げ，推定に必要な標本数 n のオーダーも緩めることを理論的に示した．この方法論は，固有ベクトルの推定にも応用された．

2　固有値と固有ベクトルの推定

平均が $\mathbf{0}$ の d 次元分布をもつ母集団から，n 個のデータベクトルを無作為に抽出して，データ行列 $\boldsymbol{X} : d \times n$ を定義する．ただし, $d > n$ と仮定する．母共分散行列 $\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, \cdots, \boldsymbol{h}_d]$ で $\boldsymbol{\Sigma} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ と分解する．ここで $\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{H}^T\boldsymbol{X}$ を定義する. $\boldsymbol{Z}$ の成分は，4次モーメントが一様有界になることを仮定する. いま, $\boldsymbol{\Sigma}$ の固有値に，Jung and Marron (2008) 等と同等な次のモデルを仮定する．

$$\lambda_i = a_i d^{\alpha_i} \ (i = 1, \cdots, m), \quad \lambda_j = c_j \ (j = m+1, \cdots, d).$$

ここで，a_i (> 0), c_j (≥ 0), α_i $(\alpha_1 \geq \cdots \geq \alpha_m > 0)$ は未知の実数，m は未知の整数とする．いま，$\alpha_i > 1/2$ なる最大の i を k_1 とし，$\lambda_1 > \cdots > \lambda_{k_1}$ と仮定する．標本共分散行列 $\boldsymbol{S} = n^{-1}\boldsymbol{X}\boldsymbol{X}^T$ と同じ固有値を有するDualな標本共分散行列を $\boldsymbol{S}_D = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ として，$\boldsymbol{S}_D$ の固有値を $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ とする．そのとき，$\hat{\lambda}_i$ $(i = 1, \cdots, k_1)$ について，以下の結果が導かれた．

定理 1. (i) $\alpha_i > 1$ ならば，$d \to \infty$, $n \to \infty$, (ii) $\alpha_i \in (1/2, 1]$ のとき, $d \to \infty$, $d^{1-\alpha_i}/n \to 0$ のもと，$\hat{\lambda}_i/\lambda_i = 1 + o_p(1)$.

HDLSS データに対して，$\boldsymbol{S}_D$ に基づく推定法では，$\alpha_i \leq 1/2$ の場合には固有値を推定できず，$\alpha_i > 1/2$ の場合であっても固有値の推定に多くの標本数が必要になる．Yata and Aoshima (2008) のアプローチを用いて，2 つの独立な $d \times n$ データ行列 $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ を使って，$\boldsymbol{S}^2 = n^{-2}\boldsymbol{X}_1\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{X}_2^T$ を定義する．（ここでの n は，$n' = n/2$ を意味する．）$\boldsymbol{S}^2$ と同じ固有値を有する Dual な 2 乗行列を $\boldsymbol{S}_D^2 = n^{-2}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{X}_2^T\boldsymbol{X}_1$ とし，$\boldsymbol{S}_D^2$ の固有値を $\tilde{\lambda}_1^2 \geq \cdots \geq \tilde{\lambda}_n^2$ とする．そのとき，以下の結果が導かれた．

定理 2. $\alpha_i > 1/4$ なる最大の i を k_2 とし，$\lambda_1 > \cdots > \lambda_{k_2}$ と仮定する．そのとき，$\tilde{\lambda}_i^2$ $(i = 1, \cdots, k_2)$ について，(i) $\alpha_i > 1/2$ ならば，$d \to \infty$, $n \to \infty$, (ii) $\alpha_i \in (1/4, 1/2]$ ならば，$d \to \infty$, $d^{2-4\alpha_i}/n \to 0$ のもと，$\sqrt{\tilde{\lambda}_i^2}/\lambda_i = 1 + o_p(1)$.

この推定量の良さは，シミュレーション実験において，推定量の分散も考慮して報告された．

本研究で得られたアプローチを用いれば，標本数 n を d の適当なオーダーで定めることで，HDLSS データに対して固有ベクトルを有効に推定することが可能である．いま，$\boldsymbol{S}$ に対して $\hat{\boldsymbol{H}}^T\boldsymbol{S}\hat{\boldsymbol{H}} = \hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_d)$ となる直交行列を，$\hat{\boldsymbol{H}} = [\hat{\boldsymbol{h}}_1, \cdots, \hat{\boldsymbol{h}}_d]$ とする．そのとき，$\hat{\boldsymbol{h}}_i$ $(i = 1, \cdots, k_1)$ について，以下の結果が導かれた．

定理 3. (i) $\alpha_i > 1$ ならば，$d \to \infty$, $n \to \infty$, (ii) $\alpha_i \in (1/2, 1]$ のとき, $d \to \infty$, $d^{1-\alpha_i}/n \to 0$ のもと，$\mathrm{Angle}(\hat{\boldsymbol{h}}_i, \boldsymbol{h}_i) \xrightarrow{p} 0$.

十分な大きさの標本が得られないとき，$\boldsymbol{S}_D$ に基づく固有値の推定量 $\hat{\lambda}_i$ は不安定になる．それに伴い，固有ベクトルの推定となる $\hat{\boldsymbol{h}}_i$ に誤差が生じ，主成分スコアにも誤差が生じることとなる．$\boldsymbol{S}_D^2$ に基づく固有値の推定量 $\sqrt{\tilde{\lambda}_i^2}$ を使えば，HDLSS データにおいて，より精度の高い固有値の推定が可能になる．本発表では，$\sqrt{\tilde{\lambda}_i^2}$ に基づくベイズ的なアプローチを使って，固有ベクトル $\hat{\boldsymbol{h}}_i$ を修正する方法も報告した．

Jung, S. and Marron, J. S. (2008). PCA consistency in high dimension, low sample size context. *Ann. Statist.*

Muller, K. E., Chi, Y.-Y., Ahn, J. and Marron, J. S. (2008). Limitations of high dimension, low sample size principal components for gaussian data. *J. Amer. Statist. Assoc.*

Yata, K. and Aoshima, M. (2008). Intrinsic dimensionality estimation of high dimension, low sample size data with *d*-asymptotics, submitted.