Condition Number Shrinkage by Joint Distributionally Robust Covariance-Precision Estimation

Renjie Chen, Viet Anh Nguyen, Huifu Xu
Speaker: Renjie Chen
Chinese University of Hong Kong, Department of Systems Engineering and Engineering Management

August, 2025 @ ICSP

Sample Covariance Matrix and Sample precision Matrix

- Let $\xi \in \mathbb{R}^p$ be a random vector with mean zero and **covariance matrix** Σ_0.
- The **precision matrix** (also called inverse covariance matrix) of ξ is Σ_0^{-1}.
- **Sample Covariance Matrix:**

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \xi_i \xi_i^\mathsf{T}$$

  where $\xi_1, \ldots, \xi_n$ are i.i.d. samples of ξ.
- If $\widehat{\Sigma}$ is non-singular, a standard estimator of the precision matrix is $\widehat{\Sigma}^{-1}$.
- However, in high-dimensional and data-deficient case, $\widehat{\Sigma}$ could be ill-conditioned or even singular, which makes the computation of $\widehat{\Sigma}^{-1}$ unstable or not feasible.

Ill-conditioning of the Sample Covariance Matrix: Overestimation and Underestimation of Eigenvalues

- The **condition number** of the sample covariance matrix $\widehat{\Sigma}$ is defined as:

$$\kappa(\widehat{\Sigma}) = \frac{\lambda_{\text{max}}(\widehat{\Sigma})}{\lambda_{\text{min}}(\widehat{\Sigma})}$$

  where $\lambda_{\text{max}}(\widehat{\Sigma})$ and $\lambda_{\text{min}}(\widehat{\Sigma})$ are the largest and smallest eigenvalues of $\widehat{\Sigma}$.

Ill-conditioning of the Sample Covariance Matrix: Overestimation and Underestimation of Eigenvalues

- It is well-known that $\lambda_{\max}(\cdot)$ is convex, the by Jensen's inequality,

$$\mathbb{E}[\lambda_{\max}(\widehat{\Sigma})] \geq \lambda_{\max}(\mathbb{E}[\widehat{\Sigma}]) = \lambda_{\max}(\Sigma_0).$$

  The largest eigenvalue tends to be **overestimated**.

- $\lambda_{\min}(\cdot)$ is concave over $\mathbb{S}^p_{++}$

$$\mathbb{E}[\lambda_{\min}(\widehat{\Sigma})] \leq \lambda_{\min}(\mathbb{E}[\widehat{\Sigma}]) = \lambda_{\min}(\Sigma_0).$$

  The smallest eigenvalue tends to be **underestimated**.

- **Overestimated Condition Number:**

$$\kappa(\widehat{\Sigma}) \text{ is often much larger than } \kappa(\Sigma_0).$$

Ledoit Wolf's Linear shrinkage estimator[1]

- The linear shrinkage estimator

$$S = \rho \nu I + (1-\rho)\widehat{\Sigma},$$

where ρ is the shrinkage intensity and νI is the shrinkage target.

- The optimal parameter is chosen by minimizing the mean squared error of S:

$$\min_{0 \le \rho \le 1, \nu \ge 0} \mathbb{E}\left[\|S - \Sigma_0\|^2\right], \text{ s.t } S = \rho \nu I + (1-\rho)\widehat{\Sigma}$$

- The optimal shrinkage target is $\nu^\star = \frac{\mathrm{tr}(\Sigma_0)}{p}$, and the optimal intensity is $\rho^\star = \frac{\mathbb{E}[\|\widehat{\Sigma}-\Sigma_0\|^2]}{\mathbb{E}[\|\widehat{\Sigma}-\nu^\star I\|^2]}$.
- **Backwards**: Linear structure is not expressive, and shrink too much (too conservative).

[1] Ledoit and Wolf 2004, *"A well-conditioned estimator for large-dimensional covariance matrices"*.

Distributionally Robust Covariance Estimation[2]

- Let $\widehat{\mathbb{P}}_n$ be the empirical distribution of i.i.d. samples $\xi_1, \ldots, \xi_n$.
- The sample covariance matrix is $\widehat{\Sigma} = \mathbb{E}_{\widehat{\mathbb{P}}_n}[\xi\xi^\top]$.
- Yue et al. proposed to view $\widehat{\Sigma}$ as the unique solution to the minimization problem

$$\widehat{\Sigma} = \arg\min_{\Sigma \in \mathbb{S}_+^p} \|\Sigma\|_F^2 - 2\langle \Sigma, \widehat{\Sigma} \rangle = \arg\min_{\Sigma \in \mathbb{S}_+^p} \underbrace{\|\Sigma\|_F^2 - 2\mathbb{E}_{\widehat{\mathbb{P}}_n}[\langle \Sigma, \xi\xi^\top \rangle]}_{\triangleq \mathrm{FrobeniusLoss}(\Sigma, \widehat{\mathbb{P}}_n)}.$$

[2] Yue et al. 2024, *"A Geometric Unification of Distributionally Robust Covariance Estimators: Shrinking the Spectrum by Inflating the Ambiguity Set"*.

Distributionally Robust Covariance Estimation

- A moment-based ambiguity set centered at the nominal distribution $\widehat{\mathbb{P}}_n$ with radius $\varepsilon > 0$:

$$\mathcal{P}_\varepsilon(\widehat{\mathbb{P}}_n) \triangleq \left\{ \mathbb{Q} : \mathbb{E}_{\mathbb{Q}}[\xi] = 0, \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] = S, D\left(S, \mathbb{E}_{\widehat{\mathbb{P}}_n}[\xi\xi^\top]\right) \leq \varepsilon \right\},$$

  where D is a divergence in the space of positive semidefinite matrices to measure the discrepancy between two matrices.
- The distributionally robust covariance estimation model:

$$\min_{\Sigma \in \mathbb{S}^p_{++}} \max_{\mathbb{Q} \in \mathcal{P}_\varepsilon(\widehat{\mathbb{P}}_n)} \|\Sigma\|_F^2 - 2\mathbb{E}_{\widehat{\mathbb{P}}_n}[\langle \Sigma, \xi\xi^\top \rangle]$$

Distributionally Robust Covariance Estimation

- Under certain assumption on choice of divergence D, the distributionally robust covariance estimation model reduces to

$$\min_{D(S,\widehat{\Sigma})\leq\varepsilon} \|S\|_F^2.$$

- It shrinks all the eigenvalues towards 0, i.e., the shrinkage target of covariance matrix is zero matrix.
- The shrinkage intensity is decided by ε. The larger ε, the more shrinkage.
- **Backwards**: (1) shrinkage target 0 is not well-conditioned; (2) it underestimates $\lambda_{\min}$ even worse.

Distributionally Robust Precision Estimation[3]

- If $\widehat{\Sigma}$ is non-singular, the sample precision matrix estimator can be viewed as the optimal solution of:

$$\widehat{X} = \arg\min_{X \in \mathbb{S}^p_{++}} -\log\det X + \langle X, \widehat{\Sigma} \rangle = \arg\min_{X \in \mathbb{S}^p_{++}} \underbrace{-\log\det X + \mathbb{E}_{\widehat{\mathbb{P}}_n}[\langle X, \xi\xi^\top \rangle]}_{\triangleq \mathrm{SteinLoss}(X, \mathbb{P}_n)}.$$

- The optimal solution is $\widehat{X} = \widehat{\Sigma}^{-1}$.

[3] Nguyen, Kuhn, and Mohajerin Esfahani 2022, *"Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator"*.

Distributionally Robust Precision Estimation

- Nguyen et al. consider the distributionally robust precision estimation model:

$$\min_{X \in \mathbb{S}^p_{++}} \; -\log\det X + \max_{\mathbb{Q} \in \mathcal{P}_\varepsilon(\widehat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\langle X, \xi\xi^\mathsf{T} \rangle].$$

- The resulting distributionally robust precision matrix estimator is also a shrinkage estiamtor.
- It shrinks all the eigenvalues of precision matrix towards 0, i.e., the shinkage target of precision matrix is zero matrix.
- It is equivalent to shrink the eigenvalues of covariance matrix towards $+\infty$.
- The shrinkage intensity is also control by radius ε.
- **Backwards**: It overestimates $\lambda_{\max}$ even worse.

Distributionally Robust Covariance-precision Estimation

- Consider the joint covariance-precision matrix estimation model:

$$\begin{aligned}\min_{\Sigma, X} \ & \mathrm{SteinLoss}(X, \widehat{\mathbb{P}}_n) + \frac{\tau}{2}\mathrm{FrobeniusLoss}(\Sigma, \widehat{\mathbb{P}}_n) \\ \text{s.t. } \ & \Sigma \in \mathbb{S}^p_+, \ X \in \mathbb{S}^p_{++}, \ X\Sigma = I.\end{aligned}$$

When $\widehat{\Sigma}$ is non-singular, the optimal solution is $\left(\Sigma^\star = \widehat{\Sigma}, X^\star = \widehat{\Sigma}^{-1}\right)$.

Distributionally Robust Covariance-precision Estimation

- Consider the moment-based ambiguity set centered at the nominal distribution $\widehat{\mathbb{P}}_n$ with radius $\varepsilon > 0$ again:

$$\mathcal{P}_\varepsilon(\widehat{\mathbb{P}}_n) \triangleq \left\{ \mathbb{Q} : \mathbb{E}_\mathbb{Q}[\xi] = 0, \mathbb{E}_\mathbb{Q}[\xi\xi^\top] = S, D\left(S, \mathbb{E}_{\widehat{\mathbb{P}}_n}[\xi\xi^\top]\right) \leq \varepsilon \right\}.$$

- Define the pair of distributionally robust covariance-precision matrix estimator $(\Sigma^\star, X^\star)$ by

$$(\Sigma^\star, X^\star) \triangleq \arg \min_{\substack{\Sigma, X \in \mathbb{S}^p_{++} \\ X\Sigma = I}} \max_{\mathbb{Q} \in \mathcal{P}_\varepsilon(\widehat{\mathbb{P}}_n)} -\log\det X + \mathbb{E}_\mathbb{Q}[\langle X, \xi\xi^\top \rangle] + \frac{1}{2}\tau \left(\|\Sigma\|_F^2 - 2\mathbb{E}_\mathbb{Q}[\langle \Sigma, \xi\xi^\top \rangle] \right) \tag{DRO}$$

Distributionally Robust Covariance-precision Estimation

- Note that the objective function depends on $\mathbb{Q}$ only through term $\mathbb{E}_{\mathbb{Q}}[\xi\xi^\mathsf{T}]$.
- The DRO model reduces to the following robust optimization (RO) problem:

$$(\Sigma^\star, X^\star) = \arg \min_{\substack{\Sigma, X \in \mathbb{S}^p_{++} \\ X\Sigma = I}} \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \le \varepsilon} \left\{ -\log\det X + \langle X, S \rangle + \frac{1}{2}\tau \left(\|\Sigma\|_F^2 - 2\langle \Sigma, S \rangle \right) \right\}. \tag{RO}$$

- Question:
- The constraint of (RO) is not convex. How can we solve it?
- Do we get a condition-number-shrunk optimal solution pair?
- If yes, what is the shrinkage target and intensity?
- How to choose parameter τ and ε?

Solvable max-min problem

- Interchange the order of min-max of (RO) gives

$$\max_{D(S,\widehat{\Sigma})\leq\varepsilon}\begin{pmatrix}\min_{X,\Sigma} & \langle\widehat{\Sigma},X\rangle-\log\det X+\frac{1}{2}\tau\left(\|\Sigma\|_F^2-2\langle\Sigma,S\rangle\right)\\ \text{s.t.} & X\in\mathbb{S}^p_{++},\ \Sigma\in\mathbb{S}^p_{+},\ X\Sigma=I\end{pmatrix}.$$

- The inner min problem is solved by $\Sigma^\star=(X^\star)^{-1}=S$, when S is non-singular. Then it becomes

$$\max_{S\in\mathbb{S}^p_+:D(S,\widehat{\Sigma})\leq\varepsilon}\ \log\det S-\frac{1}{2}\tau\|S\|_F^2. \tag{P-Mat}$$

- The following saddle-point theorem shows that the interchanging of min-max keeps the problem nature, i.e., optimal solution of (P-Mat) solves (RO).

Saddle Point Theorem

Theorem 1

Suppose that

1. $D(\cdot, \widehat{\Sigma})$ *is convex, differentiable and* $\{S \in \mathbb{S}^p_{++} : D(S, \widehat{\Sigma}) \leq \varepsilon\}$ *is compact,*
2. $(\widehat{\Sigma}, \widehat{\Sigma}) \in \mathrm{dom}(D)$, *and*
3. *Slater's condition holds, i.e., there exists* $S \in \mathbb{S}^p_+$ *such that* $D(S, \widehat{\Sigma}) < \varepsilon$.

Let $\tau > 0$ *and* $S^\star$ *be an optimal solution defined as*

$$S^\star = \arg \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \leq \varepsilon} \log\det S - \frac{1}{2}\tau \|S\|_F^2. \tag{P-Mat}$$

Then the tuple $(\Sigma = S^\star, X = (S^\star)^{-1}, S = S^\star)$ *is an optimal solution of* (RO).

Unbinding Constraint Case

Now we deal with the solvability of (P-Mat).

$$S^\star = \arg \max_{S \in \mathbb{S}_+^p : D(S, \widehat{\Sigma}) \leq \varepsilon} \log\det S - \frac{1}{2}\tau \|S\|_F^2. \tag{P-Mat}$$

We first consider a relatively straightforward case where the constraint is unbinding.

Proposition 1

If $\varepsilon \geq \varepsilon_{\max} \triangleq D\left(\sqrt{\frac{1}{\tau}} I, \widehat{\Sigma}\right)$, then $S^\star = \sqrt{\frac{1}{\tau}} I$ is the optimal solution to (P-Mat).

In the following discussion, we assume that $0 < \varepsilon < \varepsilon_{\max}$, and show that under some assumption on choice of divergence D, (P-Mat) can be solved in quasi-closed form.

Assumption 1: Convex Spectral Divergence D I

The divergence function $D : \mathbb{S}^p_+ \times \mathbb{S}^p_+ \to \mathbb{R}_+$ is non-negative, continuous, and satisfies the identity of indiscernibles, that is, for any $X, Y \in \operatorname{dom}(D)$ we have $D(X, Y) = 0$ if and only if $X = Y$. For any $Y \in \operatorname{dom}(D)$, $D(\cdot, Y)$ is convex, differentiable and for any $Y \in \operatorname{dom}(D), \varepsilon > 0$ the set $\{S \in \mathbb{S}^p_{++} : D(S, Y) \leq \varepsilon\}$ is compact. In addition, D satisfies the following structural conditions.

(i) (Orthogonal equivariance) For any $X, Y \in \mathbb{S}^p_+$ and orthogonal matrix V, we have $D(X, Y) = D(VXV^\mathsf{T}, VYV^\mathsf{T})$.

(ii) (Spectrality) There exists a function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$D(\operatorname{diag}(x), \operatorname{diag}(y)) = \sum_{i=1}^{p} d(x_i, y_i), \forall x, y \in \mathbb{R}^p_+$$

and $d(a, b)$ is twice continuously differentiable in a for every $b \geq 0$. In the following, we refer to d as the generator of D.

Assumption 1: Convex Spectral Divergence D II

(iii) (Rearrangement property) For any $x, y \in \mathbb{R}_+^p$ and $V \in \mathcal{O}(p)$ we have

$$D(V\mathrm{diag}(x^{\uparrow})V^{\mathsf{T}}, \mathrm{diag}(y^{\uparrow})) \geq D(\mathrm{diag}(x^{\uparrow}), \mathrm{diag}(y^{\uparrow})).$$

If its left side is finite, this inequality becomes an equality if and only if $V\mathrm{diag}(x^{\uparrow})V^{\mathsf{T}} = \mathrm{diag}(x^{\uparrow})$.

Remark: Consider $(aI, bI) \in \mathrm{dom}(D)$. Then by (ii), the domain of d is

$$\mathrm{dom}(d) = \{(a, b) \in \mathbb{R}_+^2 : (aI, bI) \in \mathrm{dom}(D)\},$$

and

$$D(aI, bI) = p \times d(a, b).$$

It implies that d inherits non-negativity, continuity, identity of indiscernibles, and convexity from D. Then we conclude that $d(b, b) = 0$, $\frac{\partial d}{\partial a}(b, b) = 0$ and thus b is the unique minimizer of the function $d(\cdot, b)$ for any $b > 0$.

Reduction of (P-Mat) to a Vector Space Problem

- Let $\widehat{\Sigma} = \widehat{V}\mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)\widehat{V}^\mathsf{T}$ be the spectral decomposition of $\widehat{\Sigma}$.
- Under Assumption 1, we are able to show that (P-Mat) can be reduced to a problem that optimizes over all vectors in the non-negative orthant $\mathbb{R}^p_+$:

$$\begin{aligned} \max_{s_i \geq 0} \quad & \sum_{i=1}^{p} \log s_i - \frac{1}{2}\tau \sum_{i=1}^{p} s_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^{p} d(s_i, \widehat{\lambda}_i) \leq \varepsilon. \end{aligned} \tag{P-Vec}$$

Proposition 2 (Equivalence of (P-Mat) and (P-Vec))

If Assumption 1 holds, then problem (P-Mat) *is equivalent to* (P-Vec) *in the following sense.*

(i) *If $s^\star$ solves* (P-Vec), *then* $\widehat{V}\mathrm{diag}(s^\star)\widehat{V}^\mathsf{T}$ *solves* (P-Mat).

(ii) *If $S^\star$ solves* (P-Mat), *then the vector of its eigenvalues $\lambda(S^\star)$ solves* (P-Vec).

Quasi-closed Form of Solution of (P-Vec)

- By the standard Lagrange dual approach, the dual of (P-Vec) is equivalent to

$$\min_{\gamma\geq 0}\max_{s_i\geq 0} \sum_{i=1}^{p} \log s_i - \frac{1}{2}\tau\sum_{i=1}^{p} s_i^2 - \gamma\left(\sum_{i=1}^{p} d(s_i,\widehat{\lambda}_i) - \varepsilon\right).$$

- Define a solution mapping $\varphi : \mathbb{R}_+^3 \to \mathbb{R}$ via

$$\varphi(\tau,\gamma,b) = \text{the unique solution } a^\star > 0 \text{ of the equation } 0 = \frac{1}{a^\star} - \tau a^\star - \gamma\frac{\partial d}{\partial a}(a^\star, b).$$

- For fixed γ, the inner maximization problem is solved by

$$s_i^\star = \varphi(\tau,\gamma,\widehat{\lambda}_i).$$

- By complementary slackness condition, the optimal Lagrange multipler $\gamma^\star$ must satisfies

$$\sum_{i=1}^{p} d(\varphi(\tau,\gamma^\star,\widehat{\lambda}_i),\widehat{\lambda}_i) - \varepsilon = 0.$$

Quasi-closed Form of Solution of (P-Vec)

Proposition 3 (Solution of (P-Vec))

If $\sum_{i=1}^{p} d\left(\sqrt{\frac{1}{\tau}}, \widehat{\lambda}_i\right) > \varepsilon$, *then* (P-Vec) *admits a unique optimal solution* $s^\star$ *and*

$$s_i^\star = \varphi(\tau, \gamma^\star, \widehat{\lambda}_i),$$

where $\gamma^\star$ *is the unique solution of the nonlinear equation*

$$\sum_{i=1}^{p} d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \varepsilon = 0.$$

Uniqueness of $\gamma^\star$

- Define $F : \mathbb{R}_+ \to \mathbb{R}$ through $F(\gamma) = \sum_{i=1}^{p} d(\varphi(\tau, \gamma, \widehat{\lambda}_i), \widehat{\lambda}_i)$.

Proposition 4 (Differentiable and strictly decreasing F)

The function F is differentiable and strictly decreasing over $\mathbb{R}_+$. If $\varepsilon < \varepsilon_{\max} \triangleq \sum_{i=1}^{p} d\left(\sqrt{\frac{1}{\tau}}, \widehat{\lambda}_i\right)$, then $\lim_{\gamma \downarrow 0} F(\gamma) > \varepsilon$ and $\lim_{\gamma \to \infty} F(\gamma) < \varepsilon$.

The proposition reveals that

- $F(\gamma) = \varepsilon$ admits a unique positive root,
- the equation can be solved efficiently by bisection or Newton's method.

Construction of Distributionally Robust Covariance-precision Estimators

Taking all steps so far:

(DRO) $\Leftrightarrow$ (RO) $\Leftrightarrow$ (P-Mat) $\Leftrightarrow$ (P-Vec) $\Rightarrow$ quasi-closed form solution.

- Recall that $\widehat{\Sigma} = \widehat{V}\mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)\widehat{V}^\top$ is the spectral decomposition of $\widehat{\Sigma}$.
- The theorem reveals how to construct the distributionally robust covariance-precision estimator.

Theorem 2 (Construction of covariance estimator)

Under certain assumptions, the optimal solution of (RO) *is given by* $(\Sigma^\star = \widehat{V}\Phi(\tau, \gamma^\star, \widehat{\lambda})\widehat{V}^\top, X = \Sigma^{\star -1})$, *where*

$$\Phi(\tau, \gamma^\star, \widehat{\lambda}) \triangleq \mathrm{diag}(\varphi(\tau, \gamma^\star, \widehat{\lambda}_1), \ldots, \varphi(\tau, \gamma^\star, \widehat{\lambda}_p)),$$

where $\gamma^\star$ *is the unique positive root of the equation* $\sum_{i=1}^{p} d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \varepsilon = 0$.

Proof Structure of Theorem 2

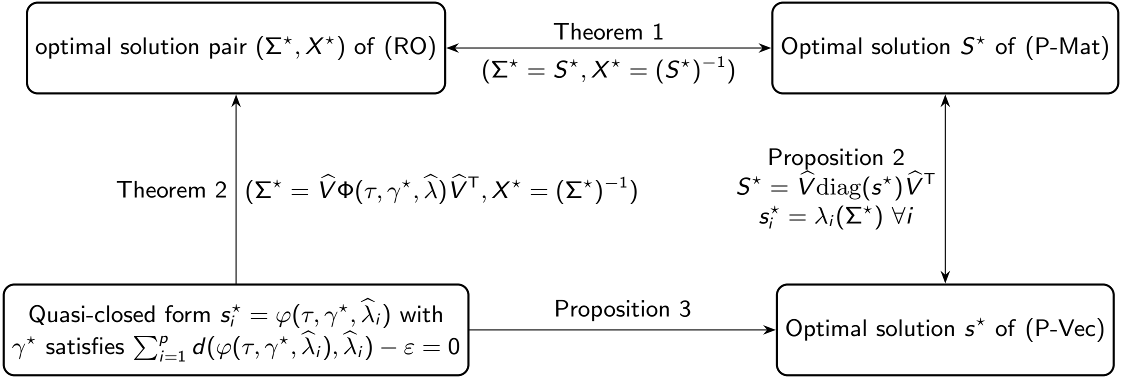

Figure 1: Structure of the proof of Theorem 2.

Nonlinear Shrinkage Estimator

- Define the covariance matrix estimator by

$$\Sigma^\star(\tau, \varepsilon) \triangleq \widehat{V} \mathrm{diag}(\varphi(\tau, \gamma_{\widehat{\lambda}}(\varepsilon), \widehat{\lambda}_1), \ldots, \varphi(\tau, \gamma_{\widehat{\lambda}}(\varepsilon), \widehat{\lambda}_p)) \widehat{V}^\mathsf{T}, \varepsilon \in (0, \varepsilon_{\max}], \tag{1}$$

where $\gamma_{\widehat{\lambda}}(\varepsilon)$ denote the unique root of $\sum_{i=1}^p d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \varepsilon = 0$.

Nonlinear Shrinkage and Improved Condition Number

Theorem 3

If Assumptions hold, then for any $\tau > 0$, $\Sigma^\star(\tau, \varepsilon)$ *is continuous over* $\varepsilon \in (0, \varepsilon_{\max}]$ *and* $\lim_{\varepsilon \to 0} \Sigma^\star(\tau, \varepsilon) = \widehat{\Sigma}$, $\Sigma^\star(\tau, \varepsilon_{\max}) = \sqrt{\frac{1}{\tau}} I$. *Moreover, if* $\lambda_{\min}(\widehat{\Sigma}) < \sqrt{\frac{1}{\tau}} < \lambda_{\max}(\widehat{\Sigma})$, *then we have*

(i) $\lambda_{\min}(\widehat{\Sigma}) < \lambda_{\min}(\Sigma^\star(\varepsilon)) < \sqrt{\frac{1}{\tau}} < \lambda_{\max}(\Sigma^\star(\varepsilon)) < \lambda_{\max}(\widehat{\Sigma})$ *for any* $\varepsilon \in (0, \varepsilon_{\max}]$,

(ii) $\kappa(\Sigma^\star(\varepsilon))$ *is strictly decreasing on* ε, *and thus* $\kappa(\Sigma^\star(\varepsilon)) < \kappa(\widehat{\Sigma})$ *for any* $\varepsilon \in (0, \varepsilon_{\max}]$.

- The proposed covariance-precision matrix estimator can be interpreted as a **nonlinear shrinkage estimator** with
  - $\sqrt{\frac{1}{\tau}} I$ is the shrinkage target,
  - ε controls the shrinkage intensity.
- The nonlinear shrinkage improves the condition number strictly.

Convex Spectral Divergences

Theorem 4

All divergences in Table 1 satisfy the Convex Spectral Divergence Assumption.

Divergence function	$D(\Sigma, \widehat{\Sigma})$	$d(a, b)$	Domain of D
Kullback-Leibler	$\frac{1}{2}\left(\mathrm{tr}(\widehat{\Sigma}^{-1}\Sigma) - p + \log\det(\widehat{\Sigma}\Sigma^{-1})\right)$	$\frac{1}{2}\left(\frac{a}{b} - 1 - \log\frac{a}{b}\right)$	$\mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$
Wasserstein	$\mathrm{tr}\left(\Sigma + \widehat{\Sigma} - 2(\Sigma^{\frac{1}{2}}\widehat{\Sigma}\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\right)$	$a + b - 2\sqrt{ab}$	$\mathbb{S}^p_{+} \times \mathbb{S}^p_{+}$
Symmetrized Stein	$\frac{1}{2}\left(\mathrm{tr}(\widehat{\Sigma}^{-1}\Sigma) + \mathrm{tr}(\Sigma^{-1}\widehat{\Sigma}) - 2p\right)$	$\frac{1}{2}\left(\frac{b}{a} + \frac{a}{b} - 2\right)$	$\mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$
Squared Frobenius	$\mathrm{tr}\left((\Sigma - \widehat{\Sigma})^2\right)$	$(a-b)^2$	$\mathbb{S}^p_{+} \times \mathbb{S}^p_{+}$
Weighted Frobenius	$\mathrm{tr}\left((\Sigma - \widehat{\Sigma})\widehat{\Sigma}^{-1}\right)$	$\frac{(a-b)^2}{b}$	$\mathbb{S}^p_{+} \times \mathbb{S}^p_{++}$

Table 1: Divergence functions and their generators.

Solution Mapping and Upper Bound of Dual Variable

Taking D as any divergence in Table 1, we can construct the optimal solution of (RO) using Theorem 2. The key steps are

- to find out the solution mapping φ as function of τ, γ and b,
- to solve the equation $\sum_{i=1}^{p} d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \varepsilon = 0$ to find out dual variable $\gamma^\star$.

Solution Mapping and Upper Bound of Dual Variable

Divergence function	Solution mapping $\varphi(\tau, \gamma, b)$	Upper bound of $\gamma^\star$
Kullback-Leibler	$\frac{-\gamma+\sqrt{\gamma^2+8\tau(2+\gamma)b^2}}{4\tau b}$	$\max\left\{\frac{2p}{\varepsilon}, \frac{2\tau\widehat{\lambda}_p^2+1}{e^{\varepsilon/p}-1}\right\}$
Wasserstein	unique positive root a of $\tau a^2 + \gamma a - \gamma\sqrt{b}\sqrt{a} - 1 = 0$	$\sqrt{\max\left\{\frac{p(1-\tau\widehat{\lambda}_1^2)^2}{\varepsilon\widehat{\lambda}_1}, \frac{p(1-\tau\widehat{\lambda}_p^2)^2}{\varepsilon\widehat{\lambda}_1}\right\}}$
Symmetrized Stein	unique positive root a of $2\tau ba^3 + \gamma a^2 - 2ba - \gamma b^2 = 0$	$\sqrt{\max\left\{\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\varepsilon\widehat{\lambda}_1^4}, \frac{p(1-\tau\widehat{\lambda}_1^2)^2}{4\varepsilon\widehat{\lambda}_1^4}\right\}}$
Squared Frobenius	$\frac{\gamma b}{\tau+2\gamma} + \frac{\sqrt{\gamma^2 b^2+\tau+2\gamma}}{\tau+2\gamma}$	$\sqrt{\max\left\{\frac{p(1-\tau\widehat{\lambda}_1^2)^2}{4\varepsilon\widehat{\lambda}_1^2}, \frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\varepsilon\widehat{\lambda}_1^2}\right\}}$
Weighted Frobenius	$\frac{\gamma b}{\tau b+2\gamma} + \frac{\sqrt{\gamma^2 b^2+b(\tau b+2\gamma)}}{\tau b+2\gamma}$	$\sqrt{\max\left\{\frac{p\widehat{\lambda}_p(1-\tau\widehat{\lambda}_1^2)^2}{4\varepsilon\widehat{\lambda}_1^2}, \frac{p\widehat{\lambda}_p(1-\tau\widehat{\lambda}_p^2)^2}{4\varepsilon\widehat{\lambda}_1^2}\right\}}$

Table 2: Solution mapping and upper bound of dual variable. $\widehat{\lambda}_1$ and $\widehat{\lambda}_p$ are the smallest and largest eigenvalues of $\widehat{\Sigma}$.

Parameter Tuning: Minimizing MSE not Working

- Our estimator $\Sigma^\star(\tau, \varepsilon)$ is parameterized by τ and ε, where τ determine the shrinkage target and ε is the shrinkage intensity.
- In Ledoit-Wolf's linear shrinkage estimator, they minimize MSE of the proposed estimator to find optimal parameters.
- If we borrow the same idea , we could solve

$$\min_{\tau \in \mathbb{R}_{++}, \varepsilon \in \mathbb{R}_+} \mathbb{E}[\|\Sigma^\star(\tau, \varepsilon) - \Sigma_0\|_F^2]. \tag{2}$$

- However, the dependence of $\Sigma^\star(\tau, \varepsilon)$ on τ and ε is nonlinear. It makes (2) hard to analyze.
- From another perspective, we hope the optimal solution of (P-Mat) is close to Σ_0.

Optimal Shrinkage Target

- In the first step, we hope the shrinkage target is near to Σ_0.
- Recall that the objective of (P-Mat) is

$$\ell_\tau(S) \triangleq \log\det S - \frac{1}{2}\tau\|S\|_F^2,$$

and the shrinkage target $\sqrt{\frac{1}{\tau}}I$ is the optimal solution of

$$\min_{S\in\mathbb{S}^p_{++}} \ \ell_\tau(S).$$

- We choose τ such that Σ_0 can nearly solve $\min_{S\in\mathbb{S}^p_{++}} \ \ell_\tau(S)$ (and thus close to $\sqrt{\frac{1}{\tau}}I$).
- We minimize the gradient norm of $\ell_\tau(S)$ evaluated at $S=\Sigma_0$:

$$\min_{\tau\in\mathbb{R}_{++}} \|\nabla_S \ell_\tau(S)\big|_{S=\Sigma_0}\|_F = \min_{\tau\in\mathbb{R}_{++}} \|\Sigma_0^{-1} - \tau\Sigma_0\|_F.$$

Optimal Shrinkage Target

- The above problem is solved by $\tau = \tau^\star \triangleq \frac{p}{\|\Sigma_0\|_F^2}$, and the induced shrinkage target is

$$S = \frac{\|\Sigma_0\|_F}{\sqrt{p}} I.$$

- Note that $\left\| \frac{\|\Sigma_0\|_F}{\sqrt{p}} I \right\|_F = \|\Sigma_0\|_F$, and thus the target can be viewed as a uniform prior with the same scale as Σ_0, i.e., the total scale is distributed evenly to each dimension.

Optimal Shrinkage Intensity I

- With the optimal target chosen as above, we tune the radius ε so that the true covariance Σ_0 is near the optimal solution of

$$S^\star_{\tau^\star} \triangleq \arg \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}_n) \le \varepsilon} \log\det S - \frac{1}{2}\tau^\star \|S\|_F^2. \tag{P-Mat-$\tau^\star$}$$

- When the constraint $D(S, \widehat{\Sigma}_n) \le \varepsilon$ is active, the first-order optimality condition of (P-Mat-$\tau^\star$) dictates that $S^\star_{\tau^\star}$ and the dual variable $\gamma^\star_{\tau^\star}$ satisfies

$$(S^\star_{\tau^\star})^{-1} - \tau^\star S^\star_{\tau^\star} - \gamma^\star_{\tau^\star} D'(S^\star_{\tau^\star}, \widehat{\Sigma}_n) = 0, \tag{3}$$

$$D(S^\star_{\tau^\star}, \widehat{\Sigma}_n) - \varepsilon = 0, \tag{4}$$

$$\gamma^\star_{\tau^\star} > 0,\ S^\star_{\tau^\star} \in \mathbb{S}^p_+. \tag{5}$$

Optimal Shrinkage Intensity II

- Recall that $S = \Sigma^\star(\tau^\star, \varepsilon)$ uniquely solves (3)-(5). By the construction of $\Sigma_n^\star(\tau^\star, \varepsilon)$, equation (4) becomes
$$\sum_{i=1}^{p} d(\varphi(\tau^\star, \gamma, \widehat{\lambda}_i), \widehat{\lambda}_i) - \varepsilon = 0.$$
- Then the optimal solution of (P-Mat-$\tau^\star$) solves
$$S^{-1} - \tau^\star S - \gamma_{\widehat{\lambda}}(\varepsilon) D'(S, \widehat{\Sigma}_n) = 0. \tag{6}$$
We call the left hand side of (6) as the *extended* gradient of problem (P-Mat-$\tau^\star$).
- To measure the optimality of S on average, we define the average extended gradient by
$$g(S) \triangleq S^{-1} - \tau^\star S - \gamma_\lambda(\varepsilon) \mathbb{E}_n[D'(S, \widehat{\Sigma}_n)].$$

Optimal Shrinkage Intensity III

- Then, again, to choose ε so that Σ_0 can nearly solve (P-Mat-$\tau^\star$) on average and thus be close to $\Sigma_n^\star(\tau^\star, \varepsilon)$, we minimize the norm of average extended gradient evaluated at $S = \Sigma_0$:

$$\varepsilon_n^\star = \arg\min_{\varepsilon > 0} \|g(\Sigma_0)\|_F^2 = \arg\min_{\varepsilon > 0} \|\Sigma_0^{-1} - \tau^\star \Sigma_0 - \gamma_\lambda(\varepsilon)\mathbb{E}_n D'(\Sigma_0, \widehat{\Sigma}_n)\|_F^2. \tag{7}$$

- Though the optimal intensity defined by (7) is not achievable, since we have no access to precise evaluation of Σ_0 and $\mathbb{E}_n D'(\Sigma_0, \widehat{\Sigma}_n)$, it is possible for us to characterize the limit behavior of $\varepsilon_n^\star$.

Assumptions

Assumption 1 (Locally-quadratic divergence)

For any $b > 0$, there exists positive constant $C_{b,d}$ that depends on the divergence d and the limit point b such that

$$\lim_{a \to b} \frac{d(a, b)}{(a - b)^2} = C_{b,d}.$$

- All the divergences in Table 1 satisfy Assumption 1.

Assumptions

Assumption 2 (Non-degenerating divergence gradient)

For all integers n sufficiently large, the derivative of divergence D satisfies $\mathbb{E}_n[D'(\Sigma_0, \widehat{\Sigma}_n)] \neq 0$ and there exists strictly positive C_1 and C_2 such that

$$\lim_{n\to\infty} n\|\mathbb{E}_n[D'(\Sigma_0, \widehat{\Sigma}_n)]\|_F = C_1$$

and

$$\lim_{n\to\infty} n\left\langle \Sigma_0^{-1} - \tau^\star\Sigma_0, \mathbb{E}_n[D'(\Sigma_0, \widehat{\Sigma}_n)]\right\rangle = C_2.$$

- Frobenius divergence does not satisfy the assumption, since $D(\Sigma_0, \Sigma_n) = \|\Sigma_0 - \Sigma_n\|_F^2$, and

$$\mathbb{E}[D'(\Sigma_0, \Sigma_n)] = \mathbb{E}[2(\Sigma_0 - \Sigma_n)] = 0.$$

$1/n^2$-order optimal radius

Theorem 5 ($1/n^2$-order optimal radius)

Under the above assumptions, the optimal intensity $\varepsilon_n^\star$ defined as (7) satisfies

$$\lim_{n\to\infty} n^2 \varepsilon_n^\star = \varepsilon^\star,$$

where the limit $\varepsilon^\star$ depends on divergence D and the underlying data-generating distribution.

Corollary: Optimal Shrinkage Intensity of Convex Divergences I

Let the optimal radius $\varepsilon_n^\star$ be defined as (7). Then under Assumptions, we have

(i) if D is taken as Kullback-Leibler divergence, then

$$\lim_{n\to\infty} n^2 \varepsilon_n^\star = \frac{(p+1)^2 \|\Sigma_0^{-1}\|_F^4}{16\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2} \sum_{i=1}^{p} \left(1 - \tau^\star \lambda_i{}^2\right)^2, \tag{8}$$

(ii) if D is taken as symmetrized Stein divergence, then

$$\lim_{n\to\infty} n^2 \varepsilon_n^\star = \frac{(p+1)^2 \|\Sigma_0^{-1}\|_F^4}{32\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2} \sum_{i=1}^{p} \left(1 - \tau^\star \lambda_i{}^2\right)^2, \tag{9}$$

Corollary: Optimal Shrinkage Intensity of Convex Divergences II

(iii) if D is taken as Wasserstein divergence, then

$$\lim_{n\to\infty} n^2 \varepsilon_n^\star = \frac{(p+1)^2 p^2}{256\left(\mathrm{tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\mathrm{tr}(\Sigma_0)\right)^2} \sum_{i=1}^{p} \frac{(1-\tau^\star\lambda_i)^2}{\lambda_i}. \tag{10}$$

Numerical Test Setup

- **Data Generation**
  - True covariance
    $$\Sigma_0 = V^\top \Lambda V \in \mathbb{S}^p_+, \quad \Lambda = \mathrm{diag}(1, 2, \ldots, p),$$
    with V drawn uniformly from the orthogonal group. Thus $\kappa(\Sigma_0) = p$.
  - Draw n samples from $\mathcal{N}(0, \Sigma_0)$ and compute the sample covariance $\widehat{\Sigma}$.
- **Estimators under Comparison**
  $$\mathrm{W}(\widehat{\Sigma}), \quad \mathrm{L}(\widehat{\Sigma}), \quad \mathrm{NL}(\widehat{\Sigma}),$$
  denoting the Wasserstein-based nonlinear shrinkage estimator, Ledoit–Wolf linear estimator, and Ledoit–Wolf nonlinear estimator, respectively.

Performance Metric & Hypothesis Tests

- **Condition Number Error**

$$\varepsilon(\mathsf{M}) = \mathbb{E}\Big[\Big|\kappa(\mathsf{M}(\widehat{\Sigma})) - \kappa(\Sigma_0)\Big|\Big], \quad \mathsf{M} \in \{\mathsf{W}, \mathsf{L}, \mathsf{NL}\}.$$

- **Hypothesis Tests** (paired t-test, $\alpha = 0.05$)

$$H^0_{\mathsf{L}} : \varepsilon(\mathsf{W}) \geq \varepsilon(\mathsf{L}) \quad \text{vs} \quad H^1_{\mathsf{L}} : \varepsilon(\mathsf{W}) < \varepsilon(\mathsf{L}), \tag{I}$$

$$H^0_{\mathsf{NL}} : \varepsilon(\mathsf{W}) \geq \varepsilon(\mathsf{NL}) \quad \text{vs} \quad H^1_{\mathsf{NL}} : \varepsilon(\mathsf{W}) < \varepsilon(\mathsf{NL}). \tag{II}$$

- Repeat sampling M times to obtain $\widehat{\varepsilon}(\mathsf{M})$.
- Test statistic: $\widehat{\varepsilon}(\mathsf{W}) - \widehat{\varepsilon}(\mathsf{L})$ or $\widehat{\varepsilon}(\mathsf{W}) - \widehat{\varepsilon}(\mathsf{NL})$.
- Reject H^0 implies W significantly outperforms the competitor.

Simulation Results

Sample size	Average of $\kappa(\mathrm{W}(\widehat{\Sigma}))$	Average of $\kappa(\mathrm{L}(\widehat{\Sigma}))$	t-value	p-value	Reject H_{L}^0 or not
300	320.60	3.00	-76.58	0.0	**Reject**
400	230.49	3.40	-306.72	0.0	**Reject**
600	189.82	4.16	-628.65	0.0	**Reject**
800	180.02	4.88	-578.36	0.0	**Reject**
1000	177.75	5.61	-642.50	0.0	**Reject**
1200	177.52	6.30	-652.57	0.0	**Reject**
1400	177.93	6.98	-690.35	0.0	**Reject**
1600	178.81	7.65	-710.80	0.0	**Reject**
1800	179.42	8.30	-736.08	0.0	**Reject**
2000	180.55	8.95	-780.41	0.0	**Reject**

Table 3: Hypothesis test for $p = 200$.

Simulation Results

Sample size	Average of $\kappa(\mathrm{W}(\widehat{\Sigma}))$	Average of $\kappa(\mathrm{NL}(\widehat{\Sigma}))$	t-value	p-value	Reject H^0_{NL} or not
300	320.60	63.96	-10.30	0.0	**Reject**
400	230.49	97.50	-69.99	0.0	**Reject**
600	189.82	128.77	-231.95	0.0	**Reject**
800	180.02	144.63	-252.25	0.0	**Reject**
1000	177.75	154.88	-187.61	0.0	**Reject**
1200	177.52	161.76	-138.53	0.0	**Reject**
1400	177.93	166.40	-106.77	0.0	**Reject**
1600	178.81	169.92	-86.35	0.0	**Reject**
1800	179.42	172.26	-77.84	0.0	**Reject**
2000	180.55	174.70	-69.77	0.0	**Reject**

Table 4: Hypothesis test for $p = 200$.