Stein's unbiased risk estimate and adaptive singular value shrinkage for estimatin problem of low-rank matrix mean with unknown covariance matrix

Yoshihiko KONNO

Osaka Metropolitan Univeristy

Conference
Mathematical Methods of Modern Statistics 3

27 June- 1st July 2022

1 Mean matrxi estimation when covariance is known

2 Sufficient conditions for Stein's identity

3 Case when the covarianc matrix is unknown

4 Simulation Study

Mean matrxi estimation when covariance is known

- Assume that m, p are postive integers s.t. $m \geq p$.
- Let

$$X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_m^T \end{pmatrix}$$

be an $m \times p$ data matrix whose row vectors are independently distributed as

$$x_i : \ p \times 1 \ \sim N(\xi_i, \sigma^2 I_p), \qquad (i = 1, 2, \ldots, m)$$

Here $\Xi^\top := (\xi_1, \ldots, \xi_m)$ and $\sigma > 0$ are unknown.

- First consider the problem of estimating Ξ under a loss fucntion and its risk

$$\mathbf{L_1}(\widehat{\Xi}, \Xi) = \mathbf{tr}\,\{(\widehat{\Xi} - \Xi)(\widehat{\Xi} - \Xi)^{\mathsf{T}}\} =: \|\widehat{\Xi} - \Xi\|_F^2$$

and

$$\mathbf{R_1}(\widehat{\Xi}, \Xi) = \mathbb{E}[\mathbf{L_1}(\widehat{\Xi}, \Xi)].$$

- Here $\widehat{\Xi}$ is an estimator based on $\boldsymbol{X}$.
- $\|\boldsymbol{A}\|_F$ is the Frobenius norm of a matrix $\boldsymbol{A}$.
- $\mathbf{tr}\,\boldsymbol{A}$ and $\boldsymbol{A}^{\mathsf{T}}$ stand for the traace and the transpose of a matrix $\boldsymbol{A}$, respectively.
- Low-rank mean matrix condition, i.e.,

$$\mathbf{rank}\,\Xi = \boldsymbol{r} < \boldsymbol{p}; \qquad \boldsymbol{r} \text{ is unknown.}$$

Eckart-Young approximation theorem

- Singular Value Decomposition: Decompose $\boldsymbol{X}$ as

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^{\top}; \quad \boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p),\ \boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p) \\ \boldsymbol{L} &= \mathbf{diag}(\ell_1, \ell_2, \ldots, \ell_p) \quad \text{with } \ell_1 \geq \ell_2 \geq \cdots \geq \ell_p \geq 0 \end{aligned}$$

where $\boldsymbol{u}_i \in \mathbb{R}^m$, $\boldsymbol{v}_i \in \mathbb{R}^p$ $(i = 1, \ldots, p)$ s.t.
$\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{I}_p$.

- The total least squres (TLS) pseudo estimator is given by

$$\widehat{\Xi}_{TLS} = \sum_{i=1}^{r} \ell_i \boldsymbol{u}_i \boldsymbol{v}_i^{\top}. \quad \Longleftrightarrow \quad \widehat{\Xi}_{TLS} \in \underset{\Xi:\ \mathbf{rank}\ \Xi \leq r}{\mathbf{argmin}} \|\Xi - \boldsymbol{X}\|_F^2.$$

A hard-shreshholding rule

- Assume that σ^2 is known.
- Solve

$$\mathbf{SVHT}_\lambda(\boldsymbol{X}) = \underset{\Xi}{\mathbf{argmin}}\left[\|\Xi - \boldsymbol{X}\|_F^2 + \lambda\,\mathbf{rank}\,(\Xi)\right]$$

  where $\lambda > 0$ is a tuning scalar parameter.
- Then the solution is given by

$$\mathbf{SVHT}_\lambda(\boldsymbol{X}) = \sum_{i=1}^{p} \ell_i\, \mathbb{1}\{\ell_i \geq \lambda\} u_i v_i^\top;\ \mathbb{1}\{\ell_i \geq \lambda\} = \begin{cases} 1 & \ell_i \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

- The optimal shreshholding is $\frac{4}{\sqrt{3}}\sqrt{p}\sigma$ when $p = m$.

A soft-thersholding rule

- Cèndes et al. define an adaptive soft-shreshholding rule based on SURE:

$$\mathbf{SVT}_\lambda(\boldsymbol{X}) = \sum_{i=1}^{p}(\ell_i - \lambda)\,\mathbb{1}\{\ell_i \geq \lambda\}\boldsymbol{u}_i\boldsymbol{v}_i^\top =: \sum_{i=1}^{p}(\ell_i - \lambda)_+\boldsymbol{u}_i\boldsymbol{v}_i^\top \tag{1}$$

  which is obtained from

$$\min_{\boldsymbol{Y}}\left\{\|\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + (2\lambda)\sum_{j=1}^{p}\lambda_j\right\} \qquad \boldsymbol{Y} = \mathbf{SVT}_\lambda(\boldsymbol{X}).$$

- The parameter λ in (**1**) is selected by minimizieng SURE, Stein's unbiased risk estimate for (**1**).

- Gaussian integration-by-parts (=Stein's identity) and a bit of algebraic calculation lead to

$$\begin{aligned}
\mathbf{R}_{\boldsymbol{B}}(\mathbf{SVT}_\lambda, \Xi) &= \mathbb{E}[\mathbf{SURE}(\mathbf{SVT}_\lambda)(\boldsymbol{X})], \\
\mathbf{SURE}(\mathbf{SVT}_\lambda)(\boldsymbol{X}) &= -mp\sigma^2 + \sum_{i=1}^{p} \min\{\ell_i^2, \lambda^2\} \\
&\quad + 2\sigma^2 \mathbf{div}(\mathbf{SVT}_\lambda(\boldsymbol{X})), \\
\mathbf{div}(\mathbf{SVT}_\lambda(\boldsymbol{X})) &= (m-p) \sum_{i=1}^{p} \left(1 - \frac{\lambda}{\ell_i}\right)_+ + \sum_{i=1}^{p} \mathbb{1}\{\ell_i > \lambda\} \\
&\quad + 2 \sum_{i=1}^{p} \sum_{j \neq i}^{p} \frac{\ell_i (\ell_i - \lambda)_+}{\ell_i^2 - \ell_j^2}
\end{aligned}$$

whenever $\ell_1 > \ell_2 > \cdots > \ell_p \geq 0$.

- An adaptive estimator is given by

$$\mathbf{SVT}_{\widehat{\lambda}}(\boldsymbol{X}) = \sum_{i=1}^{p} (\ell_i - \widehat{\lambda})_+ \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}, \tag{2}$$

$$\widehat{\lambda} = \underset{\lambda \geq 0}{\mathbf{argmin}} \left[\sum_{i=1}^{p} \min\{\ell_i^2, \lambda^2\} + 2\sigma^2 \mathbf{div}(\mathbf{SVT}_{\lambda}(\boldsymbol{X})) \right].$$

- Numerical evaluation of the risk of (**2**) was carried out by Candés et. al. Obviously

$$\mathbf{R}(\mathbf{SVT}_{\widehat{\lambda}}(\boldsymbol{X}), \Xi) < \mathbf{R}(\boldsymbol{X}, \Xi) \quad \text{for } \forall \Xi.$$

- But it is not clear if $\mathbf{R}(\mathbf{SVT}_{\widehat{\lambda}}(\boldsymbol{X}), \Xi)$ is close to $\mathbf{R}(\widehat{\Xi}_{\boldsymbol{TLS}}(\boldsymbol{X}), \Xi)$ for $\forall \Xi$ s.t. $\mathbf{rank}\, \Xi \leq r < p$.

Sufficient conditions for Stein's identity

- For $s \geq 0$ and $A \subset \mathbb{R}^p$, let $\mathcal{H}^s$ be the Haussdorff measure, i.e.,

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \left[\inf \left\{ \sum_{j=1}^{\infty} \alpha(s) \left(\frac{\mathrm{diam} C_j}{2} \right)^s ; \; A \subset \cup_{j=1}^{\infty} C_j, \; \mathrm{diam} C_j \leq \delta \right\} \right]$$

  where

$$\alpha(s) = \frac{\pi^{s/2}}{\Gamma(1 + s/2)}.$$

- Let $X = (x_1, \ldots, x_p)^\top$ be distributed as $N_p(\Xi, \sigma^2 I_p)$ and $\widehat{\Xi} = (\widehat{\xi}_1, \ldots, \widehat{\xi}_p)^\top$ where $\widehat{\xi}_i : \mathbb{R}^p \longrightarrow \mathbb{R}$ $(i = 1, 2, \ldots, p)$.

Stein's lemma by Niels R. Hansen(SPL(2018))

- Let $\boldsymbol{A} \subset \mathbb{R}^{\boldsymbol{p}}$ be a closed subset s.t. $\widehat{\Xi} : \boldsymbol{A}^{\boldsymbol{c}} \longrightarrow \mathbb{R}^{\boldsymbol{p}}$ is contineously diffrentiable. If $\mathcal{H}^{\boldsymbol{p-1}}(\boldsymbol{A}) = \mathbf{0}$ and

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \mathbb{E}\left[\left|\frac{\partial \widehat{\xi}_{\boldsymbol{i}}}{\partial \boldsymbol{x}_{\boldsymbol{i}}}\right|\right] < \infty, \qquad (2)$$

  then

$$\frac{\mathbf{1}}{\boldsymbol{\sigma^2}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \mathbf{Cov}(\widehat{\xi}_{\boldsymbol{i}}, \boldsymbol{x}_{\boldsymbol{i}}) = \frac{\mathbf{1}}{\boldsymbol{\sigma^2}} \mathbb{E}[(\boldsymbol{X} - \Xi)^{\mathsf{T}} \widehat{\Xi}] = \mathbb{E}[\mathbf{div}(\widehat{\Xi})] = \mathbb{E}\left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \frac{\partial \widehat{\xi}_{\boldsymbol{i}}}{\partial \boldsymbol{x}_{\boldsymbol{i}}}\right].$$

- The condition (**3**) is replace with

$$\mathbf{div}(\widehat{\Xi}) \geq \mathbf{0}, \quad \text{Lebesgue almost everywhere.}$$

- Hansen (2018) showed that Stein's lemma is valid for

$$\widehat{\widehat{\Xi}} = \sum_{i=1}^{p} h_i(\ell_i) u_i v_i^{\top},$$

if

$$h_i : (0, \infty) \longrightarrow \mathbb{R} \text{ is continously differentiable}$$

for $i = 1, 2, \ldots, p - 1$ and

$h_p : [0, \infty) \longrightarrow \mathbb{R}$ is continously differentiable with $h_p(0) = 0$

as well as (3) is satisfied for each h_i $(i = 1, 2, \ldots, p)$.

Case when the covarianc matrix is unknown

- Assume that $\min\{m, n\} \geq p$ and that

$$\begin{matrix} & p \\ m \\ n \end{matrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Xi \\ 0 \end{pmatrix} + E; \qquad E = \begin{matrix} p \\ \left(\begin{matrix} \longleftrightarrow \\ \longleftrightarrow \\ \vdots \\ \longleftrightarrow \\ \longleftrightarrow \end{matrix}\right) \end{matrix}$$

- The $m \times p$ mean matrix Ξ is of rank $r < p$
- The error E is an $(m + n) \times p$ error matrix(unobservable) whose rows are identically distributed as $N_p(0, \Sigma)$.
- The covariance matrix Σ is a $p \times p$ positive-definite and unknown.

- We consider the problem of estimating Ξ under low-rank mean matrix condition, i.e.,

$$\mathbf{rank}\,\Xi = r < p; \qquad r \text{ is unknown.}$$

- A loss fucntion and its risk are given by

$$L_2(\widehat{\Xi}, \Xi) = \mathbf{tr}\,\{(\widehat{\Xi} - \Xi)\Sigma^{-1}(\widehat{\Xi} - \Xi)^\top\} =: \|\widehat{\Xi} - \Xi\|^2_{F,\Sigma}$$

and

$$R_2(\widehat{\Xi}, \Xi) = \mathbb{E}[L_2(\widehat{\Xi}, \Xi)]$$

where $\widehat{\Xi}$ is an estimator based on (X, S).

- $S = Y^\top Y \sim W_p(\Sigma, n)$, which is the Wishart distribution with the degree of freedom n and the scale matrix Σ.

- To derive a class of estimators, first assume that Σ is known.
- Then we have

$$\boldsymbol{X}\Sigma^{-1/2} \sim \boldsymbol{N}_{m\times p}(\tilde{\Xi}, \boldsymbol{I}_m \otimes \boldsymbol{I}_p), \qquad \tilde{\Xi} = \Xi\Sigma^{-1/2}$$

which leads to an estimator of $\tilde{\Xi}$ given by

$$\hat{\tilde{\Xi}}_{TLS} = \underset{\text{rank}\,\Xi\le r}{\operatorname{argmin}} \|\boldsymbol{X}\Sigma^{-1/2} - \Xi\|_F^2 \implies \hat{\Xi} = \hat{\tilde{\Xi}}_{TLS}\Sigma^{1/2}.$$

- Hence we consider a class of estimators of the form

$$\hat{\Xi}_H = \left(\sum_{i=1}^{p} h_i(\ell_i)\boldsymbol{u}_i\boldsymbol{v}_i^\top\right)\boldsymbol{S}^{1/2};\ \boldsymbol{X}\boldsymbol{S}^{-1/2} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^\top$$

where $\boldsymbol{L} = \mathbf{diag}(\ell_1, \ldots, \ell_p)$, $\boldsymbol{H} = \mathbf{diag}(h_1, \ldots, h_p)$, $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ s.t. $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_p$.

- If

$$h_i(\ell_i) = \ell_i - \frac{c}{\ell_i} \ (i = 1, 2, \ldots, p);$$

c is a known positive constant,

then it results in the Efron-Morris estimator which is given by

$$\begin{aligned} \widehat{\Xi}_H &= XS^{-1/2}\left[I_p - c\{(XS^{-1/2})^\top (XS^{-1/2})\}^{-1}\right] S^{1/2} \\ &= X - cX\{X^\top X\}^{-1} S. \end{aligned}$$

- On the other hand, Tsukuma and Kubokawa (2015) considered estimators of the form

$$\widehat{\Xi}_T = X - UTU^\top X,$$

where $T = \mathbf{diag}(t_1(\ell_1^2), \ldots, t_p(\ell_p^2))$.

- Recall that

$$\boldsymbol{X}\boldsymbol{S}^{-1/2} = \boldsymbol{U}\boldsymbol{L}\boldsymbol{V}^{\top} \Longleftrightarrow \boldsymbol{L}^{-1}\boldsymbol{U}^{\top}\boldsymbol{X} = \boldsymbol{V}^{\top}\boldsymbol{S}^{1/2}.$$

  From a simple calculation we get

$$\widehat{\Xi}_{\boldsymbol{H}} \quad = \quad \boldsymbol{U}\boldsymbol{H}\boldsymbol{V}^{\top}\boldsymbol{S}^{1/2} = \boldsymbol{U}\boldsymbol{H}\boldsymbol{L}^{-1}\boldsymbol{U}^{\top}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{L}^{-1}\boldsymbol{H}\boldsymbol{X}.$$

- If we set $\boldsymbol{I_p} - \boldsymbol{T} = \boldsymbol{L}^{-1}\boldsymbol{H}\,(\boldsymbol{t_i}(\boldsymbol{x}) = \boldsymbol{h_i}(\sqrt{\boldsymbol{x}}))$, then we have

$$\widehat{\Xi}_{\boldsymbol{H}} = \widehat{\Xi}_{\boldsymbol{T}}.$$

- From this we can see that

$$\mathbf{L}_{\boldsymbol{M}}(\widehat{\Xi}_{\boldsymbol{H}}, \Xi) \quad = \quad \mathbf{L}_{\boldsymbol{M}}(\widehat{\Xi}_{\boldsymbol{T}}, \Xi).$$

- Furthermore, using the result due to Tsukuma and Kubokawa (2015), we have

$$
\begin{aligned}
\mathbf{R}_{M}(\widehat{\Xi}_{T}, \Xi) &= \mathbb{E}[\mathbf{SURE}(T)]; \\
\mathbf{SURE}(T) &= \sum_{i=1}^{p}\Bigg[m + a\ell_i^2 t_i^2 - 2bt_i - 4\ell_i^2 t_i t_i' - 4\ell_i^2 t_i' \\
&\qquad -2\sum_{j\neq i}^{p}\frac{\ell_i^4 t_i^2 - \ell_j^4 t_j^2}{\ell_i^2 - \ell_j^2} - 4\sum_{j\neq i}^{p}\frac{\ell_i^2 t_i - \ell_j^2 t_j}{\ell_i^2 - \ell_j^2}\Bigg]; \\
t_i &= 1 - \frac{h_i(\ell_i)}{\ell_i}; \qquad t_i' = -\frac{1}{2\ell_i^2}\left(h_i'(\ell_i) + \frac{h(\ell_i)}{\ell_i}\right), \\
a, b &: \text{known positive constants.}
\end{aligned}
$$

- Then we have an adaptive soft-thresholding rule

$$\widehat{\Xi}_{\widehat{\lambda}} \;=\; : \mathbf{SVT}_{\widehat{\lambda}}(XS^{-1/2})S^{1/2} = \left(\sum_{i=1}^{p}(\ell_i - \widehat{\lambda})_+ u_i v_i^\top\right) S^{1/2}$$

where

$$\widehat{\lambda} \;=\; \underset{\lambda \geq 0}{\mathbf{argmin}}\, \mathbf{SURE}(\mathbf{SVT}_\lambda)(XS^{-1/2});$$

$$\begin{aligned}\mathbf{SURE}(\mathbf{SVT}_\lambda)(XS^{-1/2}) \;=\; \sum_{i=1}^{p}\Bigg[& m + a\ell_i^2 t_i^2 - 2bt_i - 4\ell_i^2 t_i t_i' \\ & -4\ell_i^2 t_i' - 2\sum_{j\neq i}^{p}\frac{\ell_i^4 t_i^2 - \ell_j^4 t_i^2}{\ell_i^2 - \ell_j^2} - 4\sum_{j\neq i}\frac{\ell_i^2 t_i - \ell_i^2 t_i}{\ell_i^2 - \ell_j^2}\Bigg];\end{aligned}$$

$$t_i = 1 - \frac{(\ell_i - \lambda)_+}{\ell_i}\ (i = 1, \ldots, p);$$

$$t_i' = -\frac{1}{2\ell_i^2}\left(\mathbb{1}\{\ell_i > \lambda\} + \frac{(\ell_i - \lambda)_+}{\ell_i}\right).$$

Special case

- $\Sigma = \sigma^2 I_p$ where σ is postive but unknown.
- Let $s^2 = \mathrm{tr}(Y^\top Y)/p$.
- Then an adaptive soft-thresholding rule for this case is given by $\widehat{\Xi}_{\widehat{\lambda}} = \sum_{i=1}^{p} \left(\ell_i - \widehat{\lambda} s^2\right)_+ u_i v_i^\top;, \quad X = ULV^\top$, with $\widehat{\lambda} = \underset{\lambda \geq 0}{\mathrm{argmin}} \mathrm{SURE}(SVT_\lambda)(X)$ and

$$\mathrm{SURE}(SVT_\lambda)(X) = \sum_{i=1}^{p} \Bigg[m s^2 + a\ell_i^2 t_i^2 - 4\ell_i t_i' - 2 \sum_{j \neq i}^{p} \frac{\ell_i^4 t_i^2 - \ell_j^4 t_j^2}{\ell_i^2 - \ell_j^2} + s^2 \left(a\ell_i^2 t_i^2 - 4\ell_i^2 t_i t_i' - 4 \sum_{j \neq i}^{p} \frac{\ell_i^2 t_i - \ell_j t_j}{\ell_i^2 - \ell_j^2} \right) \Bigg]$$

$$t_i = 1 - \frac{(\ell_i - \lambda s^2)_+}{\ell_i}; \; t_i' = -\frac{1}{\ell_i^2} \left(\mathbb{1}\{\ell_i > \lambda s^2\} + \frac{(\ell_i - \lambda s^2)_+}{\ell_i} \right).$$

- **Remark**: It is routine to convert this result to case for complex normal distribution.

Simulation Study

References

1. Candés, E.J., Sing-Long, C.A., and Trzasko, J.D. (2013): IEEE on Signal Processing **61** 4643–4657.
2. Chételat, D. and Wells, M.T. (2012): AOS **40** 3137–3160.
3. Efron, B. (2004): JASA **99** 619–642.
4. Hansen, N.R. (2018): SPL **135** 76–88.
5. Josse, J. and Sardy, S. (2016): Stat. Comput **26** 715–724.
6. Mukherjee, A., Chen, K., Wang, N., and Zhu, L. (2015): Biometrika **102** 457–477.
7. Tsukuma, H. and Kubokawa, T. (2015): JMVA **139** 312–328.