高次元小標本データにおける次元数の推定について

矢田 和善　（筑波大学大学院数理物質科学研究科）
青嶋　誠　（筑波大学大学院数理物質科学研究科）

1　はじめに

情報化の進展に伴い，データの次元数 p が標本数 n よりも大きな，高次元小標本データの解析法が必要になってきている．高次元データ解析を行う際，データは真には高次元でなく，むしろ高次元空間に埋め込まれていて，実際は，ずっと小さな次元をもった空間において要約できる，というコンセンサスがある．そこでは，出来るだけ情報を損なうことなく，低次元空間への次元縮約を行うべく，様々な方法論が提案されている．本報告は，次元縮約のための各種方法論を使う上で鍵となる，*Intrinsic Dimension* (ID) の推定を考えた．さらに，ID まで次元縮約した低次元空間における情報量を調べるために，寄与率の推定も考えた．

2　ID と寄与率の推定

母数が未知の p 次元正規分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_p)$ において，共分散行列 $\boldsymbol{\Sigma}_p$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_p > 0$ とする．そのとき，Ahn et al. (2007) に与えられる次のモデルを仮定する．ある未知の自然数 k $(< p)$ に対して，

$$\lambda_1 = \cdots = \lambda_k = ap^{\alpha}, \quad \lambda_{k+1} = \cdots = \lambda_p = c$$

ここで，a, c (> 0)，α $(> 1/2)$ は未知の実数である．次元数 p が大きな高次元データにおいて，最初の k 番目までの固有空間は潜在的なものと考えられ，残りの $p-k$ 個の固有空間はノイズがもたらしたものと解釈できる．そこで，自然数 k を ID と考える．なお，$k = \gamma p^r$ $(\gamma > 0,\ 0 \leq r < 1)$ とおいて，ID が p に伴って増加するモデルも考慮に入れる．いま，

$$k_p := \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)} = k + O(p^{1-2\alpha}), \quad p \to \infty$$

に注意する．大きさ n $(\geq 40,\ k = 1;\ n \geq 24,\ k = 2, 3;\ n \geq 16,\ k \geq 4)$ の i.i.d. 標本を 4 つに分け，各々の大きさが $n/4$ $(= n_\star)$ の標本で標本共分散行列 $\boldsymbol{S}_{ipn_\star},\ i = 1, ..., 4$ を計算する．そのとき，k の推定として

$$\hat{k}_{n_\star} = \frac{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star})\mathrm{tr}(\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})}{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star}\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})}$$

を考えると，$k/n \to 0,\ p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}(\hat{k}_{n_\star}) = k + o(k/n) + O(p^{1-2\alpha}), \quad E_{\boldsymbol{\theta}}\{(\hat{k}_{n_\star} - k_p)^2\} = O(k/n)$$

が主張でき，この結果は高次元小標本 $(n/p \to 0)$ においても保証される．

第 k 固有空間までの潜在的な効果とノイズの影響は，第 k 固有値までの寄与率 $\delta = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$ で評価される．いま，

$$\delta_p := \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^3)\mathrm{tr}(\boldsymbol{\Sigma}_p)} = \delta + O(p^{-\alpha}), \quad p \to \infty$$

に注意する．そのとき，δ の推定として

$$\hat{\delta}_{n_\star} = \frac{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star})\mathrm{tr}(\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})}{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star}\boldsymbol{S}_{3pn_\star})\mathrm{tr}(\boldsymbol{S}_{4pn_\star})}$$

を考えると，$n \to \infty,\ p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}(\hat{\delta}_{n_\star}) = \delta + O(1/nk) + O(p^{-\alpha}), \quad E_{\boldsymbol{\theta}}\{(\hat{\delta}_{n_\star} - \delta_p)^2\} = O(1/nk)$$

が主張でき，この結果は高次元小標本 $(n/p \to 0)$ においても保証される．

以上で与えた ID と寄与率の推定が，漸近的な評価を精確に保証するための標本数を，二段階推定で決定した．その際の ID と寄与率の推定は，$p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}(\hat{k}_{N_\star}) = k + o(1), \quad E_{\boldsymbol{\theta}}\{(\hat{k}_{N_\star} - k_p)^2\} = o(1),$$
$$E_{\boldsymbol{\theta}}(\hat{\delta}_{N_\star}) = \delta + o(k^{-2}) + O(p^{-\alpha}), \quad E_{\boldsymbol{\theta}}\{(\hat{\delta}_{N_\star} - \delta_p)^2\} = o(k^{-2})$$

が主張でき，高次元小標本で結果が保証される．また，全ての結果は

$$\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_p, \quad k = \gamma p^r\ (\gamma > 0,\ 0 \leq r < 1),$$
$$\lambda_i = a p^{\alpha} + b_i p^{\beta_i} + c_i\ (a,\ b_i,\ c_i > 0;\ \alpha > 1/2;\ \alpha - \beta_i > r/2),\ i = 1, ..., k,$$
$$\lambda_j = b_j p^{\beta_j} + c_j\ (b_j,\ c_j > 0;\ \alpha - \beta_j > 1/2),\ j = k+1, ..., p$$

なる一般化モデルにおいても成立することが示された．

3　非正規分布における考察

母平均は $\boldsymbol{\mu} = \mathbf{0}$ とする．一般化モデルにおいて，k を定数 $(r = 0)$ と仮定する．適当な大きさ n の i.i.d. 標本から標本共分散行列 $\boldsymbol{S}_{pn}$ を計算し，$\mathrm{tr}(\boldsymbol{S}_{pn}) = \sum_{i=1}^{p} \lambda_i W_{in}$ なる確率変数 W_{in} $(i = 1, ..., p)$ について，$E_{\boldsymbol{\theta}}(W_{in}^4) = 1 + o(1)$ $(n \to \infty),\ E_{\boldsymbol{\theta}}(W_{in}^{-4}) < \infty$ が満たされるとする．2 節と同様に $\boldsymbol{S}_{ipn_\star},\ i = 1, ..., 4$ を計算し，k の推定として

$$\tilde{k}_{n_\star} = \frac{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star})^2}{\mathrm{tr}(\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star}\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})}$$

を考える．そのとき，$n \to \infty,\ n/p \to 0$ で

$$E_{\boldsymbol{\theta}}(\tilde{k}_{n_\star}) = k + o(1)$$

が成り立つ．さらに，$E_{\boldsymbol{\theta}}(W_{in}^8) = 1 + o(1)$ $(n \to \infty),\ E_{\boldsymbol{\theta}}(W_{in}^{-8}) < \infty$ も満たされるなら，$n \to \infty,\ n/p \to 0$ で

$$E_{\boldsymbol{\theta}}\{(\tilde{k}_{n_\star} - k_p)^2\} = o(1)$$

が成り立つ．シミュレーションによる数値的な考察も与えられた．