High-Dimensional Mean Estimation via L1-Penalized Normal Likelihood

大阪大学大学院基礎工学研究科　片山 翔太[1]

1　はじめに

2 つの p 変量正規母集団 $N_p(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}^*)$ $(i = 1, 2), \boldsymbol{\Sigma}^* > \boldsymbol{O}$ からそれぞれ独立に n_i $(i = 1, 2)$ 個の独立サンプル $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ $(i = 1, 2; j = 1, \ldots, n_i)$ が得られたとする．本研究では，高次元データ (変数の次元 p が標本サイズ n_i と同程度または大きなデータ) における，2 つの母平均ベクトルの差 $\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ に関する統計的推測問題を扱う．高次元データにおける $\boldsymbol{\delta}^*$ の仮説検定問題 $H_0 : \boldsymbol{\delta}^* = \boldsymbol{0}$ v.s. $H_1 : \boldsymbol{\delta}^* \neq \boldsymbol{0}$ に関しては，Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010) などによって，高次元データに対しても検出力の高い検定方法がこれまでに様々提案されている．そこで本報告では，帰無仮説 H_0 が棄却された後の問題，すなわち，$\boldsymbol{\delta}^*$ のどの要素が 0 ではなくかつその値はどの程度かを推定する問題を考える．

近年，回帰分析モデルにおいて Lasso (Tibshirani, 1996) とよばれる，変数選択と回帰係数の推定を同時に行うことが可能な方法が提案されており，高次元データにおいて非常に良い性質を持っていることが示されている (例えば，Meinshausen and Yu, 2009)．この方法に基づくと，母共分散行列 $\boldsymbol{\Sigma}^*$ が既知ならば，$\boldsymbol{\delta}^*$ のひとつの推定量は次の目的関数を $\boldsymbol{\delta} \in \mathbb{R}^p$ に関して最小化することによって得られる:

$$Q(\boldsymbol{\delta}; \boldsymbol{\Omega}^*) = \boldsymbol{\delta}^T \boldsymbol{\Omega}^* \boldsymbol{\delta} - 2\boldsymbol{\delta}^T \boldsymbol{\Omega}^* \bar{\boldsymbol{\delta}} + 2\tau_{n,p} \|\boldsymbol{\delta}\|_1. \tag{1.1}$$

ここに，$\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{*-1}$, $\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ ($\bar{\boldsymbol{X}}_i = n_i^{-1} \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$ は標本平均) であり，$\|\cdot\|_1$ は ℓ_1 ノルム，$\tau_{n,p}$ はあるチューニングパラメータである．すなわち，$\tau_{n,p}$ が大きければ推定量は $\boldsymbol{0}$ に縮小され，$\tau_{n,p}$ が小さければ推定量は $\bar{\boldsymbol{\delta}}$ とほぼ同じ挙動を示す．しかしながら，ほとんどの実データ解析において $\boldsymbol{\Sigma}^*$ は未知である．そこで，母共分散行列 $\boldsymbol{\Sigma}^*$ を推定し (一時的にその推定量を $\widehat{\boldsymbol{\Sigma}}$ と記す)，$\widehat{\boldsymbol{\Sigma}}$ を $Q(\boldsymbol{\delta}; \boldsymbol{\Omega}^*)$ に Plug-In した目的関数 $Q(\boldsymbol{\delta}; \widehat{\boldsymbol{\Sigma}}^{-1})$ を考える．しかしながら，$\boldsymbol{\Sigma}^*$ の典型的な推定量である標本共分散行列

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^T, \quad n = n_1 + n_2,$$

は $n \geq p$ のとき逆行列が存在せず，$n < p$ であっても $n \simeq p$ のとき性能が著しく低いことが良く知られている (例えば，Johnstone, et al., 2001).

高次元データにおいて $\boldsymbol{S}$ の性能が低いことの一つの原因としては，$\boldsymbol{\Sigma}^*$ の 0 である非対角成分を推定する際に生じる誤差の蓄積が考えられる．高次元データにおいては，$\boldsymbol{\Sigma}^*$ の多くの非対角成分が 0 であると想定することは自然であるが，標本共分散行列 $\boldsymbol{S}$ の非対

[1]日本学術振興会特別研究員 (DC1)

角成分は確率 1 で非 0 である．各々の 0 からの誤差はごく微量であると考えられるが，それらが蓄積されれば無視できない量になってしまう．このような理由から，本研究では，標本共分散行列 $\boldsymbol{S}$ の代わりに近年 Cai and Liu (2011) で提案された $\boldsymbol{\Sigma}^*$ の推定量 $\widehat{\boldsymbol{\Sigma}}_T$ を用いる．詳細は次節で述べるが，この推定量は $\boldsymbol{S}$ の (i,j) 成分が 0 にある程度近ければ 0 と推定するものである．

記法：本研究で用いる記法をここに要約する．あるベクトル $\boldsymbol{a} = (a_i)$ に対して，ℓ_q ノルム $(\sum_i |a_i|^q)^{1/q}, 1 \leq q < \infty$ を $\|\boldsymbol{a}\|_q$ で表す．また，ベクトル $\boldsymbol{a}$ に対する符号関数を $\mathrm{sgn}(\boldsymbol{a}) = (\mathrm{sgn}(a_i))$ で定義する．ここに，$\mathrm{sgn}(a_i)$ は $a_i > 0$ のとき 1, $a_i = 0$ のとき 0, $a_i < 0$ のとき -1 を返す関数である．ある行列 $\boldsymbol{A} = (a_{ij})$ に対して，その最小固有値と最大固有値をそれぞれ $\lambda_{min}(\boldsymbol{A})$, $\lambda_{max}(\boldsymbol{A})$ で表す．また，行列 $\boldsymbol{A}$ の作用素ノルム $\lambda_{max}^{1/2}(\boldsymbol{A}^T\boldsymbol{A})$ を $\|\boldsymbol{A}\|$ で表し，無限大ノルム $\max_i \sum_j |a_{ij}|$ を $\|\boldsymbol{A}\|_\infty$ で表す．明らかに，対称行列 $\boldsymbol{A}$ に対しては $\|\boldsymbol{A}\| = |\lambda_{max}(\boldsymbol{A})|$ である．

2 準備

母平均ベクトル $\boldsymbol{\delta}^*$ の推定量を提案する前に，Cai and Liu (2011) で提案された母共分散行列 $\boldsymbol{\Sigma}^*$ の推定量 $\widehat{\boldsymbol{\Sigma}}_T$ について考察する．$\boldsymbol{X}_{ij} = (X_{ijk})$, $\bar{\boldsymbol{X}}_i = (\bar{X}_{ik})$, $Z_{ijk} = X_{ijk} - \bar{X}_{ik}$, $\boldsymbol{S} = (s_{ij})$ とすると，$\widehat{\boldsymbol{\Sigma}}_T$ は以下で与えられる：

$$\widehat{\boldsymbol{\Sigma}}_T = (\hat{\sigma}_{ij}), \quad \hat{\sigma}_{ij} = s_{ij} I\left(|s_{ij}| > 2\sqrt{\frac{\hat{\theta}_{ij} \log p}{n}} \right), \tag{2.1}$$

ここで，$I(\cdot)$ は指示関数を表し，

$$\hat{\theta}_{k\ell} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(Z_{ijk}Z_{ij\ell} - \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} Z_{ijk}Z_{ij\ell} \right)^2, \quad 1 \leq k, \ell \leq p$$

である．また，行列 $\boldsymbol{\Sigma}^* = (\sigma_{ij}^*)$ の Sparsity を

$$u_p = \max_{1 \leq i \leq p} \sum_{j=1}^{p} I(\sigma_{ij}^* \neq 0). \tag{2.2}$$

で定義する．もし u_p の値が小さければ，$\boldsymbol{\Sigma}^*$ の多くの非対角成分が 0 であることが分かる．いま，n, p, u_p の発散速度と $\boldsymbol{\Sigma}^*$ の対角成分 σ_{ii}^* に次の仮定を置く：

(A1) $p \to \infty$ as $n \to \infty$.

(A2) $\log p = o(n^{1/3})$ かつ，ある $0 < \gamma < 1$ および $\xi > 0$ に対して，$u_p = O\{(\log p)^\gamma\}$ $n^\xi \leq p$.

(A3) ある $K < \infty$ が存在して $\max_{1 \leq i \leq p} \sigma_{ii}^* \leq K$.

上記の仮定のもと，Cai and Liu (2011) では，$\widehat{\boldsymbol{\Sigma}}_T$ が $\boldsymbol{\Sigma}^*$ の非常に良い推定量であることを示している．

補題 2.1 (Cai and Liu (2011) の Lemma 4) 仮定 (A1)–(A3) のもと，任意の $\varepsilon > 0$ に対して, (γ, ε, K) のみに依存する定数 C_0 が存在して

$$P\left(\|\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}^*\| \leq C_0 u_p \sqrt{\frac{\log p}{n}}\right) \geq 1 - O(p^{-\varepsilon}). \tag{2.3}$$

を満たす．さらに，

$$E(\|\widehat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}^*\|^2) = O\left(u_p^2 \frac{\log p}{n}\right). \tag{2.4}$$

1 節で述べたように，$\boldsymbol{\Sigma}^*$ が未知であるときに，目的関数 (1.1) に基づいて $\boldsymbol{\delta}^*$ の推定を行うためには $\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{*-1}$ の推定量が必要である．しかしながら，$\widehat{\boldsymbol{\Sigma}}_T$ は常に正則であるとは限らない．そこで，$\widehat{\boldsymbol{\Sigma}}_T$ にある正則化項を加え，その逆行列を取ることで $\boldsymbol{\Omega}^*$ を推定する．すなわち，$\boldsymbol{\Omega}^*$ の推定量 $\widehat{\boldsymbol{\Omega}}_{T,a}$ を次のように構成する：

$$\widehat{\boldsymbol{\Omega}}_{T,a} = \widehat{\boldsymbol{\Sigma}}_{T,a}^{-1}, \quad \widehat{\boldsymbol{\Sigma}}_{T,a} = \widehat{\boldsymbol{\Sigma}}_T + a\boldsymbol{I}_p, \tag{2.5}$$

ここに，$a = |\lambda_{min}(\widehat{\boldsymbol{\Sigma}}_T)| + \{(\log p)/n\}^{1/2}$ であり，$\boldsymbol{I}_p$ は $p \times p$ 単位行列である．明らかに $\lambda_{min}(\widehat{\boldsymbol{\Sigma}}_{T,a}) > 0$ であるから $\widehat{\boldsymbol{\Sigma}}_{T,a}$ は正則であり，かつ $\widehat{\boldsymbol{\Sigma}}_{T,a} > \boldsymbol{O}$ である．

仮定 (A1)–(A3) に加えて

(A4) $u_p \left(\frac{\log p}{n}\right)^{1/2} = o(1)$ かつ，ある $\eta > 0$ が存在して $\lambda_{min}(\boldsymbol{\Sigma}^*) \geq \eta$.

を仮定すると，$\widehat{\boldsymbol{\Omega}}_{T,a}$ の収束レートは $\widehat{\boldsymbol{\Sigma}}_T$ と同じになる．すなわち，次の補題が成立つ．

補題 2.2 仮定 (A1)–(A4) のもと，任意の $\varepsilon > 0$ に対して, $(\gamma, \eta, \varepsilon, K)$ のみに依存する定数 C_1 が存在して

$$P\left(\|\widehat{\boldsymbol{\Omega}}_{T,a} - \boldsymbol{\Omega}^*\| \leq C_1 u_p \sqrt{\frac{\log p}{n}}\right) \geq 1 - O(p^{-\varepsilon}). \tag{2.6}$$

を満たす．さらに，

$$E(\|\widehat{\boldsymbol{\Omega}}_{T,a} - \boldsymbol{\Omega}^*\|^2) = O\left(u_p^2 \frac{\log p}{n}\right). \tag{2.7}$$

3　提案手法とその漸近的性質

本節では，$\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ の推定量を提案し，その漸近的な性質である Sign Recovery および平均二乗誤差 (MSE) を明らかにする．1 節で述べたように，$\boldsymbol{\delta}^*$ の推定量 $\hat{\boldsymbol{\delta}}$ を次のように提案する：

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta} \in \mathbb{R}^p}{\operatorname{argmin}} L(\boldsymbol{\delta}; \widehat{\boldsymbol{\Omega}}_{T,a}), \tag{3.1}$$

ここに，$L(\boldsymbol{\delta};\widehat{\boldsymbol{\Omega}}_{T,a})$ は式 (1.1) で与えられる．

記述を簡単にするため，$\widehat{\boldsymbol{\Sigma}}_{T,a}$, $\widehat{\boldsymbol{\Omega}}_{T,a}$ をそれぞれ $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\Omega}}$ と書くことにする．母平均ベクトルが $\boldsymbol{\delta}_1^* \neq \mathbf{0}$ $(elementwize) \in \mathbb{R}^s$ および $\boldsymbol{\delta}_2^* = \mathbf{0} \in \mathbb{R}^{p-s}$ を用いて $\boldsymbol{\delta}^* = (\delta_1^*, \ldots, \delta_p^*)^T = (\boldsymbol{\delta}_1^{*T}, \boldsymbol{\delta}_2^{*T})^T$ のように分割されるとし，その分割に対応して $\boldsymbol{\Sigma}^*$ を分割する：

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^* & \boldsymbol{\Sigma}_{12}^* \\ \boldsymbol{\Sigma}_{21}^* & \boldsymbol{\Sigma}_{22}^* \end{pmatrix}, \quad \boldsymbol{\Sigma}_{11}^* \in \mathbb{R}^{s\times s}, \quad \boldsymbol{\Sigma}_{22}^* \in \mathbb{R}^{(p-s)\times(p-s)}.$$

また，同様にして $\widehat{\boldsymbol{\Sigma}} = (\widehat{\boldsymbol{\Sigma}}_{ij})$ のように分割する．各 $i,j = 1,2$ に対して，$r_{ij} = \|\boldsymbol{\Sigma}_{ij}^*\|_\infty$, $\tilde{r}_{ii} = \|\boldsymbol{\Sigma}_{ii}^{*-1}\|_\infty$ と置き，u_{ii} を式 (2.2) で定義される $\boldsymbol{\Sigma}_{ii}^*$ の Sparsity とする．また，各 $i = 1,2$ に対して

$$d_{n,p} = u_p\sqrt{\frac{\log p}{n}}, \quad d_{ii} = u_{ii}\sqrt{\frac{\log p}{n}}. \tag{3.2}$$

と置く．このとき，仮定

(B1) ある $0 < k_0 \leq 1$ が存在して $\|\boldsymbol{\Sigma}_{22}^{*-1}\boldsymbol{\Sigma}_{21}^*\|_\infty \leq 1 - k_0$,

(B2) $\tilde{r}_{22}d_{n,p} = o(1)$,, $r_{21}\tilde{r}_{22}^2 d_{22} = o(1)$, $r_{12}\tilde{r}_{22}^2 d_{22} = o(1)$,

(B3) $r_{12}\tilde{r}_{22}d_{n,p} = o(1)$, $r_{12}r_{21}\tilde{r}_{22}^2 d_{22} = o(1)$,

のもとで推定量 $\hat{\boldsymbol{\delta}}$ の Sign Recovery が示される．

定理 3.1 仮定 (A1)–(A3), (B1)–(B3) のもとで，さらに，ある $\alpha \geq 4\sqrt{2}\max(\sigma_2, \tilde{\sigma}_2)/k_0$, $\beta \geq \sqrt{2}\max(\sigma_2, \sigma_{11.2})$ に対して

$$\tau_{n,p} = \alpha\sqrt{\frac{\log p}{N}}, \quad \frac{1}{2}\min_{1\leq i\leq s}|\delta_i^*| > \tau_{n,p}\tilde{r}_{11} + \beta\sqrt{\frac{\log p}{N}} \tag{3.3}$$

を満たすならば,

$$\begin{aligned} &P\big(\mathrm{sgn}(\hat{\boldsymbol{\delta}}) = \mathrm{sgn}(\boldsymbol{\delta}^*)\big) \\ &\quad \geq 1 - O\left\{(\log p)^{-1/2}p^{1-\frac{k_0^2\alpha^2}{32\max(\sigma_2^2,\tilde{\sigma}_2^2)}}\right\} - O\left\{(\log p)^{-1/2}p^{1-\frac{\beta^2}{2\max(\sigma_2^2,\sigma_{11.2}^2)}}\right\} \\ &\quad \to 1. \end{aligned} \tag{3.4}$$

が成立つ．ここに，$\sigma_2^2 = \max_{s+1\leq i\leq p}\sigma_{ii}^*$, $\tilde{\sigma}_2^2 = \max_{s+1\leq i\leq p}\psi_{ii}^*$, $\sigma_{11.2}^2 = \max_{1\leq i\leq s}\rho_{ii}^*$ であり，$\boldsymbol{\Sigma}^* = (\sigma_{ij}^*)$, $\boldsymbol{\Sigma}_{22}^{*-1} = (\psi_{ij}^*)$, $\boldsymbol{\Sigma}_{11.2}^* = (\rho_{ij}^*)$ である．

この定理は，非常に高い確率で推定量 $\hat{\boldsymbol{\delta}}$ の符号が母平均ベクトル $\boldsymbol{\delta}^*$ の符号に一致することを意味している．すなわち，$\hat{\boldsymbol{\delta}}$ は漸近的に $\boldsymbol{\delta}^*$ の 0 要素を正確に推定でき，なおかつ非 0 要素の符号も正確に推定できることがわかる．

一方，$\hat{\boldsymbol{\delta}}$ の平均二乗誤差 $E(\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2^2)$ を導くためには，さらに次の仮定が要求される：

(C1) $n_1/n \to c \in (0,1)$ as $n \to \infty$

(C2) $\lambda_{max}(\mathbf{\Sigma}^*) < \infty$, $\max_{s+1 \le i \le p} \omega_{ii}^* < \infty$,

(C3) $\min(\frac{k_0^2 \alpha^2}{32 \max(\sigma_2^2, \tilde{\sigma}_2^2)}, \frac{\beta^2}{2 \max(\sigma_2^2, \sigma_{11.2}^2)}) \ge 3 - \xi^{-1}$

(C4) $\max(1, r_{12} r_{21}) d_{n,p}^2 \frac{p-s}{n} = o(1)$, $t_{n,p} d_{n,p}^4 = o(1)$,

(C5) $\max(1, r_{12} r_{21}) t_{n,p} = O(1)$.

ただし，$\mathbf{\Omega}^* = (\omega_{ij})$, $t_{n,p} = s\{(\log p)/n\}^{1/2}$ である．

定理 3.2 定理 3.1 と同じ条件を仮定し，さらに (C1)–(C5) を仮定すると，

$$E\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}^*\|_2^2 = \frac{1}{N}\left\{\mathrm{tr}\mathbf{\Sigma}_{11.2}^* + \alpha(\log p)\|\mathbf{\Sigma}_{11.2}^* \,\mathrm{sgn}(\boldsymbol{\delta}^*)\|_2^2\right\} + o(1)$$

が成立つ．ここに，$\mathbf{\Sigma}_{11.2}^* = \mathbf{\Sigma}_{11}^* - \mathbf{\Sigma}_{12}^* \mathbf{\Sigma}_{22}^{*-1} \mathbf{\Sigma}_{21}^*$ であり，α は式 (3.3) で与えられる．

他の関連手法との MSE による比較および数値実験による比較は当日報告する．

参考文献

[1] Bai, Z and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6: 311–329.

[2] Cai, T and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Society*. 106: 672–684.

[3] Chen, S. X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*. 38: 808–835.

[4] Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*. 29: 295–327.

[5] Meinshausen, N., and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*. 37: 246–270.

[6] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Jounal of Multivariate Analysis*. 99: 386–402.

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*. 58: 267–288.