数理解析研究所講究録
第1603巻 2008年 130-141

Note on Estimating the Intrinsic Dimension of High Dimension, Low Sample Size Data

筑波大学大学院・数理物質科学研究科　矢田　和善 (Kazuyoshi Yata)
Graduate School of Pure and Applied Sciences
University of Tsukuba

筑波大学・数学系　青嶋　誠 (Makoto Aoshima)
Institute of Mathematics
University of Tsukuba

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に，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，それよりもずっと小さな次元の空間に要約される，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．本論文は，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) の推定を考える．

IDの推定法として，２つのアプローチが知られている．一つはPCAに代表される射影に基づくもので，いま一つは，最短距離等で表される幾何学的な性質を利用するものである．前者では，Bruske and Sommer (1998) などに見られるように，固有値の大きさに依ってIDを推定することになるが，その方法論は探索的で、多分に経験に依るところが大きい．一方，後者には，Levina and Bickel (2005) に提唱されるMLEに基づいたIDの推定法があるが，推定が適切な性質を保証するためには相当に大きな標本が必要になり，本論文で扱うような高次元小標本データには，必ずしも適しているとは言い難い．

本論文では，Hall et al. (2005), Ahn et al. (2007) 等に与えられる高次元小標本モデルに基づいて，新しいIDの推定法を考える．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，寄与率の推定法も考える．推定量の平均と分散について漸近的な評価と，それを保証するための標本数に関する条件を与えて，高次元小標本データに対して，推定が適切な性質を有することを確認する．最後に，データに正規性が仮定できない場合についても考察し，推定法とその性質を保証するための条件を与える．なお，標本数の算出については，Yata and Aoshima (2008) の方法を応用することが考えられるが，詳細についてはここでは触れないことにする．

2　IDの推定

母数が未知の p 次元正規分布 $N_p(\mu, \Sigma_p)$ において，共分散行列 Σ_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，

$$\lambda_1 = \cdots = \lambda_k = ap^{\alpha}, \quad \lambda_{k+1} = \cdots = \lambda_p = c \tag{2.1}$$

(Johnstone (2001) も参照)．ここで，$a, c\ (>0)$, $\alpha\ (>1/2)$ は未知の実数とする．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，$k=\gamma p^r$ $(\gamma>0,\ 0\le r<1)$ 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ことにする．

2.1 共分散行列の幾何学的性質

共分散行列 $\mathbf{\Sigma}_p$ について，次が成り立つ．

$$\mathrm{tr}(\mathbf{\Sigma}_p)=akp^{\alpha}+c(p-k),\quad \mathrm{tr}(\mathbf{\Sigma}_p^2)=a^2kp^{2\alpha}+c^2(p-k),$$
$$\mathrm{tr}(\mathbf{\Sigma}_p^2)^2=a^4k^2p^{4\alpha}+c^4(p-k)^2+2a^2c^2kp^{2\alpha}(p-k),$$
$$\mathrm{tr}(\mathbf{\Sigma}_p^4)=a^4kp^{4\alpha}+c^4(p-k).$$

いま, $2\alpha>1$ なので

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

となり，p が大きい高次元データにおいて k_p と k は一致する．

注意 1　一般に，$t=1,2,\ldots$ に対して，$\alpha>1/t$ のもとで次が成り立つ．

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

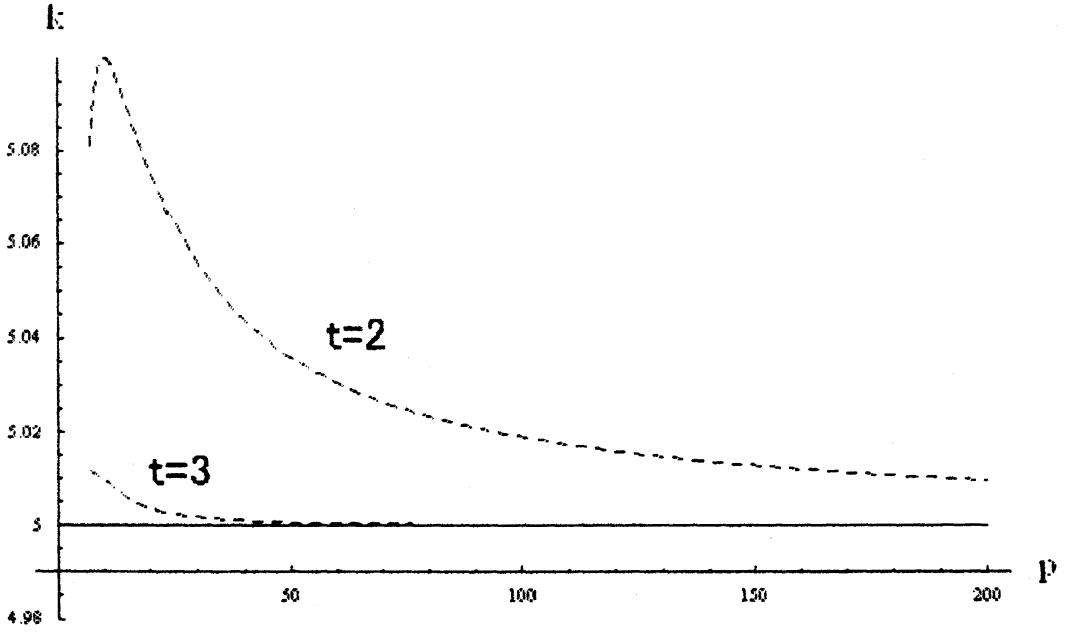

図 1　$k=5$, $\lambda_1=\cdots=\lambda_k=p$, $\lambda_{k+1}=\cdots=\lambda_p=1$ のときの $t=2,3$ に対する k_p の値

上式の左辺は，$t=1$ としたとき，球形度の尺度として知られるものとなり，正規分布における局所最強力不変検定に対する統計量を考えるときに使われる（John (1972) を参照のこと）．ID への収束は，t が大きくなると速くなる．しかしながら，$t \geq 3$ の場合，次節で構築する推定量の扱いは複雑になる．図 1 にも見られるように $t=2$ でも収束は十分に良いと考えられるので，本論文では $t=2$ と定めている．

2.2 推定量の構築

大きさ n (≥ 34) の i.i.d. 標本ベクトル $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ から標本共分散行列 $\boldsymbol{S}_{pn}$ を計算し，k_p の推定量として

$$\hat{k}_n = \frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)} \tag{2.2}$$

を考える．そのとき，次の定理が成り立つ．

定理 1 (2.1) において，$\alpha + r \geq 1$，かつ，$k^4/n \to 0$, $p \to \infty$ のとき，

$$E_{\boldsymbol{\theta}}(\hat{k}_n) = k + O(k/n^{1/2}) + O(p^{1-2\alpha}),$$
$$E_{\boldsymbol{\theta}}\{(\hat{k}_n - k_p)^2\} = O(k^2/n^{1/2}).$$

定理 1 の証明 まず，次の展開を考える．

$$E_{\boldsymbol{\theta}}(\hat{k}_n) = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)} + E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)}\right\} - E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2\left(\mathrm{tr}(\boldsymbol{S}_{pn}^4) - \mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right)}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right\},$$
$$E_{\boldsymbol{\theta}}(\hat{k}_n^2) = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2} + E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^4 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2}\right\} - E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4\left(\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2\right)}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2}\right\} \tag{2.3}$$

ここで，$\alpha + r \geq 1$，かつ，$k^4/n \to 0$, $p \to \infty$ のとき，補題 1 より

$$E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)}\right\} \leq \sqrt{E_{\boldsymbol{\theta}}\left\{\left(\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2\right)^2\right\} E_{\boldsymbol{\theta}}\left\{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^{-2}\right\}} = O(k/n^{1/2})$$

が主張できる．同様にして，補題 1 より

$$E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)}\right\} - E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2\left(\mathrm{tr}(\boldsymbol{S}_{pn}^4) - \mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right)}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right\} = O(k/n^{1/2}),$$
$$E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^4 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2}\right\} - E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4\left(\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2\right)}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^2\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2}\right\} = O(k^2/n^{1/2}) \tag{2.4}$$

が主張できる．(2.3) と (2.4) より，定理 1 が成り立つ． □

次に，大きさ n (≥ 8) の標本を4等分した各々の大きさが $n/4$ $(=n_\star)$ の i.i.d. 標本ベクトルを使って，標本共分散行列 $\boldsymbol{S}_{ipn_\star}$, $i=1,...,4$ を定義し，k_p の推定量として新たに

$$\hat{k}_{n_\star}=\frac{\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star})\mathrm{tr}(\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})}{|\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star}\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})|} \tag{2.5}$$

を考える．そのとき，次の定理が成り立つ．

定理2　(2.1) において，$k/n\to 0$, $p\to\infty$ のとき，

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

定理2の証明　まず，$\hat{U}_1=\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star})\mathrm{tr}(\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})$, $\hat{U}_2=\mathrm{tr}(\boldsymbol{S}_{1pn_\star}\boldsymbol{S}_{2pn_\star}\boldsymbol{S}_{3pn_\star}\boldsymbol{S}_{4pn_\star})$, $U_2=\mathrm{tr}(\boldsymbol{\Sigma}_p^4)$ とおく．次の展開を考える．

$$E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{\hat{U}_2}\right)=E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{U_2}\right)-E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1}{U_2^2}\left(\hat{U}_2-U_2\right)\right\}+\sum_{t=2}^{\infty}(-1)^tE_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1}{U_2^{t+1}}\left(\hat{U}_2-U_2\right)^t\right\},$$
$$E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1^2}{\hat{U}_2^2}\right)=E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1^2}{U_2^2}\right)-2E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1^2}{U_2^3}\left(\hat{U}_2-U_2\right)\right\}+\sum_{t=2}^{\infty}(-1)^t(t+1)E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1^2}{U_2^{t+2}}\left(\hat{U}_2-U_2\right)^t\right\}. \tag{2.6}$$

いま，$k/n\to 0$, $p\to\infty$ のとき，Appendix の (A.3) と (A.4) から

$$E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{U_2}\right)=k_p, \qquad E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1}{U_2^2}\left(\hat{U}_2-U_2\right)\right\}=O(1/n),$$
$$E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1^2}{U_2^2}\right)=k_p^2+O(k/n), \quad E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1^2}{U_2^3}\left(\hat{U}_2-U_2\right)\right\}=O(k/n) \tag{2.7}$$

が主張でき，さらに $t\geq 2$ について補題2から

$$E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1}{U_2^{t+1}}\left(\hat{U}_2-U_2\right)^t\right\}\leq\sqrt{E_{\boldsymbol{\theta}}\left(U_2^{-2}\hat{U}_1^2\right)E_{\boldsymbol{\theta}}\left\{U_2^{-2t}\left(\hat{U}_2-U_2\right)^{2t}\right\}}=O(n^{-t/2}k^{-t/2+1}),$$
$$E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1^2}{U_2^{t+2}}\left(\hat{U}_2-U_2\right)^t\right\}\leq\sqrt{E_{\boldsymbol{\theta}}\left(U_2^{-4}\hat{U}_1^4\right)E_{\boldsymbol{\theta}}\left\{U_2^{-2t}\left(\hat{U}_2-U_2\right)^{2t}\right\}}=O(n^{-t/2}k^{-t/2+2}) \tag{2.8}$$

が主張できることに注意する．また，$k/n\to 0$ のとき

$$\frac{1}{n}\sum_{t=2}^{\infty}(-1)^tn^{-t/2+1}k^{-t/2+1}=O(1/n), \quad \frac{k}{n}\sum_{t=2}^{\infty}(-1)^t(t+1)n^{-t/2+1}k^{-t/2+1}=O(k/n) \tag{2.9}$$

が主張できることにも注意する．そのとき，(2.8) と (2.9) から

$$\sum_{t=2}^{\infty}(-1)^t E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1}{U_2^{t+1}}\left(\hat{U}_2-U_2\right)^t\right\}=O(1/n)$$
$$\sum_{t=2}^{\infty}(-1)^t(t+1)E_{\boldsymbol{\theta}}\left\{\frac{\hat{U}_1^2}{U_2^{t+2}}\left(\hat{U}_2-U_2\right)^t\right\}=O(k/n) \tag{2.10}$$

が成り立つ．以上から，(2.7) と (2.10) を (2.6) に代入すれば

$$E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{\hat{U}_2}\right)=k_p+O(1/n),\quad E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1^2}{\hat{U}_2^2}\right)=E_{\boldsymbol{\theta}}(\hat{k}_{n_*}^2)=k_p^2+O(k/n) \tag{2.11}$$

が得られる．いま，モデル (2.1) における正数 c と indicator function $I_{\hat{U}_2<c}$ を用いて

$$E_{\boldsymbol{\theta}}(\hat{k}_{n_*})=E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{|\hat{U}_2|}\right)=E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1}{\hat{U}_2}\right)+E_{\boldsymbol{\theta}}\left(\hat{k}_{n_*}I_{\hat{U}_2<c}\right) \tag{2.12}$$

と表すと，補題 3 と (2.11) から

$$E_{\boldsymbol{\theta}}\left(\hat{k}_{n_*}I_{\hat{U}_2<c}\right)\le\sqrt{P_{\boldsymbol{\theta}}(\hat{U}_2<c)E_{\boldsymbol{\theta}}\left(\frac{\hat{U}_1^2}{\hat{U}_2^2}\right)}=O(1/n) \tag{2.13}$$

が得られる．それゆえ，(2.12) に (2.11) と (2.13) を考慮すれば

$$E_{\boldsymbol{\theta}}\left(\hat{k}_{n_*}\right)=k_p+O(1/n),\quad E_{\boldsymbol{\theta}}\{(\hat{k}_{n_*}-k_p)^2\}=E_{\boldsymbol{\theta}}(\hat{k}_{n_*}^2)-2k_pE_{\boldsymbol{\theta}}(\hat{k}_{n_*})+k_p^2=O(k/n)$$

が成り立つ． □

注意 2　定理 2 は $n/p\to 0$ の場合でも保証され，高次元小標本モデルに適用できる．下の図 2 は，$k=2([p^{1/3}]+1)$ と設定して，$\lambda_1=\cdots=\lambda_k=p^{2/3}$, $\lambda_{k+1}=\cdots=\lambda_p=1$ のときの k の推定値を，500 回のシミュレーションの平均をとってプロットしている．ここで，$[x]$ は x を越えない最大の整数である．(2.2), (2.5) 式ともに標本数は $n=2k([p^{1/5}/2]+1)$ と定めた．同じ標本数において，(2.5) 式の $\hat{k}_{n_*}$ の方が (2.2) 式の $\hat{k}_n$ よりも収束が速いことが分かる．

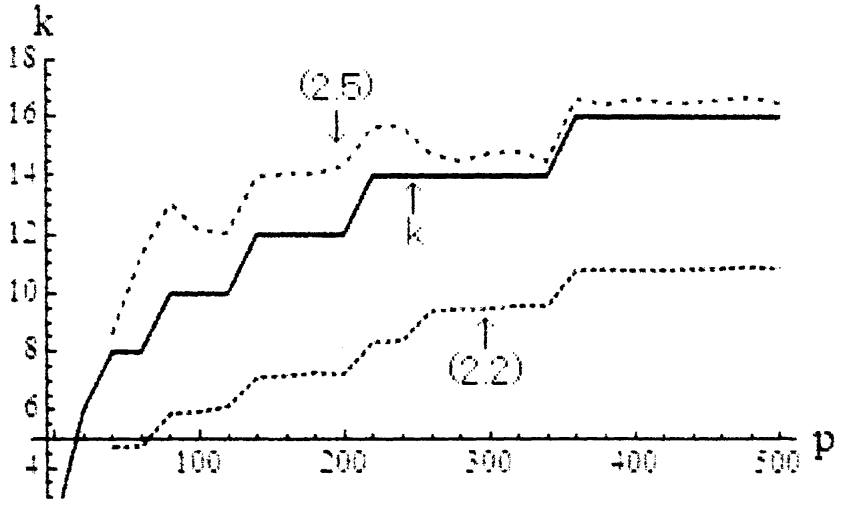

図 2　$k=2([p^{1/3}]+1)$ における k の推定値のシミュレーション

3 寄与率の推定

(2.1) のモデルにおいて，第 k 固有値までの寄与率 δ は

$$\delta = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = 1 - \frac{c(p-k)}{akp^{\alpha} + c(p-k)}$$

となる．寄与率の値は，$p \to \infty$ としたとき，(i) $\alpha + r > 1$, (ii) $\alpha + r = 1$, (iii) $\alpha + r < 1$ の 3 つの場合で，それぞれ次のように評価される．

$$\begin{aligned} &\text{(i)} \quad \delta = 1 + o(1), \\ &\text{(ii)} \quad \delta = 1 - \frac{c}{a\gamma + c} + o(1), \\ &\text{(iii)} \quad \delta = o(1) \end{aligned}$$

(i) の場合は，k 番目までの固有空間の潜在的な効果がノイズの影響よりも大きいモデルになる．(ii) の場合は，k 番目までの固有空間の潜在的な効果とノイズの影響とが同等になるモデルと解釈できる．(iii) の場合は，ノイズの影響が強く，k 番目までの固有空間の潜在的な効果が退化したモデルと解釈できる．特に (iii) の場合，寄与率に依る次元縮小法では，ID まで次元を削減することは困難になる．しかし，本論文で提案する方法論は，ノイズの影響が強い (iii) の場合においても，ID の推定量を構築することが可能なものになっている．

この節では，k 番目までの固有空間の潜在的な効果とノイズの影響を考察するために，寄与率の推定量を構築する．

3.1 寄与率の幾何学的性質

(2.1) のもと，共分散行列 $\mathbf{\Sigma}_p$ に関して，次が成り立つ．

$$\frac{\mathrm{tr}(\mathbf{\Sigma}_p^2)^2}{\mathrm{tr}(\mathbf{\Sigma}_p^3)} = \sum_{i=1}^{k} \lambda_i + \frac{2c^2(p-k)}{akp^{\alpha}} + \frac{c^4(p-k)^2}{a^3k^3p^{3\alpha}} + O(p^{1-2\alpha}), \quad p \to \infty.$$

このとき

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

となり，p が大きい高次元データにおいて δ_p と δ は一致する．

注意 3 一般に，$t = 1, 2, \ldots$ に対して，$\alpha > 1/(2t+1)$ のもとで次が成り立つ．

$$\frac{\mathrm{tr}(\mathbf{\Sigma}_p^{t+1})^2}{\mathrm{tr}(\mathbf{\Sigma}_p^{2t+1})\mathrm{tr}(\mathbf{\Sigma}_p)} = \delta + O(p^{-t\alpha}) + O(p^{1-(2t+1)\alpha}), \quad p \to \infty$$

従って，注意 1 と同様に，t が大きくなれば収束が速くなる．

3.2 推定量の構築

2.2節で計算した $\boldsymbol{S}_{pn}$ を使って，δ_p の推定量として

$$\hat{\delta}_n = \frac{\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2}{\mathrm{tr}(\boldsymbol{S}_{pn}^3)\mathrm{tr}(\boldsymbol{S}_{pn})} \tag{3.1}$$

を考える．そのとき，次の定理が成り立つ．

定理3 (2.1)において，$\alpha + r \geq 1$, かつ，$k/n \to 0$, $p \to \infty$ のとき，

$$E_{\boldsymbol{\theta}}(\hat{\delta}_n) = \delta + O(k/n) + O(p^{-\alpha}),$$
$$E_{\boldsymbol{\theta}}\{(\hat{\delta}_n - \delta_p)^2\} = O(k/n).$$

定理3の証明 定理1と同様な証明が出来る． □

一方，$\boldsymbol{S}_{ipn_*}$, $i = 1, ..., 4$ を使って δ_p の推定量

$$\hat{\delta}_{n_*} = \frac{\mathrm{tr}(\boldsymbol{S}_{1pn_*}\boldsymbol{S}_{2pn_*})\mathrm{tr}(\boldsymbol{S}_{3pn_*}\boldsymbol{S}_{4pn_*})}{|\mathrm{tr}(\boldsymbol{S}_{1pn_*}\boldsymbol{S}_{2pn_*}\boldsymbol{S}_{3pn_*})|\mathrm{tr}(\boldsymbol{S}_{4pn_*})} \tag{3.2}$$

を考えると，次の定理が成り立つ．

定理4 (2.1)において，$n \to \infty$, $p \to \infty$ のとき，

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

定理4の証明 定理2と同様な証明が出来る． □

注意4 定理3, 4は $n/p \to 0$ の場合でも保証され，高次元小標本モデルに適用できる．ただし，定理3は $\alpha + \gamma < 1$ の場合には適用できない．一方，定理4は，その場合にも適用可能である．下の図3は，$k = 12$ と設定して，$\lambda_1 = \cdots = \lambda_k = p^{2/3}$, $\lambda_{k+1} = \cdots = \lambda_p = 1$ のときの δ の推定値を，500回のシミュレーションの平均をとってプロットしている．(3.1), (3.2)式ともに標本数は $n = k([p^{1/4}] + 1)$ と定めた．同じ標本数において，(3.2)式の $\hat{\delta}_{n_*}$ の方が(3.1)式の $\hat{\delta}_n$ よりも収束が速いことが分かる．

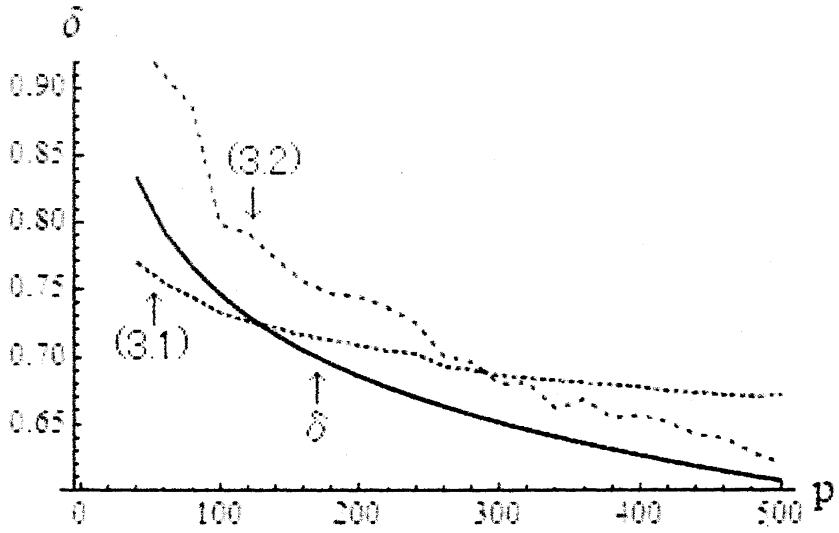

図3　$k=12$ における δ の推定値のシミュレーション

注意5　(2.1) のモデルを，次のように一般化する．

$$\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_p, \tag{3.3}$$
$$k = \gamma p^r \ (\gamma > 0,\ 0 \leq r < 1),$$
$$\lambda_i = ap^{\alpha} + b_i p^{\beta_i} + c_i \ (a,\ b_i,\ c_i > 0;\ \alpha > 1/2;\ \alpha - \beta_i > r/2),\ i = 1, ..., k,$$
$$\lambda_j = b_j p^{\beta_j} + c_j \ (b_j,\ c_j > 0;\ \alpha - \beta_j > 1/2),\ j = k+1, ..., p.$$

このモデルにおいても，定理1–4が成り立つ．

4　非正規分布における考察

データが非正規分布をもつ場合を考える．母平均ベクトルは $\boldsymbol{\mu} = \mathbf{0}$ とする．注意5で与えた (3.3) において，k を定数 $(r=0)$ としたモデルを仮定する．適当な大きさ n の i.i.d. 標本ベクトル $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ から，標本共分散行列 $\boldsymbol{S}_{pn} = n^{-1}\sum_{s=1}^{n} \boldsymbol{X}_s \boldsymbol{X}_s^T$ を計算する．

いま，

$$\mathrm{tr}(\boldsymbol{S}_{pn}) = \sum_{i=1}^{p} \lambda_i W_{in}$$

となる確率変数 W_{in} $(i = 1, ..., p)$ について，次の条件

$$E_{\boldsymbol{\theta}}(W_{in}^4) = 1 + o(1), \quad n \to \infty,$$
$$E_{\boldsymbol{\theta}}(W_{in}^{-4}) < \infty$$

が満たされるとする．2.2節と同様に標本共分散行列 $\boldsymbol{S}_{ipn_*}$，$i = 1, ..., 4$ を計算し，k_p の推定量として

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

を考える．そのとき，$\alpha \geq 1$, かつ, $n \to \infty$, $n/p \to 0$ で

$$E_{\boldsymbol{\theta}}(\tilde{k}_{n_*}) = k + o(1) \tag{4.1}$$

が成り立つ．さらに，次の条件

$$E_{\boldsymbol{\theta}}(W_{in}^{8}) = 1 + o(1), \quad n \to \infty,$$
$$E_{\boldsymbol{\theta}}(W_{in}^{-8}) < \infty$$

も満たされるなら，$\alpha \geq 1$, かつ, $n \to \infty$, $n/p \to 0$ で

$$E_{\boldsymbol{\theta}}\{(\tilde{k}_{n_*} - k_p)^2\} = o(1) \tag{4.2}$$

が成り立つ．

注意 6. 例えば，データの分布に p 次元指数分布，もしくは，p 次元ガンマ分布が仮定できるとき，(4.1) と (4.2) が主張できる．

Appendix

データの分布は $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_p)$ とし，2節で与えたモデル (2.1) を仮定する．このとき，適当な直交行列 $\boldsymbol{H}$ で，$\boldsymbol{H}^T \boldsymbol{S}_{pn} \boldsymbol{H} = (\sqrt{\lambda_i \lambda_j} W_{ij})$ と表せることに注意する．ここで，$(n-1)\boldsymbol{H}^T \boldsymbol{S}_{pn} \boldsymbol{H}$ はウィシャート分布 $W_p(n-1, \mathrm{diag}(\lambda_1, ..., \lambda_p))$ に従い，また，$(n-1)W_{ii}$, $i = 1, ..., p$ は互いに独立に自由度 $n-1$ の χ^2 分布に従い，W_{ij} $(i \neq j)$ は $E_{\boldsymbol{\theta}}(W_{ij}) = 0$, $E_{\boldsymbol{\theta}}(W_{ij}^2) = (n-1)^{-1}$, $E_{\boldsymbol{\theta}}(W_{ij}^{2t-1}) = O(n^{-t})$, $E_{\boldsymbol{\theta}}(W_{ij}^{2t}) = O(n^{-t})$ なる性質をもつ．

補題 1 $n \geq 34$, かつ, $k^4/n \to 0$, $p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}\left\{\frac{\left(tr(\boldsymbol{S}_{pn}^2)^2 - tr(\boldsymbol{\Sigma}_p^2)^2\right)^2}{tr(\boldsymbol{\Sigma}_p^4)^2}\right\} = O(k^2/n), \quad E_{\boldsymbol{\theta}}\left\{\frac{\left(tr(\boldsymbol{S}_{pn}^2)^4 - tr(\boldsymbol{\Sigma}_p^2)^4\right)^2}{tr(\boldsymbol{\Sigma}_p^4)^4}\right\} = O(k^4/n),$$
$$E_{\boldsymbol{\theta}}\left\{\frac{\left(tr(\boldsymbol{S}_{pn}^4) - tr(\boldsymbol{\Sigma}_p^4)\right)^2}{tr(\boldsymbol{\Sigma}_p^4)^2}\right\} = O(k^2/n), \quad E_{\boldsymbol{\theta}}\left\{\frac{tr(\boldsymbol{\Sigma}_p^4)^4}{tr(\boldsymbol{S}_{pn}^4)^4}\right\} < \infty.$$

証明 まず，$\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2$ と $\mathrm{tr}(\boldsymbol{S}_{pn}^4)$ が次のように表せることに注意する．

$$\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 = \left(\sum_{i=1}^{p} \lambda_i \sum_{j=1}^{p} \lambda_j W_{ij}^2\right)^2, \quad \mathrm{tr}(\boldsymbol{S}_{pn}^4) = \sum_{i=1}^{p} \lambda_i \sum_{j=1}^{p} \lambda_j \left(\sum_{l=1}^{p} \lambda_l W_{il} W_{jl}\right)^2.$$

ここで，$k^4/n \to 0$, $p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}\left\{\frac{\left(\mathrm{tr}(\boldsymbol{S}_{pn}^2)^2 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^2\right)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2}\right\} = O(k^2/n), \quad E_{\boldsymbol{\theta}}\left\{\frac{\left(\mathrm{tr}(\boldsymbol{S}_{pn}^2)^4 - \mathrm{tr}(\boldsymbol{\Sigma}_p^2)^4\right)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^4}\right\} = O(k^4/n),$$
$$E_{\boldsymbol{\theta}}\left\{\frac{\left(\mathrm{tr}(\boldsymbol{S}_{pn}^4) - \mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right)^2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^2}\right\} = O(k^2/n)$$

が得られる．また，$\mathrm{tr}(\boldsymbol{S}_{pn}^4) \geq \sum_{i=1}^k \lambda_i^4 W_{ii}^4$ なることと，さらに，$n \geq 34$ のとき

$$E_{\boldsymbol{\theta}}\left\{\frac{1}{(\sum_{i=1}^k W_{ii}^4)^4}\right\} \leq k^{-4}\left(\frac{n}{n-33}\right)^{16}$$

なることに注意すれば，$n \to \infty$, $p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}\left\{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^4}{\mathrm{tr}(\boldsymbol{S}_{pn}^4)^4}\right\} \leq \frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^4}{k^4\lambda_k^{16}}\left(\frac{n}{n-33}\right)^{16} < \infty$$

も得られる． □

補題 2　$k/n \to 0$, $p \to \infty$ のとき，$t \geq 2$ について

$$E_{\boldsymbol{\theta}}\left\{\left(\frac{tr(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn})tr(\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn})}{k\,tr(\boldsymbol{\Sigma}_p^4)}\right)^t\right\} = 1 + o(1), \tag{A.1}$$

$$E_{\boldsymbol{\theta}}\left\{\left(\frac{tr(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn}\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn}) - tr(\boldsymbol{\Sigma}_p^4)}{tr(\boldsymbol{\Sigma}_p^4)}\right)^t\right\} = O(1/(nk)^{t/2}). \tag{A.2}$$

証明　まず，次のように表せることに注意する．

$$\mathrm{tr}(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn})\mathrm{tr}(\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn}) = \left(\sum_{i=1}^p \lambda_i \sum_{j=1}^p \lambda_j W_{1ij}W_{2ij}\right)\left(\sum_{i=1}^p \lambda_i \sum_{j=1}^p \lambda_j W_{3ij}W_{4ij}\right), \tag{A.3}$$

$$\mathrm{tr}(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn}\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn}) = \sum_{i=1}^p \lambda_i \sum_{j=1}^p \lambda_j \left(\sum_{l=1}^p \lambda_l W_{1il}W_{2jl}\right)\left(\sum_{l=1}^p \lambda_l W_{3il}W_{4jl}\right). \tag{A.4}$$

$k/n \to 0$, $p \to \infty$ のとき

$$E_{\boldsymbol{\theta}}\left\{\left(\frac{\sum_{i=1}^p \lambda_i \sum_{j=1}^p \lambda_j W_{1ij}W_{2ij}}{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)}\right)^t\right\} = 1 + o(1) \quad (t \geq 2) \tag{A.5}$$

なることと，$p \to \infty$ のとき

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}_p^2)^{2t}}{k^t\mathrm{tr}(\boldsymbol{\Sigma}_p^4)^t} = 1 + o(1) \quad (t \geq 2) \tag{A.6}$$

なることに注意する．そのとき，(A.1) は (A.5) と (A.6) から得られる．一方，(A.2) には

$$\begin{aligned}
&\sum_{i=1}^{p}\lambda_i\sum_{j=1}^{p}\lambda_j\left(\sum_{l=1}^{p}\lambda_l W_{1il}W_{2jl}\right)\left(\sum_{l=1}^{p}\lambda_l W_{3il}W_{4jl}\right)-\mathrm{tr}(\boldsymbol{\Sigma}_p^4)\\
&=\sum_{i\neq j}\lambda_i\lambda_j\left(\sum_{l\neq i,\,l\neq j}\lambda_l W_{1il}W_{2jl}\right)\left(\sum_{l\neq i,\,l\neq j}\lambda_l W_{3il}W_{4jl}\right)\\
&\quad+\sum_{i\neq j}\lambda_i\lambda_j\left(\lambda_i W_{1ii}W_{2ji}+\lambda_j W_{1ij}W_{2jj}\right)\left(\lambda_i W_{3ii}W_{4ji}+\lambda_j W_{3ij}W_{4jj}\right)\\
&\quad+\sum_{i=1}^{p}\lambda_i^2\left(\sum_{l\neq i}\lambda_l W_{1il}W_{2il}\right)\left(\sum_{l\neq i}\lambda_l W_{3il}W_{4il}\right)+\sum_{i=1}^{p}\lambda_i^4(W_{1ii}W_{2ii}W_{3ii}W_{4ii}-1)\\
&(=T_1+T_2+T_3+T_4,\ \text{say}),
\end{aligned}$$

と表せることに注意する．$p\to\infty$ かつ $k/n\to 0$ のとき，$t\geq 2$ について

$$E_{\boldsymbol{\theta}}\left\{\left(\frac{T_1}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right)^t\right\}=O(k^t/n^2t),\qquad E_{\boldsymbol{\theta}}\left\{\left(\frac{T_2}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right)^t\right\}=O(1/n^t),$$

$$E_{\boldsymbol{\theta}}\left\{\left(\frac{T_3}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right)^t\right\}=O(1/n^{2t-1}),\qquad E_{\boldsymbol{\theta}}\left\{\left(\frac{T_4}{\mathrm{tr}(\boldsymbol{\Sigma}_p^4)}\right)^t\right\}=O(1/(nk)^{t/2}),$$

なので，(A.2) が得られる．□

補題 3　$k/n\to 0,\ p\to\infty$ のとき

$$P_{\boldsymbol{\theta}}(tr(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn}\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn})\leq c)=O(1/(nk)^2).$$

ここで，c はモデル *(2.1)* に与えられる正数とする．

証明　いま，$\hat{U}_2=\mathrm{tr}(\boldsymbol{S}_{1pn}\boldsymbol{S}_{2pn}\boldsymbol{S}_{3pn}\boldsymbol{S}_{4pn})$ とおく．そのとき，

$$\begin{aligned}
P_{\boldsymbol{\theta}}(\hat{U}_2\leq c)&=P_{\boldsymbol{\theta}}\left(\hat{U}_2-\mathrm{tr}(\boldsymbol{\Sigma}_p^4)\leq c-\mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right)\\
&\leq P_{\boldsymbol{\theta}}\left(\left|\hat{U}_2-\mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right|\geq \mathrm{tr}(\boldsymbol{\Sigma}_p^4)-c\right)\\
&\leq\left(\mathrm{tr}(\boldsymbol{\Sigma}_p^4)-c\right)^{-4}E_{\boldsymbol{\theta}}\left\{\left(\left|\hat{U}_2-\mathrm{tr}(\boldsymbol{\Sigma}_p^4)\right|^4\right)\right\}
\end{aligned}\tag{A.7}$$

となり，補題 2 から

$$E_{\boldsymbol{\theta}}\left\{\left(\left|\hat{U}_2/\mathrm{tr}(\boldsymbol{\Sigma}_p^4)-1\right|^4\right)\right\}=O(1/(nk)^2)\tag{A.8}$$

が成り立つので，(A.7) と (A.8) から結果を得る．□

参考文献

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, **94**, 760–766.

Bruske, J. and Sommer, G. (1998). Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. on PAMI*, **20**, 572–575.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc.*, **B 67**, 427–444.

John, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika*, **59**, 169–173.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, **29**, 295–327.

Levina, E. and Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in NIPS 17* (Eds. L. K. Saul, Y. Weiss, L. Bottou), Vancouver, Canada.

Yata, K. and Aoshima, M. (2008). Double shrink methodologies to determine the sample size via covariance structures. *J. Statist. Plan. Infer.*, in press.