National Central University, Taiwan (4 Dec 2012)

Shrinkage estimators for large covariance matrices in multivariate normal distribution

Yoshihiko KONNO, Japan Women's University

4 Dec. 2012

Outline

In this talk we consider the problem of estimating large covariance matrices in a decision-theoretic manner when the dimension of variables p is larger than the number of observations n (both n and p are finite). Population distributions include not only real multivariate distributions but also complex multivariate distributions. The resulsts of real and complex cases parallel each other. See Konon (JMVA, 2009).

Outline

(1) Wishart matrix and its distribution
(2) Notation and settup of the problem
(3) Class of estimators and Stein's Unbiased Risk Estmate (SURE) method.
(4) Improved estimators
(5) Simulation Study

Wishart distribution (1)

Assume that $\mathbb{R}^p$-valued random column vectors $X_1,\ X_2,\ \ldots,\ X_n$ are distributed as multivariate normal distribution with mean zero vector and covariance $\boldsymbol{\Sigma}$:

$$N_p(0,\ \boldsymbol{\Sigma})(dx) = \frac{1}{(2\pi)^{p/2}\mathrm{det}(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}x'\boldsymbol{\Sigma}^{-1}x\right\}\, dx$$

where dx is Lebesgue measure on $\mathbb{R}^p$ and

$$\boldsymbol{\Sigma} \in \mathrm{Sym}_p^+(\mathbb{R}) = \{\boldsymbol{S} \in \mathbb{R}^{p\times p} :\ x'\boldsymbol{S}x > 0 \quad \text{for } x \in \mathbb{R}^p \setminus \{0\}\}.$$

Here x' is the transpose of column vector x.

Wishart distribution (2)

$\boxed{\text{Wishart Matrix}}$

Let $\boldsymbol{X} : n \times p = [X_1, \ldots, X_n]'$ and put

$$\boldsymbol{W} : p \times p = \boldsymbol{X}'\boldsymbol{X} = \sum_{j=1}^{n} X_i X_i'.$$

⋆ By definition $\boldsymbol{W} \in \overline{\mathrm{Sym}_p^+(\mathbb{R})}$ (the closure of $\mathrm{Sym}_p^+(\mathbb{R})$).

⋆ It is known that

$$\mathbb{P}\{\boldsymbol{W} \in \mathrm{Sym}_p^+(\mathbb{R})\} = 1 \Longleftrightarrow n \geq p.$$

⋆ When $n < p$ the distribution of $\boldsymbol{W}$ concentrates on the boundary of $\mathrm{Sym}_p^+(\mathbb{R})$.

Wishart distribution (3)

Normal distribution with mean zero	$\stackrel{\text{sum of square}}{\Longrightarrow}$	Gamma distribution (χ^2-distribution)
$\Downarrow$	(multivariate generation)	$\Downarrow$
Multivariate normal distribution with mean zero vector	$\stackrel{\text{Wishart matrix}}{\Longrightarrow}$	Wishart distribution

Wishart distribution (4)

Wishart density ($\boldsymbol{\Sigma} \in \mathrm{Sym}_p^+(\mathbb{R})$ and $n \geq p$)

$$W_p(\boldsymbol{\Sigma},\, n)(dx) = \frac{1}{c_{p,\,n}(\det \boldsymbol{\Sigma})^{n/2}}(\det x)^{n/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}x\right)\right) \times 1_{\mathrm{Sym}_p^+(\mathbb{R})}(x)\frac{dx}{(\det\, x)^{(p+1)/2}},$$

where dx is Lebesgue measure on the space of $p \times p$ symmetric matrices,

$$c_{p,\,n} = 2^{np}\pi^{p(p-1)/4}\prod_{i=0}^{p-1}\Gamma((n-i)/2)$$

$$1_{\mathrm{Sym}_p^+(\mathbb{R})}(x) = \begin{cases} 1 & (x \in \mathrm{Sym}_p^+(\mathbb{R})), \\ 0 & (\text{otherwise}) \end{cases}$$

Decomposition of Wishart matrix

$$\boldsymbol{W} = \boldsymbol{H}\boldsymbol{L}\boldsymbol{H}', \quad \boldsymbol{H} \in \mathrm{O}_p(\mathbb{R}), \quad \boldsymbol{L} = \mathrm{diag}(\ell_1, \ell_2, \ldots, \ell_p)$$

with $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_p \geq 0$

Here

$$\mathrm{O}_p(\mathbb{R}) = \{\boldsymbol{H} \in \mathbb{R}^{p\times p} : \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_p\}$$

and $\boldsymbol{I}_p$ is the identity matrix of size p ($p \times p$ matrix).

Remark Observe that $\ell_{n+1} = \cdots = \ell_p = 0$ when $n < p$ and that the distribution of eigenvalues is complicated except $\boldsymbol{\Sigma} = c\boldsymbol{I}_p\,(c > 0)$.

Distribution of the eigenvalues of Wishart matrix (1)

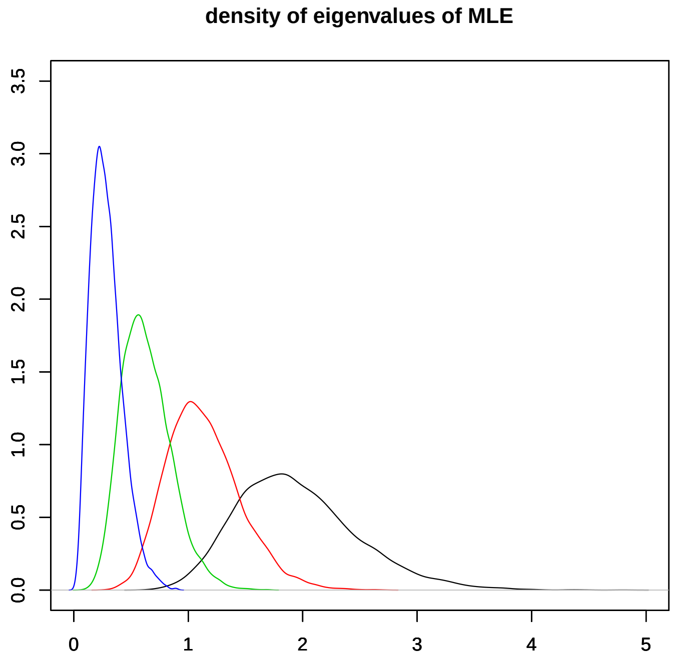

Figure 1: $(\ell_1, \ldots, \ell_4)/10$ $(p = 4,\ n = 10)$: Plot of simulated eigenvalues (divided by 10) of $W_4(\boldsymbol{I}_4, 10)$.

Distribution of the eigenvalues of Wishart matrix (2)

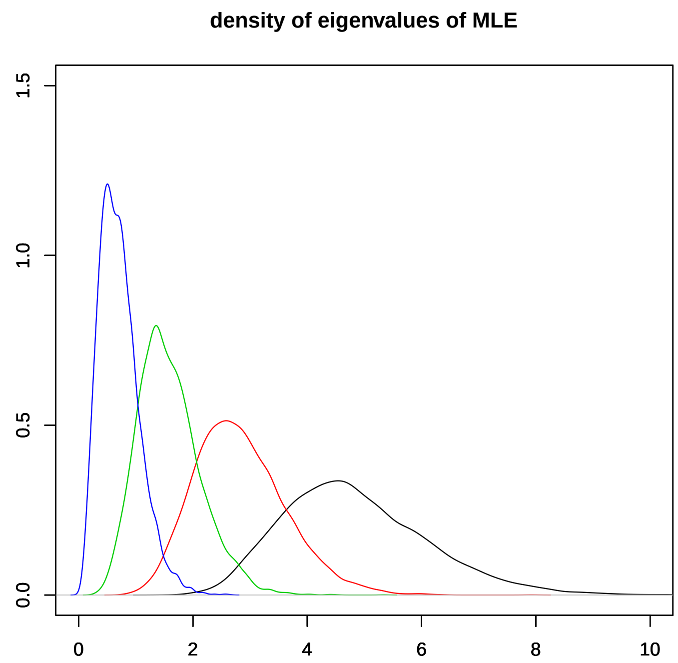

Figure 2: $(\ell_1, \ldots, \ell_{10})/4$ $(p = 10,\ n = 4)$: Plot of simulated eigenvalues (divided by 4) of $W_{10}(\boldsymbol{I}_{10}, 4)$.

Distribution of the eigenvalues of Wishart matrix (3)

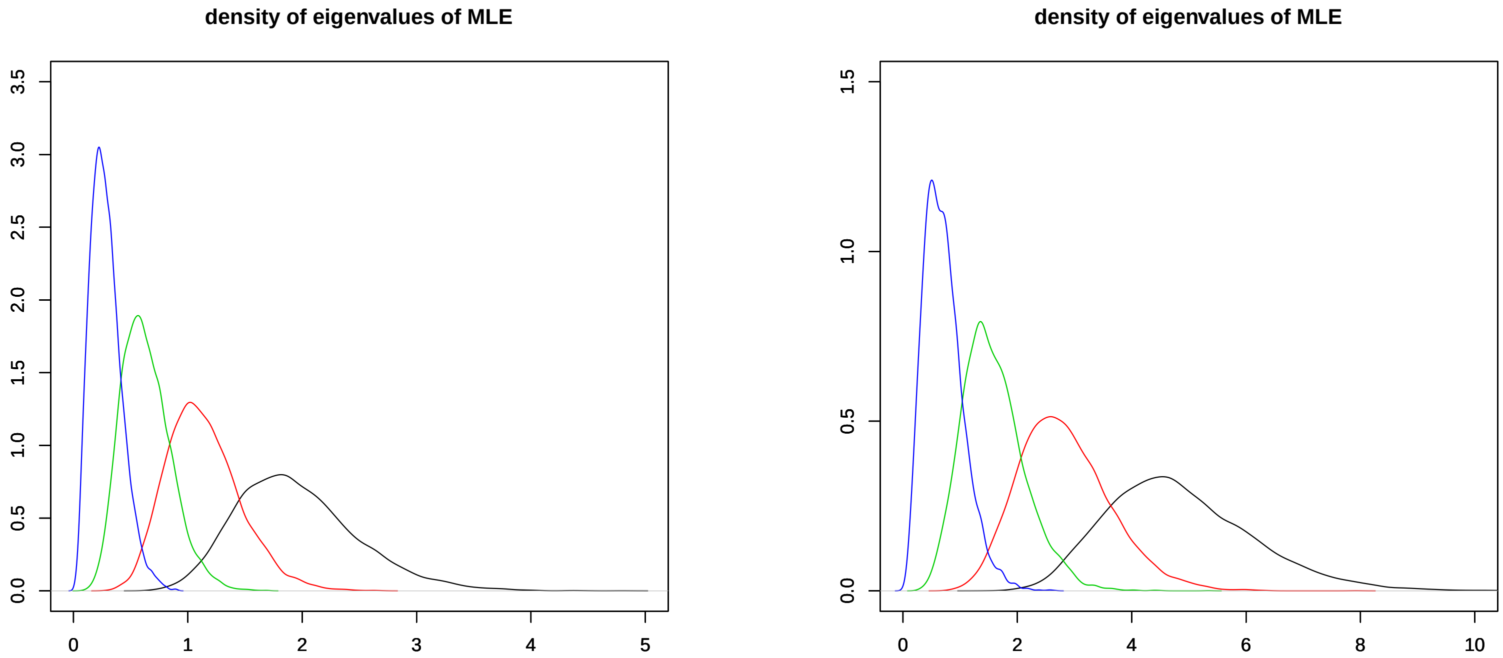

Figure 3: left: $W_4(\boldsymbol{I}_4,\, 10)$, right: $W_{10}(\boldsymbol{I}_{10},\, 4)$

Distribution of the eigenvalues of Wishart matrix: from RMT view

Marčhenko and Pastur law[1]

When $p = p(n)$ and $p/n \to \tau \leq 1 (n \to \infty)$,

$$\frac{1}{p}\sum_{i=1}^{p} 1_{\lambda_i \leq x} \to \int_{-\infty}^{x} \frac{\sqrt{(b-t)(t-a)}}{2\pi t\tau} 1_{a \leq t \leq b}\, dt$$

(almost surely).

Here $a = (1-\tau^{1/2})^2$, $b = (1+\tau^{1/2})^2$.

See Lecture 4 of [12].

Distribution of the eigenvalues of Wishart matrix (5)

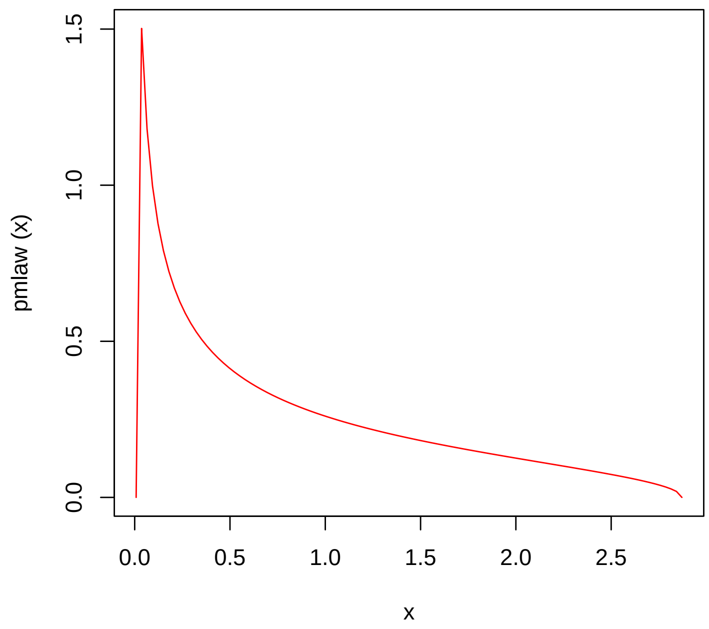

Figure 4: Density of Marčhenko-Pastur's quater-circle law ($\tau = 4/7 \doteqdot 0.57$, $a \doteqdot 0.06$, $b \doteqdot 3.086$).

Shrinkage estimators (1)

⋆ $\boxed{n \geq p}$ Assume that $\boldsymbol{W} : p \times p \sim W_p(\boldsymbol{\Sigma}, n)$ $(\Sigma > 0$and $n \geq p)$ and $\ell_1 > \ell_2 > \cdots > \ell_p > 0$ (with probability one) are the eigenvalues of $\boldsymbol{W}$. Then $\boldsymbol{W}/n$ is unbiased estimator for $\boldsymbol{\Sigma}$.

⋆ However $(\ell_1/n, \ldots, \ell_p/n)$, the eigenvalues of $\boldsymbol{W}/n$, are biased estimator for the eigenvalues of $\boldsymbol{\Sigma}$.

$$\Downarrow$$

Shrink eigenvalues toward a center!

⋆ Many shrinkage estimators have been proposed under different risk functions.

⋆ Stein's unbiased risk estimate method works well to evaluate the performance of the shrinkage estimators.

Shrinkage estimators (2)

⋆ $\boxed{n < p}$ Problem is more difficult since $\boldsymbol{W}$ is not invertable and $\ell_{n+1} = \cdots = \ell_p = 0$.

⋆ Ledoit and Wolf(2004, JMVA) proposed a shrinkage estimator under a quadratic loss function and evaluate the performace of the estimator asymptotically as the number of observations and the number of variables go to infinity together.

⋆ Our goal is to evaluate the performace of different shrinkage estimator (modification of Haff's estimator for $n \geq p$) under different loss function in a finite sample settup.

Notation and settup

⋆ Let $X_1,\, X_2,\, \ldots,\, X_n \sim N_p(0,\, \mathbf{\Sigma})$, independently. Each $X_i\,(i = 1,\, 2, \ldots,\, n)$ is a p-variate column vector. Here $\mathbf{\Sigma}$ is unknown $p \times p$ positive definite matrix.

⋆ n equals "sample size minus 1" and p is dimension of the variables.

⋆ Defien Wishart matrix($p \times p$ matrix) as $\boldsymbol{W} := \sum_{k=1}^{n} X_i X_i'$. Here "$'$" denotes the transpose of the matrix and the vector.

⋆ Consider the estimation problem of covariance matrix $\mathbf{\Sigma}$ under a loss function

$$\text{loss: } L(\widehat{\mathbf{\Sigma}},\, \mathbf{\Sigma}) = \mathrm{tr}\,(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1} - I_p)^2; \qquad \text{risk: } R(\widehat{\mathbf{\Sigma}},\, \mathbf{\Sigma}) := \mathbb{E}[L(\widehat{\mathbf{\Sigma}},\, \mathbf{\Sigma})].$$

Here $\widehat{\mathbf{\Sigma}}$ is an estimator for $\mathbf{\Sigma}$, and $\boldsymbol{I}_p$ denotes $p \times p$ identity matrix.

⋆ Our goal is to compare the risk among the estimators uniformly in $\mathbf{\Sigma}$.

Some remarks (1)

⋆ Without loss of generality, we can assume that the mean is zero.

⋆ The Wishart matrix Wishart $\boldsymbol{W}$ is positive-deffinite $\Longleftrightarrow n \geq p$.

⋆ When $p > n$, the distribution of $\boldsymbol{W}$ exists, but its density does not exists in usual sense.

⋆ Examples of loss functions which are invariant under the group of the transformations $\widehat{\boldsymbol{\Sigma}} \mapsto \boldsymbol{A}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A}'$; $\boldsymbol{\Sigma} \mapsto \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ ($\boldsymbol{A}$ is $p \times p$ nonsingular)

$$L(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - I_p)^2;\ L_S(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log\det(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - p.$$

But we can not use L_S if $p > n$ as we can not evaluate the expected value of L_S of $n^{-1}\boldsymbol{W}$.

⋆ Ledoit and Wolf (JMVA, 2004) used a loss function $L_F(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) = \mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^2$.

Some remarks (2)

$\star$ $\mathbb{E}[n^{-1}\boldsymbol{W}] = \boldsymbol{\Sigma}$. But the eigenvalues of $n^{-1}\boldsymbol{W}$ are more spread-out than those of $\boldsymbol{\Sigma}$.

$\star$ When $p > n$, $n^{-1}\boldsymbol{W}$ semi-p.d (not invertable!). This makes everything difficult!

$\star$ Under the loss L_S, the eigenvalues of $n^{-1}\boldsymbol{W}$ are shrunk and expanded to obtain improved estimators when $n > p+1$(because of technical reason). See Stein (Prague Symp, 1977), Dey and Srivastava (AOS, 1985), Haff (AOS, 1991) *et al.*

$\star$ To evaluate risk, SURE(Stein's unbiased risk estimate) and eigenvalue-caluculus were worked well. $\rightarrow$ Do these senario work for the case when $p > n$ and the loss L?

$\star$ When $p > n$, Ledoit and Wolf (JMVA, 2004) considered a loss $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^2$ (not invariant!) and obtained asymptotically efficient estimators(roughly speaking) among the linear combinations of $n^{-1}\boldsymbol{W}$ and $\boldsymbol{I}_p$.

Class of estimators

⋆ Decompose $\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'$; $\ell_1 \geq \cdots \geq \ell_n$ are the nonzero eigenvalues of $\boldsymbol{W}$, and

$$\boldsymbol{W} = \boldsymbol{O}_1 \boldsymbol{L} \boldsymbol{O}_1', \qquad \boldsymbol{L} = \mathrm{Diag}(\ell_1, \ldots, l_n);$$

$\boldsymbol{O}_1$ is a $p \times n$ semi-orthogonal s.t. $\boldsymbol{O}_1' \boldsymbol{O}_1 = \boldsymbol{I}_n$.

⋆ $\boxed{\text{Our estimators}}$

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1 \boldsymbol{\Psi}(\boldsymbol{L}) \boldsymbol{O}_1', \tag{1}$$

Here $\boldsymbol{\Psi} := \boldsymbol{\Psi}(\boldsymbol{L}) = \mathrm{Diag}(\psi_1, \psi_2, \ldots, \psi_n)$ and $\psi_k := \psi_k(\boldsymbol{L}) (k = 1, 2, \ldots, n)$ are differentiable functions from $\mathbb{R}^n_{\geq}$ to $\mathbb{R}$.

⋆ $\boxed{\text{Our goal}}$ Evaluate the risk $\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}} \boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2]$.

⋆ Generally the risk depends on $\boldsymbol{\Sigma}$ generally. Hence it is impossible to derive analytical expression of risk for the class of estimators (1).

⋆ Instead, we derive the unbiased risk estimate for the class of estimators (1) and compare the perfomance of estimators via the unbiased risk estimate.

Idea of SURE(Stein's Unbiased risk estimate) method

⋆ For $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1 \boldsymbol{\Psi}(\boldsymbol{L}) \boldsymbol{O}_1'$, derive unbiased estimator $\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}})$ (which depends only on $\psi_1, \ldots, \psi_n$ and $\ell_1, \ldots, \ell_n$) for the risk $\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2]$, i.e.,

$$\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] = \mathbb{E}[\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}})].$$

⋆ As $\mathbb{E}[\mathrm{tr}\,(n^{-1}\boldsymbol{W}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2]$ is constant,

$$\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}}) \le \mathbb{E}[\mathrm{tr}\,(n^{-1}\boldsymbol{W}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] = \text{constant}$$

$$\Downarrow$$

$$\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] \le \mathbb{E}[\mathrm{tr}\,(n^{-1}\boldsymbol{W}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] = \text{constant}.$$

⋆ To derive unbiased risk estimate $\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}})$ for the class (1), we need some technical lemmas.

SURE

$\boxed{\text{Result}}$ Recall that $\boldsymbol{W} = \boldsymbol{O}_1\boldsymbol{L}\boldsymbol{O}_1'$, $\boldsymbol{L} = \mathrm{Diag}(\ell_1, \ldots, l_n)$. For the class of estimators $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1\mathrm{Diag}(\psi_1(\boldsymbol{L}), \ldots, \psi_n(\boldsymbol{L}))\boldsymbol{O}_1'$ and $p > n+1$, we have

$$
\begin{aligned}
R(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) &= \mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] = \mathbb{E}[\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}})] \\
\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}}) &= \sum_{k=1}^{n}\Big\{(p-n-1)\Big(\frac{\widetilde{\psi}_k^{(1)}}{\ell_k} - 2\frac{\psi_k}{\ell_k}\Big) + 2\Big(\frac{\partial\widetilde{\psi}_k^{(1)}}{\partial\ell_k} - 2\frac{\partial\psi_k}{\partial\ell_k}\Big) \\
&\qquad + \sum_{b\neq k}^{n}\frac{(\widetilde{\psi}_k^{(1)} - 2\psi_k) - (\widetilde{\psi}_b^{(1)} - 2\psi_b)}{\ell_k - \ell_b}\Big\} + p.
\end{aligned}
$$

Here, for $k = 1, 2, \ldots, n$,

$$
\widetilde{\psi}_k^{(1)} = \frac{(p-n-1)\psi_k^2}{\ell_k} + 4\psi_k\frac{\partial\psi_k}{\partial\ell_k} + 2\psi_k\sum_{b\neq k}^{n}\frac{\psi_k - \psi_b}{\ell_k - \ell_b}.
$$

Some technical lemmas for derivation of SURE (1)

⋆ Expand the risk for $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1' = \boldsymbol{O}_1\mathrm{Diag}(\psi_1(\boldsymbol{L}), \ldots, \psi_n(\boldsymbol{L}))\boldsymbol{O}_1'$ as

$$\begin{aligned}&\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2]\\&= \mathbb{E}[\mathrm{tr}\,\{\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\}] - 2\mathbb{E}[\mathrm{tr}\,\{\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\}] + p.\end{aligned}$$

⋆ We need unbiased risk estimate for

$$\mathbb{E}[\mathrm{tr}\,\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\}]$$

and

$$\mathbb{E}[\mathrm{tr}\,\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\}].$$

Some technical lemmas for derivation of SURE (2)

⋆ Let $\boldsymbol{X} : n \times p = (x_{ij}) = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n]' \sim N_{n\times p}(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$ and $\boldsymbol{W} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i'$.

⋆ $\boxed{\text{Loh's identity}}$ For $p \times p$ matrix function $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{W}) = (g_{ij}(\boldsymbol{W}))$

$$\mathbb{E}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{G}\right] = \mathbb{E}\left[n\boldsymbol{G} + (\boldsymbol{X}'\nabla_{\boldsymbol{X}})'\boldsymbol{G}\right].$$

⋆ For $\boldsymbol{X} : n \times p = (x_{ij})$, $n \times p$ differential matrix operator $\nabla_{\boldsymbol{X}}$ is defined as

$$\nabla_{\boldsymbol{X}} = \left(\frac{\partial}{\partial x_{ij}}\right)_{\substack{i=1,2,\ldots,n\\ j=1,2,\ldots,p}}; \qquad ((\boldsymbol{X}'\nabla_{\boldsymbol{X}})'\boldsymbol{G})_{ij} = \sum_{k_1=1}^{p}\sum_{k_2=1}^{p} x_{k_1k_2}\frac{\partial g_{k_2j}}{\partial x_{k_1i}}.$$

⋆ Put $\boldsymbol{G} = \boldsymbol{O}_1 \mathrm{Diag}(\ell_1^{-1}\psi_1, \ldots, \ell_n^{-1}\psi_n)\boldsymbol{O}_1'$ in Loh's identity. Recall that our class of estimator is of the form $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1\mathrm{Diag}(\psi_1(\boldsymbol{L}), \ldots, \psi_n(\boldsymbol{L}))\boldsymbol{O}_1'$ where $\boldsymbol{W} = \boldsymbol{O}_1\boldsymbol{L}\boldsymbol{O}_1'$, $\boldsymbol{L} = \mathrm{Diag}(\ell_1, \ldots, l_n)$.

Some technical lemmas for derivation of SURE (3)

$\star$ Put $\boldsymbol{G} = \boldsymbol{O}_1 \mathrm{Diag}(\ell_1^{-1}\psi_1, \ldots, \ell_n^{-1}\psi_n)\boldsymbol{O}_1'$ in Loh's identity

$$\mathbb{E}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{G}] = \mathbb{E}\,[n\boldsymbol{G} + (\boldsymbol{X}'\nabla_{\boldsymbol{X}})'\boldsymbol{G}],$$

$\star$ $\boxed{\text{Stein-Haff identity for } n < p}$ For $\boldsymbol{\Psi} = \mathrm{Diag}(\psi_1, \ldots, \psi_n)$, one has

$$\mathbb{E}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'] = \mathbb{E}\,\left[\boldsymbol{O}_1 \mathrm{Diag}(\psi_1^{(1r)}, \ldots, \psi_n^{(1r)})\boldsymbol{O}_1' + (\boldsymbol{I}_p - \boldsymbol{O}_1\boldsymbol{O}_1')\mathrm{tr}\,(\boldsymbol{L}^{-1}\boldsymbol{\Psi})\right].$$

where

$$\psi_k^{(1r)} = \sum_{b \neq k}^{n} \frac{\psi_k - \psi_b}{\ell_k - \ell_b} + 2\frac{\partial \psi_k}{\partial \ell_k} - \frac{\psi_k}{\ell_k}, \qquad k = 1, 2, \ldots, n.$$

Some technical lemmas for derivation of SURE (4)

⋆ Taking the trace of

$$\mathbb{E}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\right] = \mathbb{E}\left[\boldsymbol{O}_1\boldsymbol{\Psi}^{(1r)}\boldsymbol{O}_1' + (\boldsymbol{I}_p - \boldsymbol{O}_1\boldsymbol{O}_1')\mathrm{tr}\left(\boldsymbol{L}^{-1}\boldsymbol{\Psi}\right)\right],$$

one has Kubokawa and Srivastava (2008)'s identity:

$$\mathbb{E}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\right\}\right] = \mathbb{E}\left[\sum_{k=1}^{n}\left\{(p-n-1)\frac{\psi_k}{\ell_k} + 2\frac{\partial\psi_k}{\partial\ell_k} + \sum_{b\neq k}^{n}\frac{\psi_k-\psi_b}{\ell_k-\ell_b}\right\}\right].$$

Some technical lemmas for derivation of SURE (5)

⋆ Put $\boldsymbol{G} = \boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\mathrm{Diag}(\ell_1^{-1}\psi_1, \ldots, \ell_n^{-1}\psi_n)\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\mathrm{Diag}(\psi_1, \ldots, \psi_n)\boldsymbol{O}_1'$ in Loh's identity $\mathbb{E}\,[\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{G}] = \mathbb{E}\,[n\boldsymbol{G} + (\boldsymbol{X}'\nabla_{\boldsymbol{X}})'\boldsymbol{G}]$.

⋆ For $\boldsymbol{\Psi} = \mathrm{Diag}(\psi_1, \ldots, \psi_n)$,

$$\mathbb{E}[\mathrm{tr}\,\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\}] = \mathbb{E}[\mathrm{tr}\,\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\mathrm{Diag}(\tilde{\psi}_1^{(1)}, \ldots, \tilde{\psi}_n^{(1)})\boldsymbol{O}_1'\}].$$

where $\tilde{\psi}_k^{(1)} = (p-n-1)\frac{\psi_k^2}{\ell_k} + 4\psi_k \cdot \frac{\partial \psi_k}{\partial \ell_k} + 2\psi_k \cdot \sum_{b\neq k}^n \frac{\psi_k - \psi_b}{\ell_k - \ell_b}$ for $k = 1, \ldots, k$.

⋆ Apply Kubokawa and Srivastava's identity for $\tilde{\boldsymbol{\Psi}}^{(1)} = \mathrm{Diag}(\tilde{\psi}_1^{(1)}, \ldots, \tilde{\psi}_n^{(1)})$ to get

$$\mathbb{E}\,[\mathrm{tr}\,\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\tilde{\boldsymbol{\Psi}}^{(1)}\boldsymbol{O}_1'\}] = \mathbb{E}\,\left[\sum_{k=1}^n \left\{(p-n-1)\frac{\tilde{\psi}_k^{(1)}}{\ell_k} + 2\frac{\partial\tilde{\psi}_k^{(1)}}{\partial \ell_k} + \sum_{b\neq k}^n \frac{\tilde{\psi}_k^{(1)} - \tilde{\psi}_b^{(1)}}{\ell_k - \ell_b}\right\}\right].$$

Some technical lemmas for derivation of SURE (6)

⋆ Put

$$\mathbb{E}\left[\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\right\}\right] = \mathbb{E}\left[\sum_{k=1}^{n}\left\{(p-n-1)\frac{\psi_k}{\ell_k} + 2\frac{\partial\psi_k}{\partial\ell_k} + \sum_{b\neq k}^{n}\frac{\psi_k-\psi_b}{\ell_k-\ell_b}\right\}\right]$$

and

$$\begin{aligned}&\mathbb{E}[\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\right\}]\\ &= \mathbb{E}\left[\sum_{k=1}^{n}\left\{(p-n-1)\frac{\tilde{\psi}_k^{(1)}}{\ell_k} + 2\frac{\partial\tilde{\psi}_k^{(1)}}{\partial\ell_k} + \sum_{b\neq k}^{n}\frac{\tilde{\psi}_k^{(1)}-\tilde{\psi}_b^{(1)}}{\ell_k-\ell_b}\right\}\right]\end{aligned}$$

into

$$\begin{aligned}&\mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}-\boldsymbol{I}_p)^2]\\ &= \mathbb{E}[\mathrm{tr}\left\{\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\right\}] - 2\mathbb{E}[\mathrm{tr}\left\{\boldsymbol{O}_1\boldsymbol{\Psi}\boldsymbol{O}_1'\boldsymbol{\Sigma}^{-1}\right\}] + p.\end{aligned}$$

SURE

$\boxed{\text{Result}}$ Recall that $\boldsymbol{W} = \boldsymbol{O}_1\boldsymbol{L}\boldsymbol{O}_1'$, $\boldsymbol{L} = \mathrm{Diag}(\ell_1, \ldots, l_n)$. For the class of estimators $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{O}_1\mathrm{Diag}(\psi_1(\boldsymbol{L}), \ldots, \psi_n(\boldsymbol{L}))\boldsymbol{O}_1'$, we have

$$
\begin{aligned}
R(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}) &= \mathbb{E}[\mathrm{tr}\,(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I}_p)^2] = \mathbb{E}[\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}})] \\
\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}}) &= \sum_{k=1}^{n}\Big\{(p-n-1)\big(\frac{\widetilde{\psi}_k^{(1)}}{\ell_k} - 2\frac{\psi_k}{\ell_k}\big) + 2\big(\frac{\partial\widetilde{\psi}_k^{(1)}}{\partial\ell_k} - 2\frac{\partial\psi_k}{\partial\ell_k}\big) \\
&\qquad + \sum_{b\neq k}^{n}\frac{(\widetilde{\psi}_k^{(1)} - 2\psi_k) - (\widetilde{\psi}_b^{(1)} - 2\psi_b)}{\ell_k - \ell_b}\Big\} + p.
\end{aligned}
$$

Here, for $k = 1, 2, \ldots, n$,

$$
\widetilde{\psi}_k^{(1)} = \frac{(p-n-1)\psi_k^2}{\ell_k} + 4\psi_k\frac{\partial\psi_k}{\partial\ell_k} + 2\psi_k\sum_{b\neq k}^{n}\frac{\psi_k - \psi_b}{\ell_k - \ell_b}.
$$

Best scalar multiple

To get risk for estimators of the form $a\boldsymbol{W}$ where a is a positive constant, apply

$$\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\Sigma}}) = \sum_{k=1}^{n}\Bigg\{(p-n-1)\Big(\frac{\widetilde{\psi}_k^{(1)}}{\ell_k} - 2\frac{\psi_k}{\ell_k}\Big) + 2\Big(\frac{\partial\widetilde{\psi}_k^{(1)}}{\partial\ell_k} - 2\frac{\partial\psi_k}{\partial\ell_k}\Big) + \sum_{b\neq k}^{n}\frac{(\widetilde{\psi}_k^{(1)} - 2\psi_k) - (\widetilde{\psi}_b^{(1)} - 2\psi_b)}{\ell_k - \ell_b}\Bigg\} + p.$$

with $\varphi_k = a\ell_k\,(k = 1, \ldots, n)$. Then

$$R(a\boldsymbol{W}, \boldsymbol{\Sigma}) = np(p+n+1)\left(a - \frac{1}{p+n+1}\right)^2 + \frac{p^2+p}{p+n+1}.$$

$\boxed{\text{Result}}$ $\frac{1}{p+n+1}\boldsymbol{W}$ is the best among estimators of the form $a\boldsymbol{W}$.

Improved estimators

Estimators Assume that $p > n+1$ and consider estimators of the form

$$\widehat{\boldsymbol{\Sigma}}_t = \frac{1}{p+n+1}\Big(\boldsymbol{W} + \frac{t}{\operatorname{tr}\boldsymbol{W}^{+}}\boldsymbol{O}_1\boldsymbol{O}_1'\Big).$$

Here $\boldsymbol{O}_1$ is $p \times n$ semi-orthogonal matrix consisting of column eigenvectors coresponding to the positive eigenvalues of $\boldsymbol{W}$ and $\boldsymbol{W}^{+}$ is the Moore-Penrose inverse of $\boldsymbol{W}$, and t is positve constant.

Results We use SURE and evaluate the risk of $\widehat{\boldsymbol{\Sigma}}_t$ to get the following: If

$$0 < t \le 2(n-1)(p-n-1)/\{(p-n+1)(p-n+2)\},$$

then

$$R(\widehat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}) < R((p+n+1)^{-1}\boldsymbol{W}, \boldsymbol{\Sigma})$$

for all $\boldsymbol{\Sigma}$.

Remark that $R((p+n+1)^{-1}\boldsymbol{W}, \boldsymbol{\Sigma}) < R(n^{-1}\boldsymbol{W}, \boldsymbol{\Sigma})$

Modification of improved estimators

⋆ $\widehat{\boldsymbol{\Sigma}}_t$ improves upon the usual estimator $(p+n+1)^{-1}\boldsymbol{W}$. But $\widehat{\boldsymbol{\Sigma}}_t$ is not p.d.

⋆ Modify $\widehat{\boldsymbol{\Sigma}}_t = \frac{1}{p+n+1}\left(\boldsymbol{W} + \frac{t}{\mathrm{tr}\,\boldsymbol{W}^+}\boldsymbol{O}_1\boldsymbol{O}_1'\right)$ $(0 < t \le \frac{2(n-1)(p-n-1)}{(p-n+1)(p-n+2)})$ to get:

$$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{HF}} = \frac{1}{p+n+1}\{\boldsymbol{W} + \frac{t_0}{\mathrm{tr}\,\boldsymbol{W}^+}\boldsymbol{I}_p\}, \qquad t_0 = \frac{2(n-1)(p+n+1)}{(p-n+1)(p-n+3)}.$$

⋆ Unfortunately, we can not evaluate the risk of the estimators $\widetilde{\boldsymbol{\Sigma}}_{HF}$ by using SURE! But we can see that this estimator is better than the usual estimator via the simulation results we did.

Simulation study (1)

$\boxed{\text{the percentage relative improvement in average loss of } \widehat{\boldsymbol{\Sigma}}_{\mathrm{HF}} \text{ over } \boldsymbol{W}/(n+p+1)}$

$$\mathrm{PRIAL}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{HF}}) = \frac{\text{average loss of } \boldsymbol{W}/(n+p+1) - \text{average loss of } \widehat{\boldsymbol{\Sigma}}_{\mathrm{HF}}}{\text{average loss of } \boldsymbol{W}/(n+p+1)}$$

Simulation study (2)

Table 1: Result of 1000 Monte-Carlo simulations for $p/n = 1/2$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

p	n	$\boldsymbol{W}/(n+p+1)$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{HF}}$	PRIAL
10	5	6.863 (0.013)	6.3465 (0.017)	7.5%
20	10	13.549 (0.014)	12.654 (0.017)	6.6%
40	20	26.891 (0.014)	25.647 (0.016)	4.6%
60	30	40.216 (0.014)	38.821 (0.016)	3.5%
80	40	53.570 (0.014)	52.095 (0.015)	2.8%
100	50	66.894 (0.014)	65.362 (0.015)	2.3%

The values in parentheses refer to the standard error on average loss.

References

[1] Z.D. Bai, Methodologies in spectral analysis of large dimensional random matrices, a reivew. *Statistica Sinica* **9** (1999) 611-677.

[2] Y. Chen, A. Wiesel, Y. Eldar, Shrinkage algorithms for MMSE Covariance Estimation. *IEEE Transactions on Signal Processing* **58** (2010) 5016–5029.

[3] D.K. Dey and C. Srinivasan, Estimation of a covariance matrix under Stein's loss, *Ann. Statist.* **13** (1985) 1581-1591.

[4] M.L. Eaton, Group Invariance Application in Statistics, Regional conference series in Probability and Statistics Vol. 1, Institute of Mathematical Statistics (1989).

[5] L.R. Haff, Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** (1980) 586–597.

[6] L.R. Haff, The variational form of certain Bayes estimators, *Ann. Statist.* **19** (1992) 1163-1190.

[7] W. James and C. Stein, Estimation with quadratic loss, in PROC. FOURTH BERKELEY SYMP. MATH. STATIST. PROB. **1** 361-380, Univ. California Press (1961).

[8] Y. Konno, Shrinkage estimators for large covariance matrices in multivariate real and complex normal distributions under an invriant quadratic loss. *J. Multivariate Anal.* **100** (2009), 2237-2253.

[9] T. Kubokawa and M. Srivastava, Estimation of the precision matrix of s singular Wishart distribution and its application in high dimenstional data. *J. Multivariate Anal.* **99** (2008) 1906-1928.

[10] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices. (English summary) *J. Multivariate Analysis* **88** (2004) 365–411.

[11] R.J. Muirhead, ASPECTS OF MULTIVARIATE STATISTICAL ANALYSIS, John Wiley & Sons, Inc (1982).

[12] C. Stein, Lectures on the theory of estimation of many parameters, in STUDIES IN THE STATISTICAL THEORY OF ESTIMATION I(I. A. Ibragimov and M. S. Nikulin, eds.) (1977).

Thank you!