Adaptive shrinkage of singular values of a low-rank mean matrix when a covariance matrix is unknown

Yoshihiko KONNO

Osaka Metropolitan Univeristy/JWU

Workshop

Mathematical optimazaiton and statistical theories using geometric methods

20-21 Octorber 2022

1. The reconstruction of a low-rank matrix from its noisy observation is useful in many applications. This problem is reformulated into a constrained nuclear norm minimization problem (regularized problem).
2. An important ingrident of this problem is how to choose a regularization parameter based on data. Usually the data is independently and identically distributed with unknown variance.
3. (1) The discrepacy principle approach, (2) Stein's Unbiased risk estimator(SURE) approach.
4. Inspired by approach(2) we consider the problem of estimating a low-rank matrix mean in MANOVA(Mulitivariate Analysis of Variance) setting [1]when a positive-definite covariance matrix of error is unknown.

[1] We have data for unknown covariance matrix. The distribution of this data is mean-zero.

MANOVA model and its canonical model

Let $m, n, p \in \mathbb{N}$ such that $\min(m, p) \geq p$. Consider a multivariate regression model

$$\underbrace{W}_{(m+n)\times p} = \underbrace{A}_{(m+n)\times m} \underbrace{B}_{m\times p} + \underbrace{\mathrm{Err}}_{(m+n)\times p},$$

where A is a known design matrix of full rank, B is an unknown regression matrix of rank r $(< \min(m, p)$ and r is unknown$)$, and Err is an unobservable error matrix. Here rows of Err are independently and identically distributed as $N_p(0_p, \Sigma)$ where Σ is a $p \times p$ positive-definite unknown matrix.

Notation

1.
$$\mathrm{Err} = \begin{bmatrix} e_1^\top \\ e_2^\top \\ \vdots \\ e_{m+n}^\top \end{bmatrix} : \ (m+n) \times p, \quad \mathrm{vec}(\mathrm{Err}) := \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{m+n} \end{bmatrix}$$
where e_j's are independently and identically distributed as $N_p(0_p, \Sigma)$ $(j = 1, 2, \ldots, (m+n))$.

2. Write
$$\begin{aligned} \mathbb{COV}(\mathrm{Err}) &= \mathbb{E}\Big[\Big\{\mathrm{vec}(\mathrm{Err} - \mathbb{E}[\mathrm{Err}]\Big\}\Big\{\mathrm{vec}(\mathrm{Err} - \mathbb{E}[\mathrm{Err}]\Big\}^\top\Big] \\ &= I_{m+n} \otimes \Sigma, \\ \mathrm{Err} &\sim N_{(m+n)\times p}(0_{(m+n)\times p}, I_{m+n} \otimes \Sigma). \end{aligned}$$

1. Let
$$P = (A^\top A)^{-1/2} A^\top : \; m \times (m+n)$$
and take $P^\perp : \; n \times (m+n)$ s.t.
$$P(P^\perp)^\top = 0_{m\times n} \quad \text{and} \quad P^\perp (P^\perp)^\top = I_n,$$
Note that
$$\begin{bmatrix} P \\ P^\perp \end{bmatrix} [P^\top, (P^\perp)^\top] = I_{m+n}.$$

2. Put $\Xi := \left(A^\top A\right)^{1/2} B$ and
$$\begin{bmatrix} X \\ Y \end{bmatrix} := \begin{bmatrix} P \\ P^\perp \end{bmatrix} W \sim N_{(m+n)\times p}\left(\begin{bmatrix} \Xi \\ 0_{n\times p} \end{bmatrix}, I_{m+n} \otimes \Sigma \right).$$

Problem set-up

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

$$\begin{array}{cc} & p \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} X \\ Y \end{pmatrix} \end{array} = \begin{pmatrix} \Xi \\ 0 \end{pmatrix} + E; \qquad E = \begin{array}{c} p \\ \begin{pmatrix} \longleftrightarrow \\ \longleftrightarrow \\ \vdots \\ \longleftrightarrow \\ \longleftrightarrow \end{pmatrix} \end{array}$$

where $\begin{bmatrix} X \\ Y \end{bmatrix}$ is observation and Ξ is an $m \times p$ non-random matrix(unknown) of rank $r < p$, E is an $(m + n) \times p$ error matrix(unobservable) whose rows are identically distributed as $N_p(0, \Sigma)$. Here Σ is a $p \times p$ positive-definite and unknown matrix.

We consider the problem of estimating $\boldsymbol{\Xi}$ under a low-rank mean matrix condition, i.e.,

$$\mathrm{rank}\, \boldsymbol{\Xi} = r < p; \quad r \text{ is unknown}$$

under a loss fucntion and its risk

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

and

$$\mathbf{R}_{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi}) = \mathbb{E}[\mathbf{L}_{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Xi}}, \boldsymbol{\Xi})]$$

where $\widehat{\boldsymbol{\Xi}}$ is an estimator based on $(\boldsymbol{X}, \boldsymbol{S})$. Here $\boldsymbol{S} = \boldsymbol{Y}^{\top}\boldsymbol{Y} \sim W_p(\boldsymbol{\Sigma}, n)$, which is the Wishart distribution with the degree of freedom n and the scale matrix $\boldsymbol{\Sigma}$.

Mean matrix estimation when a covariance is known

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

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{z}_1^{\mathrm{T}} \\ \boldsymbol{z}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{z}_m^{T} \end{pmatrix}$$

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

$$\boldsymbol{z}_i : \; \boldsymbol{p} \times \mathbf{1} \sim \boldsymbol{N}_{(}\widetilde{\boldsymbol{\xi}}_i, \sigma^2 \boldsymbol{I}_p), \qquad (\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{m})$$

Here $\widetilde{\boldsymbol{\Xi}}^{\mathrm{T}} := (\widetilde{\boldsymbol{\xi}}_1, \ldots, \widetilde{\boldsymbol{\xi}}_m)$ is unknown but $\sigma > \mathbf{0}$ are known.

- We assume that low-rank mean matrix condition, i.e.,

$$\mathrm{rank}\left(\widetilde{\Xi}\right) = r < p; \quad r \text{ is unknown}.$$

- Consider the problem of estimating $\widetilde{\Xi}$ under a loss fucntion and its risk

$$\mathrm{L}_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi}) = \mathrm{tr}\,\{(\widehat{\widetilde{\Xi}} - \widetilde{\Xi})(\widehat{\widetilde{\Xi}} - \widetilde{\Xi})^\top\} =: \|\widehat{\widetilde{\Xi}} - \widetilde{\Xi}\|_F^2$$

and

$$\mathrm{R}_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi}) = \mathbb{E}[\mathrm{L}_1(\widehat{\widetilde{\Xi}}, \widetilde{\Xi})].$$

- Here $\widehat{\widetilde{\Xi}}$ is an estimator based on $\boldsymbol{Z}$.
- $\mathrm{tr}\,\boldsymbol{A}$ and $\boldsymbol{A}^\top$ stand for the traace and the transpose of a matrix $\boldsymbol{A}$, respectively.
- $\|\boldsymbol{A}\|_F := \sqrt{\mathrm{tr}\left(\boldsymbol{A}^\top \boldsymbol{A}\right)}$, the Frobenius norm of a matrix $\boldsymbol{A}$.

Eckart-Young approximation theorem

- Singular Value Decomposition: We can assume that $m \geq p$ without loss of generality. Decompose Z as

$$\begin{aligned} Z &= ULV^\top; \quad U = (u_1, \ldots, u_p),\ V = (v_1, \ldots, v_p) \\ L &= \mathrm{diag}(\ell_1, \ell_2, \ldots, \ell_p) \quad \text{with } \ell_1 \geq \ell_2 \geq \cdots \geq \ell_p \geq 0 \end{aligned}$$

  where $u_i \in \mathbb{R}^m$, $v_i \in \mathbb{R}^p$ $(i = 1, \ldots, p)$ s.t. $U^\top \underbrace{U}_{m \times p} = V^\top V = I_p$.

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

$$\widehat{\widetilde{\Xi}}_{TLS} = \sum_{i=1}^{r} \ell_i u_i v_i^\top. \quad \Longleftrightarrow \quad \widehat{\widetilde{\Xi}}_{TLS} \in \operatorname*{argmin}_{\widetilde{\Xi}:\, \mathrm{rank}\,(\widetilde{\Xi}) \leq r} \|\widetilde{\Xi} - Z\|_F^2.$$

$\boxed{\text{Notaton}}$ $\sigma_j(A) > 0$ $(j = 1, 2, \ldots, r)$ are non-zero singular values of a matrix A with $r = \mathrm{rank}\,(A)$.

Regularization approach

- We consider an estimator which minimizes the penalized least squares criterion

$$\mathrm{Mat}(m, p; \mathbb{R}) \ni \Xi \mapsto \frac{1}{2}\|Z - \Xi\|_F^2 + \mathbf{pen}_\lambda\left(\Xi\right) \in [0, \infty)$$

where $\mathbf{pen}_\lambda\left(\cdot\right)$ (≥ 0) is a penality function of Ξ and λ (≥ 0) is a tuning parameter.

- Examples of penalities: For a positive $\lambda > 0$,

★ $\mathbf{pen}_\lambda(\Xi) = \lambda\mathbf{rank}\,(\Xi)$

$\Longrightarrow$ a hard-theshholding rule, i.e.,$\mathbf{SVHT}_\lambda(Z) = \sum_{j=1}^{p} \ell_j \mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top$,

where $\mathbb{1}\{\text{event}\} = \begin{cases} 1 & \text{if event is true,} \\ 0 & \text{otherwise} \end{cases}$.

★ $\mathbf{pen}_\lambda(\Xi) = \lambda\|\Xi\|_1 := \lambda \sum_{j=1}^{p} \left|\sigma(\Xi)_j\right| \quad \left(\sigma(\Xi)_j\text{: SV's of } \Xi\right)$

where $\left\{(\sigma(\Xi)_j,\ u_j,\ v_j)\right\}_{j=1}^{\min(m,p)}$ is a system of singular values of Ξ

$\Longrightarrow$ a soft-thresholding rule, i.e.,

$$\mathbf{SVST}_\lambda(Z) = \sum_{j=1}^{p} (\ell_j - \lambda)\mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top.$$

A hard-shreshholding rule

- Assume that σ^2 is known.
- Solve

$$\mathbf{SVHT}_\lambda(Z) = \operatorname*{argmin}_{\Xi}\left[\frac{1}{2}\|\Xi - Z\|_F^2 + \lambda\,\mathrm{rank}\,(\Xi)\right]$$

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

$$\mathbf{SVHT}_\lambda(Z) = \sum_{j=1}^{p} \ell_j\, \mathbb{1}\{\ell_j \geq \lambda\} u_j v_j^\top;\ \mathbb{1}\{\ell_j \geq \lambda\} = \begin{cases} 1 & \ell_j \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

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

  (See Donoho and Garvish (2017, IEEE, Trans. Inform Theory).

Steps to obtain an adaptive thresholding esitmator

1. Solve regularized minimizaton problem

$$\widehat{\widetilde{\Xi}}_\lambda \in \underset{\Xi \in \mathrm{Mat}(m,p;\mathbb{R})}{\arg\min} \left\{ \|Z - \Xi\|_F^2 + \mathrm{pen}_\lambda(\Xi). \right\}.$$

2. Calculate SURE if possible $\left(\text{a closed form of } \widehat{\widetilde{\Xi}}_\lambda\right)$:

$$\mathrm{R}_1\left(\widehat{\widetilde{\Xi}}_\lambda, \widetilde{\Xi}\right) = \mathbb{E}\left[\mathrm{SURE}\left(\widehat{\widetilde{\Xi}}_\lambda\right)\right]$$

   Note that $\mathbf{SURE}\left(\widehat{\widetilde{\Xi}}_\lambda\right)$ is a function of λ and observable data.

3. Solve minimization problem

$$\widehat{\lambda} \in \mathrm{SURE}\left(\widehat{\widetilde{\Xi}}_\lambda\right) \Longrightarrow \widehat{\widetilde{\Xi}}_{\widehat{\lambda}}.$$

Remarks

1. This method works for the soft-thresholding rule. See Cándes *el al.* (2013).
2. **SURE** does not work for the hard-thresholding rule since Stein's identity, integration-by-parts formula with respect to multivariate normal distribution, fails for the hard-thresholding rule becuase of discontinuity of estimator.

A soft-thersholding rule

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

$$\mathbf{SVST}_\lambda(Z) = \sum_{j=1}^{p} (\ell_i - \lambda)\, \mathbb{1}\{\ell_i \geq \lambda\} u_j v_j^\top =: \sum_{j=1}^{p} (\ell_j - \lambda)_+ u_j v_j^\top \quad (1)$$

which is obtained from

$$\min_Y \left\{ \frac{1}{2} \|Z - Y\|_F^2 + \lambda \sum_{j=1}^{p} \lambda_j \right\} \qquad Y = \mathbf{SVST}_\lambda(Z).$$

- 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

$$
\begin{aligned}
\mathrm{R}_1(\mathrm{SVST}_\lambda, \Xi) &= \mathbb{E}[\mathrm{SURE}(\mathrm{SVST}_\lambda)(Z)], \\
\mathrm{SURE}(\mathrm{SVST}_\lambda)(Z) &= -mp\sigma^2 + \sum_{j=1}^{p} \min\{\ell_j^2, \lambda^2\} \\
&\quad +2\sigma^2 \operatorname{div}(\mathrm{SVST}_\lambda(X)), \\
\operatorname{div}(\mathrm{SVST}_\lambda(Z)) &= (m-p)\sum_{j=1}^{p}\left(1-\frac{\lambda}{\ell_j}\right)_+ + \sum_{j=1}^{p} \mathbb{1}\{\ell_j > \lambda\} \\
&\quad +2\sum_{j=1}^{p}\sum_{k\neq i}^{p} \frac{\ell_j(\ell_j-\lambda)_+}{\ell_j^2-\ell_k^2}
\end{aligned}
$$

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

- An adaptive estimator is given by

$$\mathbf{SVST}_{\widehat{\lambda}}(Z) = \sum_{j=1}^{p} (\ell_i - \widehat{\lambda})_+ u_j v_i^{\top}, \tag{2}$$

$$\widehat{\lambda} \in \operatorname*{arg\,min}_{\lambda \geq 0} \left[\sum_{i=1}^{p} \min\{\ell_i^2, \lambda^2\} + 2\sigma^2 \mathrm{div}(\mathbf{SVST}_{\lambda}(Z)) \right].$$

- Numerical evaluation of the risk of **(2)** was carried out by Candés et. al.
- But it is not clear if $\mathbf{R_1}(\mathbf{SVST}_{\widehat{\lambda}}(Z), \widetilde{\Xi})$ is close to $\mathbf{R_1}(\widehat{\Xi}_{TLS}(Z), \widetilde{\Xi})$ for $\forall \widetilde{\Xi}$ s.t. $\mathrm{rank}\left(\widetilde{\Xi}\right) \leq r < \min(m, p)$.

Mean matrxi estimation when a covarianc matrix is unknown

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

$$\begin{array}{cc} & p \\ m & \\ n & \end{array} \!\!\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Xi \\ 0 \end{pmatrix} + E; \qquad E = \begin{array}{c} p \\ \begin{pmatrix} \longleftrightarrow \\ \longleftrightarrow \\ \vdots \\ \longleftrightarrow \\ \longleftrightarrow \end{pmatrix} \end{array}$$

- 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.,

$$\mathrm{rank}\,\Xi = r < \min(m, p); \quad r \text{ is } \text{unknown}.$$

- A loss fucntion and its risk are given by

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

and

$$R_{\Sigma}(\widehat{\Xi}, \Xi) = \mathbb{E}[L_{\Sigma}(\widehat{\Xi}, \Xi)]$$

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 $\boldsymbol{\Sigma}$ is known.
- Then we have

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

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

$$\widehat{\widetilde{\boldsymbol{\Xi}}}_{TLS} \in \operatorname*{arg\,min}_{\text{rank}\,\boldsymbol{\Xi}\le r} \|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{\Xi}\|_F^2 \implies \widehat{\boldsymbol{\Xi}} = \widehat{\widetilde{\boldsymbol{\Xi}}}_{TLS}\boldsymbol{\Sigma}^{1/2}.$$

- Hence we consider a class of estimators of the form

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

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

Regularized minimization problem

- Known $\boldsymbol{\Sigma}$ case: For $\lambda \geq \mathbf{0}$,

$$\mathrm{Mat}(m,\ p;\ \mathbb{R}) \ni \boldsymbol{\Xi}\boldsymbol{\Sigma}^{-1/2} \mapsto \|\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} - \widetilde{\boldsymbol{\Xi}}\boldsymbol{\Sigma}^{-1/2}\|_F{}^2 + 2\lambda\|\widetilde{\boldsymbol{\Xi}}\boldsymbol{\Sigma}^{-1/2}\|_1$$

- Unknonw $\boldsymbol{\Sigma}$ case: For $\lambda \geq \mathbf{0}$, find a minimizer $\widehat{\widetilde{\boldsymbol{\Xi}}}$ of a regularized minimization problem

$$\mathrm{Mat}(m,\ p;\ \mathbb{R}) \ni \widetilde{\boldsymbol{\Xi}} \mapsto \|\boldsymbol{X}\boldsymbol{S}^{-1/2} - \widetilde{\boldsymbol{\Xi}}\|_F{}^2 + 2\lambda\|\widetilde{\boldsymbol{\Xi}}\|_1$$

and

$$\widehat{\boldsymbol{\Xi}} = \widehat{\widetilde{\boldsymbol{\Xi}}}\boldsymbol{S}^{1/2} = \left(\sum_{j=1} \ell_j(\ell_j - \lambda)_+ \boldsymbol{u}_j \boldsymbol{v}_j^\top\right)\boldsymbol{S}^{1/2}$$

where $\{(\ell_j,\ \boldsymbol{u}_j,\ \boldsymbol{v}_j)\}$ is a system of singular values of $\boldsymbol{X}\boldsymbol{S}^{-1/2}$

- If

$$h_j(\ell_j) \quad = \quad \ell_j - \frac{c}{\ell_j}\,(j = 1, 2, \ldots, p);$$

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 $T = \mathrm{diag}(t_1(\ell_1^2), \ldots, t_p(\ell_p^2))$ and $XS^{-1/2} = ULV^\top$ with $m \times \min(m, p)$ matrix U s.t. $U^\top U = I_{\min(m,p)}$.

- Recall that

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

From a simple calculation we get

$$\widehat{\Xi}_H \;=\; UHV^\top S^{1/2} = UHL^{-1}U^\top X = UL^{-1}HX.$$

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

$$\widehat{\Xi}_H = \widehat{\Xi}_T.$$

- From this we can see that

$$\mathbf{L}_\Sigma(\widehat{\Xi}_H, \Xi) \;=\; \mathbf{L}_\Sigma(\widehat{\Xi}_T, \Xi).$$

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

$$
\begin{aligned}
R_\Sigma(\widehat{\Xi}_T, \Xi) &= \mathbb{E}[\mathrm{SURE}(T)]; \\
\mathrm{SURE}(T) &= \sum_{j=1}^{p}\Big[m + a\ell_j^2 t_j^2 - 2bt_i - 4\ell_j^2 t_j \widetilde{t}_j - 4\ell_i^2 \widetilde{t}_j \\
&\quad -2\sum_{k \neq j}^{p} \frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} - 4\sum_{k \neq j}^{p} \frac{\ell_j^2 t_j - \ell_k^2 t_k}{\ell_j^2 - \ell_k^2}\Big]; \\
t_i &= 1 - \frac{h_i(\ell_j)}{\ell_j}; \qquad t_i' = -\frac{1}{2\ell_i^2}\left(\widetilde{h}_j'(\ell_j) + \frac{h(\ell_j)}{\ell_j}\right), \\
a, b &: \text{ known positive constants.}
\end{aligned}
$$

- Then we have an adaptive soft-thresholding rule

$$\widehat{\Xi}_{\widehat{\lambda}} \;=\; : \mathrm{SVST}_{\widehat{\lambda}}(XS^{-1/2})S^{1/2} = \left(\sum_{j=1}^{p}(\ell_j - \widehat{\lambda})_+ u_j v_j^{\top}\right) S^{1/2}$$

where $\widehat{\lambda} = \operatorname{argmin}_{\lambda \geq 0} \mathrm{SURE}(\mathrm{SVST}_{\lambda})(XS^{-1/2})$;

$$\mathrm{SURE}(\mathrm{SVST}_{\lambda})(XS^{-1/2}) = \sum_{j=1}^{p}\Bigg[m + a\ell_j^2 t_j^2 - 2bt_j - 4\ell_j^2 t_j \widetilde{t}_j - 4\ell_j^2 \widetilde{t}_j - 2\sum_{k \neq j}^{p} \frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} - 4\sum_{j \neq i} \frac{\ell_i^2 t_i - \ell_i^2 t_i}{\ell_i^2 - \ell_j^2}\Bigg];$$

$$t_j = 1 - \frac{(\ell_j - \lambda)_+}{\ell_j} \; (j = 1, \ldots, p);$$

$$\widetilde{t}_j = -(2\ell_j)^{-2}\left(\mathbb{1}\{\ell_j > \lambda\} + \frac{(\ell_j - \lambda)_+}{\ell_j}\right).$$

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 $\widehat{\Xi}_{\widehat{\lambda}} = \sum_{j=1}^{p} \left(\ell_i - \widehat{\lambda} s^2\right)_+ u_j v_j^\top;$, $\quad X = ULV^\top$, with $\widehat{\lambda} = \underset{\lambda \geq 0}{\mathrm{argmin}} \mathrm{SURE}(SVST_\lambda)(X)$ and

$$
\begin{aligned}
\mathrm{SURE}(SVST_\lambda)(X) &= \sum_{j=1}^{p} \Bigg[m s^2 + a\ell_j^2 t_j^2 - 4\ell_j \widetilde{t}_j - 2\sum_{k \neq j}^{p} \frac{\ell_j^4 t_j^2 - \ell_k^4 t_k^2}{\ell_j^2 - \ell_k^2} \\
&\quad + s^2 \left(a\ell_j^2 t_j^2 - 4\ell_j^2 t_j \widetilde{t}_j - 4 \sum_{k \neq j}^{p} \frac{\ell_j^2 t_j - \ell_k t_k}{\ell_j^2 - \ell_k^2} \right) \Bigg]
\end{aligned}
$$

where

$$t_j = 1 - \frac{(\ell_j - \lambda s^2)_+}{\ell_i}; \ \widetilde{t}_j = -\frac{1}{\ell_j^2}\left(\mathbb{1}\{\ell_j > \lambda s^2\} + \frac{(\ell_j - \lambda s^2)_+}{\ell_j}\right).$$

Concluding remarks

1. Derivation of an adaptive threshholding rule:
   - For $\lambda \geq 0$, solve a regularized minimizaton problem(random one) $\widehat{\widehat{\Xi}} \in \underset{\Xi \in \mathrm{Mat}(m,p;\mathbb{R})}{\arg\min} \in \Xi \left\{ \frac{1}{2} \| X S^{-1/2} - \Xi \|_F{}^2 + \lambda \| \Xi \|_1 \right\}.$
   - We have $\widehat{\Xi}_\lambda = \widehat{\widehat{\Xi}} S^{1/2} = \left(\sum_{j=1} \ell_j (\ell_j - \lambda)_+ u_j v_j^T \right) S^{1/2}.$

     wherer $\left\{ (\ell_j,\ u_j,\ v_j \right\}_{j=1,2,\ldots,m}$ is a system of singular values of $XS^{-1/2}$.
   - Obtain SURE $\quad \mathbf{R}_\Sigma \left(\widehat{\Xi}_\lambda,\ \Xi \right) = \mathbb{E}\left[\mathrm{SURE}\left(\widehat{\Xi}_\lambda \right) \right]$
   - Solve the minimization problem $\widehat{\lambda} \in \underset{\lambda \geq 0}{\arg\min}\, \mathrm{SURE}\left(\widehat{\Xi}_\lambda \right) \Longrightarrow \widehat{\Xi}_{\widehat{\lambda}}.$.
2. It is routine to convert this result to case for complex normal distribution.

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. Li, K., Li, H., Chen, R.H., and Wen, Y.-W. (2022): SIAM JSC **44** A2204-A2225.
7. Mukherjee, A., Chen, K., Wang, N., and Zhu, L. (2015): Biometrika **102** 457–477.
8. Tsukuma, H. and Kubokawa, T. (2015): JMVA **139** 312–328.