SCOPE: Spectral Concentration by Distributionally Robust Joint Covariance-Precision Estimation

Renjie Chen* Viet Anh Nguyen* Huifu Xu*

Abstract

We propose a distributionally robust formulation for simultaneously estimating the covariance matrix and the precision matrix of a random vector. The proposed model minimizes the worst-case weighted sum of the Frobenius loss of the covariance estimator and Stein's loss of the precision matrix estimator against all distributions from an ambiguity set centered at the nominal distribution. The radius of the ambiguity set is measured via convex spectral divergence. We demonstrate that the proposed distributionally robust estimation model can be reduced to a convex optimization problem, thereby yielding quasi-analytical estimators. The joint estimators are shown to be nonlinear shrinkage estimators. The eigenvalues of the estimators are shrunk nonlinearly towards a positive scalar, where the scalar is determined by the weight coefficient of the loss terms. By tuning the coefficient carefully, the shrinkage corrects the spectral bias of the empirical covariance/precision matrix estimator. By this property, we call the proposed joint estimator the **S**pectral concentrated **CO**variance and **P**recision matrix **E**stimator (**SCOPE**). We demonstrate that the shrinkage effect improves the condition number of the estimator. We provide a parameter-tuning scheme that adjusts the shrinkage target and intensity that is asymptotically optimal. Numerical experiments on synthetic and real data show that our shrinkage estimators perform competitively against state-of-the-art estimators in practical applications.

1 Introduction

The covariance matrix $\Sigma_0 \triangleq \mathbb{E}_{\mathbb{P}}[(\xi - \mathbb{E}_{\mathbb{P}}[\xi])(\xi - \mathbb{E}_{\mathbb{P}}[\xi])^\top]$ of a random vector $\xi \in \mathbb{R}^p$ with distribution $\mathbb{P}$ is a widely-used summary statistic that captures the dispersion of ξ. The inverse of the covariance matrix, Σ_0^{-1}, is called the precision matrix. It is intuitive from the terminology that a random vector with a large covariance matrix is of low precision, and vice versa (Nguyen et al., 2022). The covariance matrix and the precision matrix, and their estimators, are ubiquitous in data-driven applications. In Markowitz's mean-variance model (Markowitz, 1952), the covariance matrix defines the risk of the given portfolio, and the precision matrix determines the optimal portfolio that achieves the minimal risk. In risk-sensitive control, the policy optimization objective itself depends on both the covariance matrix and the precision matrix of the Gaussian noise (Wang et al., 2020, Appendix 1). In Gaussian graphical model (Yuan and Lin, 2007; Zhao and Duan, 2019), a Gaussian random variable is associated with a graph. The covariance matrix encodes the marginal correlations of the nodes, and the precision matrix encodes the conditional independence of the nodes on the graph. In moment-based distributionally robust optimization (Delage and Ye, 2010), the first moment uncertainty is captured by an

*The authors are with the Chinese University of Hong Kong (rchen@se.cuhk.edu.hk,nguyen@se.cuhk.edu.hk,hfxu@se.cuhk.edu.hk).

ellipsoid whose shape matrix is the precision matrix, and the second moment uncertainty involves a matrix inequality defined by the covariance matrix. These examples show that the covariance matrix and the precision matrix are often used in conjunction in practical applications. However, in most data-driven problems, the true covariance matrix and precision matrix are unknown. This requires a statistical model that provides stable estimators of the covariance and precision matrix.

The estimation of the covariance matrix and the precision matrix is widely studied in the literature. Since the true distribution of ξ is usually not accessible, and the calculation of multi-dimensional integration could be prohibitively expensive (even if we have access to $\mathbb{P}$), we typically have to estimate the true distribution via n independent and identically distributed samples $\xi_1, \ldots, \xi_n \sim \mathbb{P}$. Let $\widehat{\mu} \triangleq \frac{1}{n}\sum_{i=1}^n \xi_i$ be the sample mean. The most straightforward unbiased estimator of the covariance matrix is the sample covariance matrix $\widehat{\Sigma} \triangleq \frac{1}{n-1}\sum_{i=1}^n (\xi_i - \widehat{\mu})(\xi_i - \widehat{\mu})^\top$. When $\widehat{\Sigma}$ is non-singular, a standard estimator of the precision matrix is $\widehat{\Sigma}^{-1}$. However, this approach does not give a valid estimator of the precision matrix when $\widehat{\Sigma}$ is singular, which is a circumstance that occurs frequently in the high-dimensional and data-deficient regime, i.e., when the sample size is less than the dimension of the problem ($n < p$) (Wainwright, 2019). Another drawback of the sample covariance matrix is that it suffers from systematic spectral bias. To see this, we start from the estimation bias of the largest and smallest eigenvalues of Σ_0. Let $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ be the largest eigenvalue and the smallest eigenvalue of the matrix X, respectively. Since $\lambda_{\max}(\cdot)$ is a convex function and $\lambda_{\min}(\cdot)$ is a concave function over the space of symmetric matrices, then by Jensen's inequality, it holds that

$$\lambda_{\max}(\Sigma_0) = \lambda_{\max}(\mathbb{E}_n[\widehat{\Sigma}]) \leq \mathbb{E}_n[\lambda_{\max}(\widehat{\Sigma})] \text{ and } \lambda_{\min}(\Sigma_0) = \lambda_{\min}(\mathbb{E}_n[\widehat{\Sigma}]) \geq \mathbb{E}_n[\lambda_{\min}(\widehat{\Sigma})], \tag{1}$$

where $\mathbb{E}_n$ denotes the expectation with respect to the n-product measures of the data-generating distribution $\mathbb{P}$ and the expectation is taken on $\widehat{\Sigma}$. Inequalities in (1) reveal that, on average, $\lambda_{\max}(\widehat{\Sigma})$ overestimates its true counterpart $\lambda_{\max}(\Sigma_0)$, while $\lambda_{\min}(\widehat{\Sigma})$ underestimates its true counterpart $\lambda_{\min}(\Sigma_0)$. The spectral bias occurs not only in the estimation of the largest and smallest eigenvalues but also in the estimation of all other eigenvalues. Let $\lambda(\Sigma_0) \triangleq (\lambda_1, \ldots, \lambda_p)$ be the eigenvalues of Σ_0, and $\lambda(\widehat{\Sigma}) \triangleq (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)$ be the eigenvalues of $\widehat{\Sigma}$, both in increasing order. In Ledoit and Wolf (2004), the authors point out that, on average, $\lambda(\widehat{\Sigma})$ is more dispersed than $\lambda(\Sigma_0)$, in the sense that

$$\frac{1}{p}\mathbb{E}_n\left[\sum_{i=1}^p (\widehat{\lambda}_i - \overline{\lambda})^2\right] = \frac{1}{p}\sum_{i=1}^p (\lambda_i - \overline{\lambda})^2 + \mathbb{E}_n[\|\widehat{\Sigma} - \Sigma_0\|_F^2] > \frac{1}{p}\sum_{i=1}^p (\lambda_i - \overline{\lambda})^2,$$

where $\overline{\lambda} \triangleq \frac{1}{p}\sum_{i=1}^p \lambda_i$. The inequality above implies that systematic bias may occur in the estimation of all eigenvalues in a similar manner of (1):

$$\begin{cases} \widehat{\lambda}_i \text{ overestimates } \lambda_i, & \text{when } \lambda_i \text{ is large,} \\ \widehat{\lambda}_i \text{ underestimates } \lambda_i, & \text{when } \lambda_i \text{ is small.} \end{cases} \tag{2}$$

From a distributional point of view, the dispersion of $\lambda(\widehat{\Sigma})$ can also be explained by the Marchenko–Pastur law (Marčenko and Pastur, 1967). The law states that in the high-dimensional regime where $p, n \to \infty, p/n \to c$ and $\Sigma_0 = I$, the eigenvalues of $\widehat{\Sigma}$ are supported on the interval $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$. It is intuitive to see that the ratio p/n governs the degree of dispersion: when the sample size n is large relative to p, the eigenvalues of $\widehat{\Sigma}$ concentrate around 1; conversely, when the sample size is relatively small, the eigenvalues spread over a wider interval.

The spectral bias poses challenges on both the computational and modeling sides. Firstly, it leads to an inflated condition number of $\widehat{\Sigma}$, making the calculation of $\widehat{\Sigma}^{-1}$ computationally unstable. Recall that the condition number of $\widehat{\Sigma}$ is defined by $\kappa(\widehat{\Sigma}) \triangleq \widehat{\lambda}_p/\widehat{\lambda}_1$. A combination of the spectral bias effects in (1) inflates the condition number $\kappa(\widehat{\Sigma})$. Such spectral bias and ill-conditioning not only bring in the difficulty of calculating $\widehat{\Sigma}^{-1}$, but also amplify the estimation error in downstream precision-matrix-related models. Markowitz's mean-variance model could be an example, since the construction of the optimal portfolio involves the precision matrix estimator (Ledoit and Wolf, 2003b; Michaud, 1989). From a modeling perspective, one may view ξ as random returns of risky assets and interpret the eigenvalues of $\widehat{\Sigma}$ as the risk or volatility of the portfolios induced by the eigenvectors of $\widehat{\Sigma}$. The spectral bias, as indicated in (2), suggests that we overestimate the potential risk of the more risky portfolio and underestimate the potential risk of the less risky portfolio. Such estimation bias could be misleading and may result in suboptimal decision-making in risk-averse applications.

An effective and promising way to correct the spectral bias and reduce the mean square error of the covariance-precision matrix estimator is to shrink the sample covariance matrix towards a well-conditioned and data-*in*sensitive target matrix. In Stein (1975, 1986), Charles Stein first proposed that shrinking the sample covariance matrix toward a constant matrix helps reduce the mean squared error of the estimation. Since Stein's seminal work, the study of Stein-type shrinkage estimators has been an important research area in statistics. Formally, let $\widehat{\Sigma} = \widehat{V}\mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)\widehat{V}^\top$ be the spectral decomposition of the sample covariance matrix. A covariance matrix estimator $\widetilde{\Sigma}$ is called a Stein-type shrinkage estimator if it is of the form $\widetilde{\Sigma} = \widehat{V}\mathrm{diag}(\widetilde{\lambda}_i, \ldots, \widetilde{\lambda}_p)\widehat{V}^\top$, where $\widetilde{\lambda}_i \triangleq \varphi(\widehat{\lambda}_i), i = 1, \ldots, p$ are the shrunk eigenvalues induced by the eigenvalue mapping $\varphi(\cdot)$ (Won et al., 2012). Note that Stein-type shrinkage estimators are *rotation-equivariant*, i.e., let $\widetilde{\Sigma}(\{\xi_i\}_{i=1}^n)$ denote the estimator on the dataset $\{\xi_i\}_{i=1}^n$ and R be a rotation matrix. Then the estimator $\widetilde{\Sigma}(\{R\xi_i\}_{i=1}^n)$ on the rotated dataset $\{R\xi_i\}_{i=1}^n$ is equivalent to $R\widetilde{\Sigma}(\{\xi_i\}_{i=1}^n)R^\top$.

In the study of Stein-type shrinkage covariance-precision matrix estimators, there are two main research directions: linear shrinkage and nonlinear shrinkage, which are determined by whether the eigenvalue mapping φ is linear or nonlinear. A linear shrinkage estimator is usually induced by a convex combination of the sample covariance matrix $\widehat{\Sigma}$ and a data-insensitive shrinkage target, i.e., $\widetilde{\Sigma} = (1-t)\widehat{\Sigma}+tT$. Typical choices of shrinkage target T include: (a) $\frac{1}{p}\mathrm{Tr}(\Sigma_0)I$ (Ledoit and Wolf, 2004), which is the optimal scalar matrix that minimize the mean squared error of the shrinkage estimator, (b) the constant correlation model (Ledoit and Wolf, 2003a), in which the target is a sample constant correlation matrix, (c) the single index model (Ledoit and Wolf, 2003b), in which the target is the sum of a rank-one matrix and a diagonal matrix that represent systematic and idiosyncratic risk factors of Sharpe's single index model (Sharpe, 1963), (d) semantic similarity shrinkage (Becquin and Esmeir, 2023), in which the target is constructed from semantic similarity of the textual descriptions or knowledge graphs. The combination coefficient t is usually chosen by minimizing the mean squared error of the induced shrinkage estimator. A linear shrinkage estimator is Stein-type (rotation-equivariant) only if the target commutes with $\widehat{\Sigma}$.

As an extension of the linear shrinkage estimator, Ledoit and Wolf (2012, 2020a) proposed a Stein-type nonlinear shrinkage estimator with a well-constructed eigenvalue mapping that is asymptotically optimal with respect to the mean squared error. The optimal shrinkage estimator of the precision matrix can be constructed in a similar manner (Bodnar et al., 2016; Ledoit and Wolf, 2022). Won et al. (2012) proposed a condition-number-regularized covariance estimation model. The model seeks to maximize the log-likelihood of the estimator, subject to the constraint that the condition number

of the estimator does not exceed a predetermined upper bound. It turns out that the optimizer can be viewed as a Stein-type nonlinear shrinkage estimator with a piece-wise linear eigenvalue mapping. Recently, Liu and Liu (2025) points out that in financial markets, the returns of multiple assets are usually positively correlated, and considers the covariance matrix estimation problem under this decision scenario. Unlike traditional (linear or nonlinear) shrinkage estimators that shrink all the eigenvalues towards the same target, the proposed Stein-type shrinkage estimator adjusts the eigenvalues in pairs and shrinks the gap between pairs of eigenvalues. For a review of shrinkage covariance estimators, see Ledoit and Wolf (2020b).

Recently, it has been demonstrated that Stein-type shrinkage estimators can also be constructed from a distributionally robust optimization (DRO) perspective. Yue et al. (2024) use a second-moment-based distributionally robust approach to model the estimation of covariance matrices, and show that the resulting estimator is a Stein-type estimator that shrinks the eigenvalues of the nominal covariance estimator towards zero in a nonlinear sense. Similarly, Nguyen et al. (2022) considers the Wasserstein distributionally robust precision matrix estimation problem. The proposed Stein-type precision matrix estimator shrinks the eigenvalues of the precision matrix toward zero in a nonlinear manner. These two seminal works establish a connection between the distributionally robust optimization models from the OR society and the Stein-type estimators from the statistics society. However, neither of them addresses the issue of systematic spectral bias (2). Both methods shrink the covariance or precision matrix estimator towards the zero matrix, and thus only correct one side of the spectral bias in (2). Such a one-sided correction will aggravate the spectral bias on the other side and lead to a misunderstanding about the variance. For a comparison of the shrinkage estimators proposed in the literature, refer to Table 1.

The tuning of the radius of the divergence-based ambiguity set is another crucial aspect of the DRO model. The radius represents the conservativeness of the model, specifically controlling the intensity of shrinkage in the Stein-type distributionally robust covariance/precision matrix estimators mentioned above. A larger radius implies greater shrinkage. From a theoretical perspective, the radius of the ambiguity set can be selected by using a finite sample guarantee principle (Mohajerin Esfahani and Kuhn, 2018; Yue et al., 2024), in which case the radius is chosen so that the ground truth lies in the ambiguity set with high probability. In practice, the cross-validation (CV) method is widely used for parameter tuning. For applications of the CV method in radius tuning in DRO, see the numerical experiment parts in Mohajerin Esfahani and Kuhn (2018); Yue et al. (2024); Gao and Kleywegt (2023); Nguyen et al. (2022). The numerical experiments show that, compared to the optimal radius suggested by the CV method, the finite sample guarantee principle tends to be more conservative. In a recent paper, Blanchet and Si (2019) propose a customized radius tuning scheme for the Wasserstein distributionally robust precision matrix estimation model (Nguyen et al., 2022). They select the optimal radius by minimizing the estimation loss of the DRO estimator induced by the radius. They demonstrate that the order of the radius proposed by their scheme aligns with the order suggested by the CV method, thereby validating the empirical findings and providing theoretical support for them.

Contributions. We propose a moment-based distributionally robust optimization model that jointly estimates the covariance matrix and precision matrix. The proposed DRO model minimizes the worst-case sum of the Frobenius loss of the covariance estimator and weighted Stein's loss of the precision matrix estimator. The ambiguity set contains distributions whose covariance matrix is close to the nominal covariance matrix. The distance between matrices is measured via divergence functions over the space of positive semi-definite matrices.

Estimator	Estimator Type	Shrinkage Type	Shrinkage Target	Shrinkage Intensity
Ledoit and Wolf (2004)	Covariance Estimator	Linear	$\frac{1}{p}\mathrm{Tr}(\Sigma_0)I$	Controlled by Combination Coefficient
Ledoit and Wolf (2012)	Covariance Estimator	Nonlinear	N.A.	N.A
Yue et al. (2024)	Covariance Estimator	Nonlinear	0	Controlled by Radius of Ambiguity Set
Nguyen et al. (2022)	Precision Estimator	Nonlinear	0	Controlled by Radius of Ambiguity Set
SCOPE (this paper)	Covariance & Precision Estimator	Nonlinear	$\sqrt{\frac{1}{\tau}}I$	Controlled by Radius of Ambiguity Set

Table 1: A comparison of the shrinkage estimators of the covariance-precision matrix in the literature and the estimator proposed in this paper. The parameter τ in the shrinkage target of SCOPE is chosen by the statistician.

Because the precision matrix is the inverse of the covariance matrix, the estimation problem involves a bilinear constraint restricting the matrix product of the covariance and precision estimates to be the identity matrix. This bilinear constraint injects nonconvexities into the estimation problem, which is fundamentally different from the nonconvexity in earlier distributionally robust estimation frameworks. The main contributions of this paper can be summarized as follows.

- Under regularity assumptions on the divergence function, the DRO covariance-precision matrix estimator can be found by solving a convex optimization model, despite the nonconvex bilinear constraint.
- Additionally, when the divergence function is a convex spectral divergence, we show that the DRO covariance-precision matrix estimator is a Stein-type nonlinear shrinkage estimator and admits a quasi-closed form. We provide an explicit expression of the shrinkage target of the estimator, parametrized by the weight coefficient of the objective function. By tuning the coefficient τ judiciously, we can ensure that the target is a well-conditioned matrix. Consequently, the spectral bias (2) of the estimator is corrected from two sides: both the large and the small eigenvalues are shrunk to the middle. As a result of the bias correction, the condition number of the estimator is also improved. By this spectral correction effect, we call the proposed estimator the **S**pectrum concentrated **CO**variance and **P**recision matrix **E**stimator (**SCOPE**).
- We provide explicit expressions of the estimator when the divergence function is a commonly-used convex spectral divergence, such as the Kullback-Leibler divergence, Wasserstein divergence, Symmetrized Stein divergence, Squared Frobenius, and Weighted Frobenius divergence. See Table 2 for definitions of these divergences.
- We propose a parameter-tuning scheme that adjusts the shrinkage target and the shrinkage intensity by minimizing the violation of the optimality of the underlying true covariance Σ_0.

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 2: Popular divergence functions and their generators discussed in this paper. See Assumption 3 for the definition of the generator of a divergence.

The result shows that the optimal shrinkage target is a scalar matrix with the same norm as Σ_0, and the optimal order to the radius of the ambiguity set is $O(n^{-2})$.

The remainder of the paper is organized as follows. In Section 2, we propose our distributionally robust joint covariance-precision estimation model, demonstrate the difficulty of the model, and compare it with the distributionally robust covariance estimation model proposed by Yue et al. (2024). In Section 3, we show that under regularity assumptions on the divergence functions, the covariance-precision matrix estimator induced by the DRO model admits a quasi-closed form. We further demonstrate that the DRO model shrinks the estimator towards a well-conditioned target, thereby correcting the spectral bias from both sides. Specific covariance-precision matrix estimators induced by spectral convex divergence functions are listed in Table 3. In Section 4, we propose a practical parameter tuning scheme and adjust the shrinkage target and the shrinkage intensity accordingly. Finally, we report numerical results based on synthetic and real data in Section 5.

Notations. By convention, we use $\mathbb{R}, \mathbb{R}_+, \mathbb{R}_{++}$, and $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ to denote the entire real line, the set of non-negative real numbers, the set of strictly positive real numbers, and the extended real line. The p-dimensional Euclidean space and its subsets of non-negative real vectors and strictly positive real vectors are denoted by $\mathbb{R}^p, \mathbb{R}^p_+, \mathbb{R}^p_{++}$, respectively. The space of $p \times p$ real matrices is denoted by $\mathbb{R}^{p\times p}$, and the space of $p \times p$ symmetric matrices, the cone of positive semi-definite matrices, and the cone of strictly positive definite matrices are denoted by $\mathbb{S}^p, \mathbb{S}^p_+$, and $\mathbb{S}^p_{++}$, respectively. The identity matrix is denoted by I. For a vector $x \in \mathbb{R}^p$, we use $x^\uparrow$ to denote the vector obtained by rearranging the entries of x in non-decreasing order. The trace of a matrix $S \in \mathbb{R}^{p\times p}$ is defined as $\mathrm{Tr}(S) \triangleq \sum_{i=1}^p S_{ii}$, and the inner product of two matrices $X, Y \in \mathbb{R}^{p\times p}$ is defined via $\langle X, Y\rangle \triangleq \mathrm{Tr}(X^\top Y)$. The Frobenius norm of a matrix $S \in \mathbb{R}^{p\times p}$ is defined by $\|S\|_F \triangleq \sqrt{\langle S, S\rangle}$. By slight abuse of notation, we use $\mathrm{diag}(S)$ to denote the vector of diagonal entries of matrix $S \in \mathbb{R}^{p\times p}$, and $\mathrm{diag}(x)$ to denote the diagonal matrix whose diagonal is given by vector $x \in \mathbb{R}^p$. The domain of a non-negative function f is defined as $\mathrm{dom}(f) := \{x : f(x) < \infty\}$.

2 Distributionally Robust Joint Covariance-Precision Estimation

Let ξ be a random vector in $\mathbb{R}^p$ with mean zero and covariance matrix Σ_0. Let $\widehat{\mathbb{P}}$ be a nominal distribution constructed from n samples of ξ, and $\widehat{\Sigma} \triangleq \mathbb{E}_{\widehat{\mathbb{P}}}[\xi\xi^\top]$ be the nominal covariance matrix estimator. A typical choice of $\widehat{\mathbb{P}}$ could be the empirical distribution of n independent realizations of ξ, and $\widehat{\Sigma}$ is the sample covariance matrix of the n samples in this case. However, at this stage, we make no assumptions about the choice of $\widehat{\mathbb{P}}$ and $\widehat{\Sigma}$ to preserve the flexibility of our model. For convenience in modeling, Yue et al. (2024) proposed to view $\widehat{\Sigma}$ as the unique solution to the minimization problem:

$$\widehat{\Sigma} = \arg\min_{\Sigma \in \mathbb{S}^p_+} \|\Sigma\|_F^2 - 2\langle \Sigma, \widehat{\Sigma}\rangle = \arg\min_{\Sigma \in \mathbb{S}^p_+} \underbrace{\|\Sigma\|_F^2 - 2\mathbb{E}_{\widehat{\mathbb{P}}}[\langle \Sigma, \xi\xi^\top\rangle]}_{\triangleq \text{FrobeniusLoss}(\Sigma, \widehat{\mathbb{P}})}, \tag{3}$$

where the expectation is taken over ξ. The objective function of (3) is termed the Frobenius loss in the variable Σ, evaluated under the distribution $\widehat{\mathbb{P}}$. On the other hand, if the underlying distribution of ξ is assumed to be a Gaussian distribution, the maximum likelihood estimator $\widehat{X}$ for the precision matrix Σ_0^{-1} is given by the unique minimizer (if it exists) of the minimization problem:

$$\widehat{X} = \arg\min_{X \in \mathbb{S}^p_{++}} -\log\det X + \langle X, \widehat{\Sigma}\rangle = \arg\min_{X \in \mathbb{S}^p_{++}} \underbrace{-\log\det X + \mathbb{E}_{\widehat{\mathbb{P}}}[\langle X, \xi\xi^\top\rangle]}_{\triangleq \text{SteinLoss}(X, \widehat{\mathbb{P}})}. \tag{4}$$

The objective function of (4) is known as Stein's loss in the variable X (James et al., 1961). The model (4) can also be viewed as a Bregman divergence minimization problem that minimizes the divergence between X and its true counterpart Σ_0^{-1}, and thus is also applicable for the estimation of non-Gaussian random variables (Ravikumar et al., 2011). A downside of problem (4) is that when $\widehat{\Sigma}$ is singular, the minimizer does not exist, hence it requires that the sample covariance matrix $\widehat{\Sigma}$ be nonsingular.

Motivated by problems (3) and (4), we estimate the covariance matrix and the precision matrix simultaneously by solving the following problem:

$$\begin{array}{ll} \min\limits_{\Sigma, X} & \text{SteinLoss}(X, \widehat{\mathbb{P}}) + \frac{\tau}{2}\text{FrobeniusLoss}(\Sigma, \widehat{\mathbb{P}}) \\ \text{s.t.} & \Sigma \in \mathbb{S}^p_{++},\ X \in \mathbb{S}^p_{++},\ X\Sigma = I. \end{array} \tag{5}$$

The objective function is a linear combination of the Stein loss and the Frobenius loss under a non-negative weighting coefficient τ, where τ balances the two losses. The constraint $X\Sigma = I$ ensures that the estimators of covariance and precision matrices are inverses of each other, and thus consistent with their definition. When $\widehat{\Sigma}$ is nonsingular, the optimal solution to (5) is $(\Sigma^\star, X^\star) = (\widehat{\Sigma}, \widehat{\Sigma}^{-1})$. This recovers the classical covariance and inverse covariance estimators. However, if $\widehat{\Sigma}$ is singular (also known as rank deficient), problem (5) is unbounded below and there is no optimal solution.[1]

Given the above argument, the joint estimation model (5) is not suitable for a high-dimensional setting, where the sample size n is less than or comparable to the problem dimension p. Moreover, the classical model (5) also suffers from distribution ambiguity: In data-driven problems, the available data are often limited, and the empirical distribution $\widehat{\mathbb{P}}$ may not approximate the data-generating distribution $\mathbb{P}$ well. As a consequence, the out-of-sample performance is not reliable.

[1] See Lemma 1 for a formal statement.

To address these issues, we consider a distributionally robust estimation formulation motivated by model (5). Consider a moment-based ambiguity set "centered" at the nominal distribution $\widehat{\mathbb{P}}$ with radius $\rho > 0$:

$$\mathcal{P}_\rho(\widehat{\mathbb{P}}) \triangleq \left\{\mathbb{Q} : \mathbb{E}_{\mathbb{Q}}[\xi] = 0, \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] = S, D\left(S, \mathbb{E}_{\widehat{\mathbb{P}}}[\xi\xi^\top]\right) \leq \rho\right\}. \tag{6}$$

The set $\mathcal{P}_\rho(\widehat{\mathbb{P}})$ contains all the distributions with mean zero and second-moment matrix S that is of a divergence at most ρ from the nominal covariance matrix estimator $\mathbb{E}_{\widehat{\mathbb{P}}}[\xi\xi^\top] = \widehat{\Sigma}$. Here, we utilize a divergence function D in the space of positive semi-definite matrices to measure the discrepancy between two matrices. Divergences are generalized distance functions that need not be symmetric or satisfy the triangle inequality. When we construct the ambiguity set $\mathcal{P}_\rho(\widehat{\mathbb{P}})$, the choice of ρ reveals the conservativeness of the set: when we believe the nominal value is accurate, we shall set a smaller ρ, and if not, we shall set a larger ρ.

To robustify the empirical model (5) against all potential mismatches against distributions in $\mathcal{P}_\rho(\widehat{\mathbb{P}})$, we define the pair of distributionally robust covariance-precision matrix estimators $(\Sigma^\star, X^\star)$ as the solution of a min-max problem

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

The uniqueness of the solution is guaranteed under some moderate conditions on the divergence function D. We will formally prove this result in Theorem 1. Since the objective function of the problem above depends linearly on $\mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$, we can reformulate the DRO as a robust optimization (RO) problem:

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

The objective function $f(\Sigma, X, S)$ is convex in (Σ, X) and concave in S. Later, we will see that the weighting coefficient τ parameterizes the shrinkage target of the model, while the radius ρ controls the shrinkage effect. We will discuss their choices in Sections 4. In the remainder of this section, we discuss the main difficulty of (RO) and delineate the key differences between the DRO joint estimation problem (DRO) and the existing formulations in the literature.

Difficulty of (RO) and comparison with Yue et al. (2024). In Yue et al. (2024, equation (4)), the authors propose a DRO formulation of the covariance estimation problem (3) using the same ambiguity set $\mathcal{P}_\rho(\widehat{\mathbb{P}})$ as in (6). Their model reduces to the following robust optimization problem

$$\min_{\Sigma \in \mathbb{S}^p_+} \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \leq \rho} \left\{g(\Sigma, S) \triangleq \|\Sigma\|_F^2 - 2\langle\Sigma, S\rangle\right\}, \tag{7}$$

where the divergence D is chosen from a family of (possibly nonconvex) divergence functions. The main difficulty in (7) arises from the non-convexity of the divergence function D. To obtain a tractable reformulation of (7), the authors exploit the structure of the divergence function and propose a (generalized) minimax theorem for geodesically convex-concave functions that is applicable to (7); see (Yue et al., 2024, theorem 3). In contrast, the difficulty in problem (RO) arises from the bilinear constraint $X\Sigma = I$, which makes it difficult to justify whether exchanging the order of min-max will preserve the optimal value and optimal solution to the problem. This bilinear term cannot be convexified by elementary operations. We consider two standard ways to eliminate the bilinear constraint.

(i) Eliminate Σ by substituting it with X^{-1}. Consequently, problem (RO) becomes an optimization problem with variables (X, S) only:

$$\min_{X\in\mathbb{S}^p_{++}} \max_{S\in\mathbb{S}^p_+:D(S,\widehat{\Sigma})\leq\rho} \left\{ f_1(X,S) \triangleq -\log\det X + \langle X, S\rangle + \frac{1}{2}\tau\left(\|X^{-1}\|_F^2 - 2\langle X^{-1}, S\rangle\right)\right\}. \tag{8}$$

The objective function $f_1(X, S)$ is, in general, not convex in X. The non-convexity arises from the term $\|X^{-1}\|_F^2 - 2\langle X^{-1}, S\rangle$. To see this, let us consider the 1-dimensional case and take $S = 1$, this term becomes $h(x)x^{-2} - 2x^{-1}$. Since $h''(2) = -\frac{1}{8}$, then $h(x)$ is nonconvex in x. So for a large τ, $f_1(X, S)$ is not convex in X. Moreover, consider the inner maximization problem of (8) in the variable S, i.e.,

$$\max_{S\in\mathbb{S}^p_+:D(S,\widehat{\Sigma})\leq\rho} \langle X, S\rangle - \tau\langle X^{-1}, S\rangle \;=\; \max_{S\in\mathbb{S}^p_+:D(S,\widehat{\Sigma})\leq\rho} \langle X - \tau X^{-1}, S\rangle.$$

This problem is not even a geodesically convex problem because the matrix coefficient $X - \tau X^{-1}$ can be indefinite for some $X \in \mathbb{S}^p_+$. So the *geodesically convex-concave* minimax theorem in (Yue et al., 2024, theorem 3) does not apply to (8).

(ii) Alternatively, we can eliminate X by substituting X with Σ^{-1}. In this case, we obtain an optimization in only the (Σ, S) variable:

$$\min_{\Sigma\in\mathbb{S}^p_{++}} \max_{S\in\mathbb{S}^p_+:D(S,\widehat{\Sigma})\leq\rho} \left\{ f_2(\Sigma,S) \triangleq \log\det\Sigma + \langle \Sigma^{-1}, S\rangle + \frac{1}{2}\tau\left(\|\Sigma\|_F^2 - 2\langle \Sigma, S\rangle\right)\right\}. \tag{9}$$

The objective function f_2, again, is neither convex-concave nor geodesically convex-concave, and the existing methods fail to derive the minimax property of (9).

The discussions above show that the bilinear constraint has made the model intrinsically nonconvex. To ensure a tractable formulation of (RO), we may have to restrict the choice of function D. In the forthcoming discussions, we will concentrate on the case where the divergence D is convex. It turns out that problem (RO) admits a convex reformulation under the convexity assumption of the divergence, see the details in the next section.

3 Representation and Characterization of the Estimators

In this section, we formally derive the semi-analytical form of the DRO joint covariance-precision estimator induced by (RO). Toward that goal, Section 3.1 shows that under a convexity assumption on the divergence D, (RO) reduces to a convex optimization problem, i.e., (P-Mat). In section 3.3, we introduce the concept of spectral divergence, and show that when divergence D is further assumed to be spectral divergence, the optimal solution to problem (P-Mat) admits a quasi-closed form. Finally, we derive the quasi-closed form of the distributionally robust covariance-precision matrix estimator $(\Sigma^\star, X^\star)$, and propose a computationally efficient framework for the estimator.

3.1 Convex Reformulation

Before delving into the details of the reformulation, we first note that to ensure a well-posed (RO) problem, we assume the following throughout the paper.

Assumption 1 (Regularity of nominal)**.** *For the nominal covariance matrix estimator* $\widehat{\Sigma}$*, it holds that* $(\widehat{\Sigma}, \widehat{\Sigma}) \in \mathrm{dom}(D)$*.*

Assumption 1 is to ensure well-definedness of $D(\widehat{\Sigma}, \widehat{\Sigma})$. We make this assumption as $D(\widehat{\Sigma}, \widehat{\Sigma}) = \infty$ for some divergence functions, such as Kullback-Leibler divergence, when $\widehat{\Sigma}$ is singular. Moreover, the convex reformulation of (RO) relies on the following convexity assumption on the choice of divergence D.

Assumption 2 (Convex divergence)**.** *The divergence* $D : \mathbb{S}^p_+ \times \mathbb{S}^p_+ \to \mathbb{R}_+$ *satisfy:*

(i) D *is non-negative, and satisfies the identity of indiscernibles, that is, for any* $(X, Y) \in \mathrm{dom}(D)$*,* $D(X, Y) = 0$ *if and only if* $X = Y$*,*

(ii) $D(\cdot, \widehat{\Sigma})$ *is convex, continuous, and differentiable,*

(iii) $D(\cdot, \widehat{\Sigma})$ *is coercive, i.e.,* $D(X, \widehat{\Sigma}) \to +\infty$ *as* $\|X\|_F \to \infty$*.*

Under Assumptions 1 and 2, it holds that $D(\widehat{\Sigma}, \widehat{\Sigma}) = 0$, and thus the feasible set $\{\Sigma \in \mathbb{S}^p_{++} : D(\Sigma, \widehat{\Sigma}) \leq \rho\}$ is non-empty. The following theorem states that under these assumptions, the optimal solution to (RO) can be derived by the optimal solution to the following convex optimization problem:

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

Theorem 1 (Convex reformulation of (RO))**.** *Let Assumptions 1 and 2 hold. Then*

(i) (P-Mat) *admits a unique optimal solution* $\Sigma^\star$*,*

(ii) (RO) *admits a unique optimal solution pair, which can be represented* $(\Sigma = \Sigma^\star, X = (\Sigma^\star)^{-1})$*.*

The theorem enables us to solve (RO) by solving (P-Mat). In the next subsections, we discuss the construction of the optimal solution to (P-Mat). First, we show in Section 3.2 that when the constraint of (P-Mat) is redundant, the optimal solution to (P-Mat) is a scaled identity matrix. Next, we demonstrate, in Section 3.3, that when the constraint is active, the optimal solution to (P-Mat) admits a quasi-closed form. This gives us an efficient way to construct the estimator induced by (RO), which is stated in the forthcoming Theorem 2.

3.2 A Trivial Solution

We first observe that when the radius ρ is sufficiently large, the constraint $D(\Sigma, \widehat{\Sigma}) \leq \rho$ becomes redundant in (P-Mat) in the sense that the global maximizer of the objective function lies in the feasible set. In that case, the optimal solution is $\Sigma^\star = \sqrt{\frac{1}{\tau}} I$. The next proposition states this.

Proposition 1 (Unbinding constraint)**.** *Let* ρ *be such that* $D\left(\sqrt{\frac{1}{\tau}} I, \widehat{\Sigma}\right) \leq \rho$*. Then* $\Sigma^\star = \sqrt{\frac{1}{\tau}} I$ *is the optimal solution to* (P-Mat)*.*

Proposition 1 states that when ρ is sufficiently large, the optimal solution to (P-Mat) is a scaled identity matrix. The magnitude of this scaling depends on τ, and we will see in Section 3.4 that the scaled identity matrix can be viewed as the shrinkage target of the proposed covariance estimator.

3.3 Quasi-closed Form of the Estimators

We now turn to the construction of $\Sigma^\star$ when $D(\sqrt{\frac{1}{\tau}}I, \widehat{\Sigma}) \geq \rho$, in which case the trivial solution in Proposition 1 is no longer valid. To ensure the quasi-closed form of the optimal solution to (P-Mat), we make the following *spectral divergence* assumption on divergence D. The idea is from Yue et al. (2024, Assumption 2), which summarizes a common property that is satisfied by many widely-used divergences: the divergence between two positive semi-definite matrices is determined by their spectra. For examples of divergences that satisfy Assumption 3, see Table 2.

Assumption 3 (Spectral divergence)**.** *The divergence $D : \mathbb{S}_+^p \times \mathbb{S}_+^p \to \mathbb{R}_+$ satisfies the following structural conditions:*

(i) (Orthogonal equivariance) For any $X, Y \in \mathbb{S}_+^p$ and $V \in \mathcal{O}(p)$, $D(X, Y) = D(VXV^\top, VYV^\top)$.

(ii) (Spectrality) There exists a function $d : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, called the generator of D, such that

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

and $d(a, b)$ is twice continuously differentiable in a for every $b \geq 0$.

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

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

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

In the remainder of this paper, we refer to a divergence D satisfying Assumptions 2 and 3 as a convex spectral divergence. The following remark on the generator of convex spectral divergences collects some basic properties of d, and will facilitate the upcoming derivation in this paper.

Remark 1 (Properties of generator d)**.** *Let D be a divergence that satisfies Assumption 3. Consider $(aI, bI) \in \mathrm{dom}(D)$. Then by Assumption 3(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 some basic calculus shows 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 $b > 0$.

Remark 2 (Regularity of spectrum of nominal estimator)**.** *Let $0 \leq \widehat{\lambda}_1 \leq \widehat{\lambda}_2 \leq \ldots \leq \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\Sigma}$ and $\widehat{\Sigma} = \widehat{V}\mathrm{diag}(\widehat{\lambda})\widehat{V}^\top$ be its spectral decomposition. Under Assumptions 1– 3, it holds that $(\widehat{\lambda}_i, \widehat{\lambda}_i) \in \mathrm{dom}(d)$ for all $i = 1, \ldots, p$. To see this, note that*

$$\sum_{i=1}^{p} d(\widehat{\lambda}_i, \widehat{\lambda}_i) = D(\mathrm{diag}(\widehat{\lambda}), \mathrm{diag}(\widehat{\lambda})) = D(\widehat{V}\mathrm{diag}(\widehat{\lambda})\widehat{V}^\top, \widehat{V}\mathrm{diag}(\widehat{\lambda})\widehat{V}^\top) = D(\widehat{\Sigma}, \widehat{\Sigma}) = 0.$$

Compared with Yue et al. (2024, Assumption 3), we do not require $\sum_{i=1}^{p} d(0, \widehat{\lambda}_i) > \rho$, as the log det term in the objective automatically excludes the trivial solution $\Sigma = \mathbf{0}$. Instead, in the following assumption, we exclude the case where $\widehat{\Sigma}$ is proportional to identity to ensure that the robustification is meaningful, since being proportional to identity is already robust enough.

Assumption 4 (Regularity of nominal estimator)**.** *The nominal covariance matrix estimator satisfies* $\widehat{\Sigma} \neq \sigma I$ *for any* $\sigma \in \mathbb{R}_{++}$.

The last assumption is to avoid the relatively trivial case as in Proposition 1, where we exploit the fact that under Assumption 3(i), we have

$$D\left(\sqrt{\frac{1}{\tau}}I, \widehat{\Sigma}\right) = D\left(\sqrt{\frac{1}{\tau}}I, \operatorname{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)\right) = \sum_{i=1}^{p} d\left(\sqrt{\frac{1}{\tau}}, \widehat{\lambda}_i\right).$$

Assumption 5 (Regularity of radius)**.** *The choice of* ρ *satisfies* $0 < \rho \leq \rho_{\max} \triangleq \sum_{i=1}^{p} d\left(\sqrt{\frac{1}{\tau}}, \widehat{\lambda}_i\right)$.

Under the above assumptions, we are able to construct the optimal solution to (P-Mat), and thus the covariance-precision matrix estimator induced by (RO), in a quasi-closed form. The construction involves a eigenvalue mapping corresponding to the generator d of the divergence function, which is defined by

$$\varphi(\tau, \gamma, b) \triangleq \text{the unique solution } a^\star > 0 \text{ of the equation } 0 = \frac{1}{a^\star} - \tau a^\star - \gamma \frac{\partial d}{\partial a}(a^\star, b). \tag{10}$$

The following proposition ensures that the mapping is well-defined.

Proposition 2 (Well-definedness of eigenvalue mapping φ)**.** *Let Assumption 2 and 3 hold. Then for* $\tau > 0$, $\gamma \geq 0$ *and* $b \geq 0$, *the equation* $\frac{1}{a} - \tau a - \gamma \frac{\partial d}{\partial a}(a, b) = 0$ *admits a unique solution in* $\mathbb{R}_{++}$.

After specifying the mapping φ, we are ready to propose the following construction theorem formally.

Theorem 2 (Construction of the covariance-precision matrix estimator)**.** *Under Assumptions 1-5, the optimal solution to* (RO) *is*

$$\Sigma^\star = \widehat{V}\Phi(\tau, \gamma^\star, \widehat{\lambda})\widehat{V}^\top, \quad X^\star = (\Sigma^\star)^{-1}, \tag{11}$$

where $\widehat{V}$ *and* $\widehat{\lambda}$ *are defined as in Remark 2 and* $\Phi(\tau, \gamma^\star, \widehat{\lambda}) \triangleq \operatorname{diag}(\varphi(\tau, \gamma^\star, \widehat{\lambda}_1), \ldots, \varphi(\tau, \gamma^\star, \widehat{\lambda}_p))$, *and* $\gamma^\star$ *is the unique non-negative root of the equation* $\sum_{i=1}^{p} d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \rho = 0$.

From the theorem, we can see that $\Sigma^\star(\tau, \rho)$ shares the same eigenbasis as the sample covariance matrix $\widehat{\Sigma}$, i.e., it is a Stein-type rotation equivariant estimator. The proof of Theorem 2 involves a number of intermediate results to be presented in Propositions 3-5. Figure 1 visualizes the flow of these results in the proof of Theorem 2. We begin with Proposition 3, which shows that (P-Mat) can be reduced to a convex problem that optimizes over all vectors in the non-negative orthant $\mathbb{R}^p_+$:

$$\begin{array}{ll} \max\limits_{s \in \mathbb{R}^p_+} & \sum\limits_{i=1}^{p} \log s_i - \frac{1}{2}\tau \sum\limits_{i=1}^{p} s_i^2 \\ \text{s.t.} & \sum\limits_{i=1}^{p} d(s_i, \widehat{\lambda}_i) \leq \rho. \end{array} \tag{P-Vec}$$

Proposition 3 (Equivalence of (P-Mat) and (P-Vec))**.** *Let Assumptions 1-3 hold. Then problem* (P-Mat) *is equivalent to* (P-Vec) *in the following sense:*

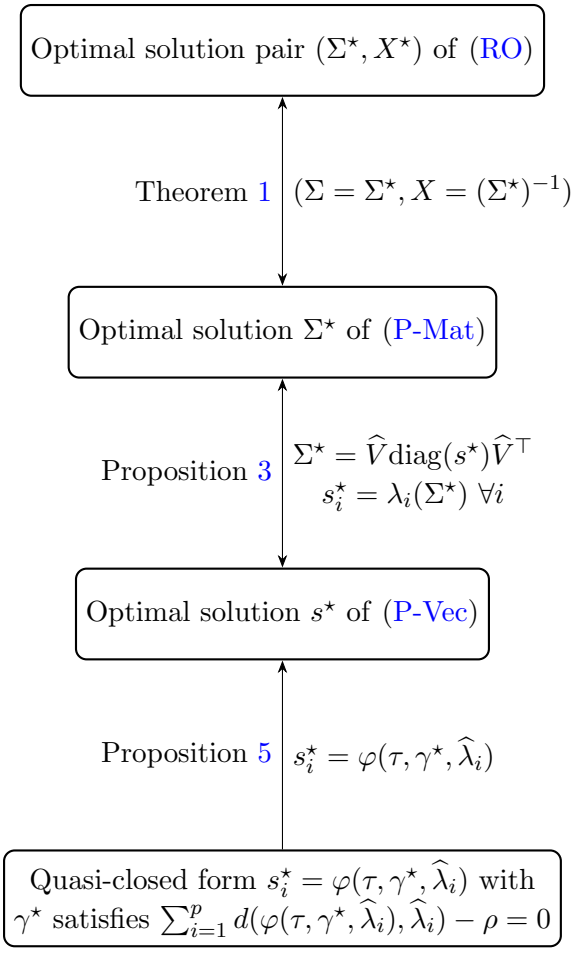


Figure 1: Proof structure of Theorem 2. Theorem 1 states that (RO) reduces to a convex optimization problem (P-Mat). Proposition 3 shows that the optimal solution of (P-Mat) can be constructed by solving (P-Vec), which is over a vector space. Proposition 5 reveals that the optimal solution to (P-Vec) admits a quasi-closed form, which is specified by eigenvalue mapping φ.

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

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

Next, we discuss how to solve (P-Vec). The representation of the solution of (P-Vec) involves the eigenvalue mapping φ defined in (10). We begin with the Lagrange function of (P-Vec)

$$\begin{aligned} L(\gamma, s) &= \sum_{i=1}^p \log s_i - \frac{1}{2}\tau \sum_{i=1}^p s_i^2 - \gamma\left(\sum_{i=1}^p d(s_i, \widehat{\lambda}_i) - \rho\right) \\ &= \sum_{i=1}^p \left(\log s_i - \frac{1}{2}\tau s_i^2 - \gamma d(s_i, \widehat{\lambda}_i)\right) - \gamma\rho, \end{aligned}$$

where $\gamma \geq 0$ is the Lagrange multiplier associated with the constraint $\sum_{i=1}^p d(s_i, \widehat{\lambda}_i) \leq \rho$. Then for fixed γ, the eigenvalue mapping $\varphi(\tau, \gamma, b)$ can be viewed as the minimizer of the summation term

$$\psi(\tau, \gamma, a, b) = \log a - \frac{1}{2}\tau a^2 - \gamma d(a, b).$$

Taking the derivative of ψ with respect to a and setting it to 0, we obtain the definition of φ. The following proposition 4 collects some properties of the mapping φ that will facilitate the upcoming discussion and analysis in this paper.

Proposition 4 (Properties of eigenvalue mapping φ)**.** *If Assumptions 1-3 hold, then it follows that*

(i) Let $\varphi_{\tau,b}(\gamma) \triangleq \varphi(\tau,\gamma,b)$. *If* $\tau>0$, $b>0$, *then for any* $\gamma \geq 0$*:*

$$\begin{cases} b < \varphi_{\tau,b}(\gamma) \leq 1/\sqrt{\tau}, & \text{for } b < 1/\sqrt{\tau}, \\ 1/\sqrt{\tau} \leq \varphi_{\tau,b}(\gamma) < b, & \text{for } b > 1/\sqrt{\tau}, \\ \varphi_{\tau,b}(\gamma) = 1/\sqrt{\tau}, & \text{for } b = 1/\sqrt{\tau}, \end{cases}$$

(ii) If $\tau>0$ *and* $b \geq 0$, *then* $\varphi_{\tau,b}(\gamma)$ *is continuous, strictly monotonic and differentiable on* $\mathbb{R}_+$. *Specifically,* $\varphi_{\tau,b}(\cdot)$ *is strictly increasing if* $b > 1/\sqrt{\tau}$ *and strictly decreasing if* $b < 1/\sqrt{\tau}$,

(iii) If $\tau>0$ *and* $b \geq 0$, *then* $\varphi_{\tau,b}(0) = 1/\sqrt{\tau}$ *and* $\lim_{\gamma\to\infty} \varphi_{\tau,b}(\gamma) = b$,

(iv) If $\tau>0$ *and* $\gamma \geq 0$, *then* $\phi(b) \triangleq \frac{\varphi(\tau,\gamma,b)}{b}$ *is non-increasing on* b.

The following proposition shows that the optimal solution to (P-Vec) admits a quasi-closed form, given that the eigenvalue mapping φ is known. The result shows that φ can be viewed as the eigenvalue mapping that distorts the empirical eigenvalues $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_p$.

Proposition 5 (Solution of (P-Vec))**.** *If Assumptions 1-5 hold, then* (P-Vec) *admits a unique optimal solution* $s^\star$ *and* $s_i^\star = \varphi(\tau,\gamma^\star,\widehat{\lambda}_i), i=1,\ldots,p$, *where* $\gamma^\star$ *is the unique solution of the nonlinear equation* $\sum_{i=1}^p d(\varphi(\tau,\gamma^\star,\widehat{\lambda}_i),\widehat{\lambda}_i) - \rho = 0$.

Remark 3 (Verification of the construction when $\rho=\rho_{\max}$)**.** *In the case where* $\rho=\rho_{\max}$, *we note that the characterization* $\Sigma^\star = \widehat{V}\Phi(\tau,\gamma^\star,\widehat{\lambda})\widehat{V}^\top$ *given by Theorem 2 is compatible with Proposition 1. To see this, recall that* $D\left(\sqrt{\frac{1}{\tau}}I,\widehat{\Sigma}\right) = \sum_{i=1}^p d\left(\sqrt{\frac{1}{\tau}},\widehat{\lambda}_i\right) = \rho_{\max}$. *Then by Proposition 4(iii), it holds that* $\gamma^\star = 0$, $\varphi(\tau,\gamma^\star,\widehat{\lambda}_i) = \sqrt{\frac{1}{\tau}}, i=1,\ldots,p$ *solve the equation* $\sum_{i=1}^p d(\varphi(\tau,\gamma^\star,\widehat{\lambda}_i),\widehat{\lambda}_i) - \rho_{\max} = 0$. *So the optimal solution characterized by Theorem 2 is* $\Sigma^\star = \sqrt{\frac{1}{\tau}}\widehat{V}\widehat{V}^\top = \sqrt{\frac{1}{\tau}}I$, *and it agrees with the result stated in Proposition 1.*

To construct the optimal solution $s^\star$ of (P-Vec), it remains to show that the nonlinear equation $\sum_{i=1}^p d(\varphi(\tau,\gamma^\star,\widehat{\lambda}_i),\widehat{\lambda}_i) - \rho = 0$ can be solved efficiently. Define $F_{\widehat{\lambda}} : \mathbb{R}_+ \to \mathbb{R}$ by $F_{\widehat{\lambda}}(\gamma) \triangleq \sum_{i=1}^p d(\varphi(\tau,\gamma,\widehat{\lambda}_i),\widehat{\lambda}_i)$. The following proposition reveals that $F_{\widehat{\lambda}}(\gamma) = \rho$ admits a unique positive root, and the equation can be solved efficiently by bisection or Newton's method.

Proposition 6 (Differentiable and strictly decreasing $F_{\widehat{\lambda}}$)**.** *If Assumptions 1- 4 hold, then the function* $F_{\widehat{\lambda}}$ *is differentiable and strictly decreasing over* $\mathbb{R}_+$. *If in addition, Assumption 5 holds, then* $F_{\widehat{\lambda}}(0) = \rho_{\max} \geq \rho$ *and* $\lim_{\gamma\to\infty} F_{\widehat{\lambda}}(\gamma) = 0 \leq \rho$.

To sum up the derivation so far, Theorem 1 reduces the robust problem (RO) to a more tractable formulation (P-Mat). Proposition 3 shows that the unique optimal solution to (P-Mat) can be constructed from $s^\star$, the optimal solution to (P-Vec). Proposition 5 provides a quasi-closed form of (P-Vec) parameterized by the dual variable $\gamma^\star$, which can be characterized by solving a univariate equation. Proposition 6 ensures that the equation can be solved efficiently. Finally, taking all these together proves Theorem 2, the construction of distributionally robust covariance estimators.

3.4 Nonlinear Shrinkage and Spectral Bias Correction

In this section, we show that the covariance matrix estimator $\Sigma^\star$ introduced in Theorem 2 can be interpreted as a nonlinear shrinkage estimator with $\sqrt{\frac{1}{\tau}}I$ as the shrinkage target and ρ as the shrinkage intensity. In Theorem 2, the estimator is characterized by the Lagrange multiplier $\gamma^\star$. To see how the estimator can be parameterized by the radius ρ when ρ varies in $(0, \rho_{\max}]$, we note that $\gamma^\star$ can be viewed as a function of ρ in the sense of

$$\gamma^\star_{\widehat{\lambda}}(\rho) \triangleq \text{the unique solution } \gamma^\star \geq 0 \text{ of the equation } F_{\widehat{\lambda}}(\gamma^\star) = \rho, \tag{12}$$

where $F_{\widehat{\lambda}}$ is defined as in Proposition 6. By Proposition 6, $\gamma^\star_{\widehat{\lambda}}(\rho_{\max}) = 0$, $\lim_{\rho \downarrow 0} \gamma^\star_{\widehat{\lambda}}(\rho) = \infty$, and $\gamma^\star_{\widehat{\lambda}}(\rho)$ is strictly decreasing in ρ. Moreover, it follows by (11) in Theorem 2 that

$$\Sigma^\star(\tau, \rho) = \widehat{V}\Phi(\tau, \gamma^\star_{\widehat{\lambda}}(\rho), \widehat{\lambda})\widehat{V}^\top \triangleq \widehat{V}\mathrm{diag}(\varphi(\tau, \gamma^\star_{\widehat{\lambda}}(\rho), \widehat{\lambda}_1), \ldots, \varphi(\tau, \gamma^\star_{\widehat{\lambda}}(\rho), \widehat{\lambda}_p))\widehat{V}^\top, \rho \in (0, \rho_{\max}]. \tag{13}$$

Here we write $\Sigma^\star(\tau, \rho)$ to emphasize its dependence on τ and ρ. Consequently, we can show that $\Sigma^\star(\tau, \rho)$ is a nonlinear shrinkage estimator, and the spectral bias of the estimator is corrected. The next theorem states these.

Theorem 3 (Nonlinear shrinkage and spectral bias correction)**.** *Let Assumptions 1-4 hold. Then for any $\tau > 0$, $\Sigma^\star(\tau, \rho)$ is continuous in ρ over $(0, \rho_{\max}]$ and $\lim_{\rho \downarrow 0} \Sigma^\star(\tau, \rho) = \widehat{\Sigma}$, $\Sigma^\star(\tau, \rho_{\max}) = \sqrt{\frac{1}{\tau}}I$. Moreover, the following assertions hold.*

(i) if $\widehat{\lambda}_i < \sqrt{\frac{1}{\tau}}$, then $\widehat{\lambda}_i < \lambda_i(\Sigma^\star(\tau, \rho))$ for any $\rho \in (0, \rho_{\max}]$,

(ii) if $\widehat{\lambda}_i > \sqrt{\frac{1}{\tau}}$, then $\widehat{\lambda}_i > \lambda_i(\Sigma^\star(\tau, \rho))$ for any $\rho \in (0, \rho_{\max}]$,

(iii) $\kappa(\Sigma^\star(\tau, \rho)) < \kappa(\widehat{\Sigma})$ for any $\rho \in (0, \rho_{\max}]$, and $\kappa(\Sigma^\star(\tau, \rho))$ is strictly decreasing on ρ.

Theorem 3 ensures that all the eigenvalues are shrunk towards $\sqrt{\frac{1}{\tau}}$, i.e., $\sqrt{\frac{1}{\tau}}I$ is the shrinkage target of $\Sigma^\star(\tau, \rho)$. The choice of τ balances the shrinkage effects stemming from two different sources: the Stein loss and the Frobenius loss, as modeled in (DRO). If τ is set large, then the shrinkage effect from the Frobenius loss is significant, and the estimation tends to shrink all the eigenvalues of $\Sigma^\star(\tau, \rho)$ to 0. If τ is set close to 0, on the other hand, then the shrinkage effect is dominated by the effect caused by the Stein loss, and thus the eigenvalues of the inverse of $\Sigma^\star(\tau, \rho)$, i.e., the precision matrix estimator, are nearly shrunk to 0. If we choose τ suitably so that $\sqrt{\frac{1}{\tau}}$ lying between interval $(\widehat{\lambda}_1, \widehat{\lambda}_p)$, then the shrinkage effects are balanced, the spectral bias (2) is corrected from both sides, and thereby the condition number of $\Sigma^\star(\tau, \rho)$ is improved strictly. Finally, the radius ρ controls the shrinkage intensity. When ρ is large, the eigenvalues of $\Sigma^\star(\tau, \rho)$ are close to $\sqrt{\frac{1}{\tau}}$, and thus condition number of $\Sigma^\star(\tau, \rho)$ is close to one. When ρ is small, the shrinkage effect is weak and $\Sigma^\star(\tau, \rho)$ is close to the nominal matrix $\widehat{\Sigma}$.

3.5 Covariance-Precision Matrix Estimator Induced by Spectral Convex Divergences

In this section, we concretize the previous results by taking the specific forms of convex spectral divergence D in Table 2. These divergences are widely used in statistics, machine learning, and

Divergence function	Eigenvalue mapping $\varphi(\tau,\gamma,b)$	Upper bound of $\gamma^\star$
Kullback-Leibler	$\frac{-\gamma+\sqrt{\gamma^2+8\tau(2+\gamma)b^2}}{4\tau b}$	$\max\left\{\frac{2p}{\rho}, \frac{2\tau\widehat{\lambda}_p^2+1}{e^{\rho/p}-1}\right\}$
Wasserstein	unique positive root a of $\tau a^2+\gamma a-\gamma\sqrt{b}\sqrt{a}-1=0$	$\max\left\{\frac{\eta_1+\sqrt{\eta_1^2+16\eta_1\tau^{2.5}}}{8\tau^2}, \frac{\eta_2+\sqrt{\eta_2^2+16\eta_2\tau^{2.5}}}{8\tau^2}\right\}$
Symmetrized Stein	unique positive root a of $2\tau b a^3+\gamma a^2-2ba-\gamma b^2=0$	$\sqrt{\max\left\{\frac{p\widehat{\lambda}_p^2}{2\rho\widehat{\lambda}_1^4}, \frac{p\widehat{\lambda}_p^2(1-\tau\widehat{\lambda}_p^2)^2}{2\rho\widehat{\lambda}_1^4}\right\}}$
Squared Frobenius	$\frac{\gamma b}{\tau+2\gamma}+\frac{\sqrt{\gamma^2b^2+\tau+2\gamma}}{\tau+2\gamma}$	$\frac{p}{4\rho}+\sqrt{\frac{p^2}{16\rho^2}+\tau\frac{p}{4\rho}} \vee \frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho}+\sqrt{\frac{p^2(1-\tau\widehat{\lambda}_p^2)^4}{16\rho^2}+\tau\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho}}$
Weighted Frobenius	$\frac{\gamma b}{\tau b+2\gamma}+\frac{\sqrt{\gamma^2b^2+b(\tau b+2\gamma)}}{\tau b+2\gamma}$	$\sqrt{\max\left\{\frac{p\widehat{\lambda}_p}{4\rho\widehat{\lambda}_1^2}, \frac{p\widehat{\lambda}_p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho\widehat{\lambda}_1^2}\right\}}$

Table 3: Eigenvalue mappings and upper bounds of dual variables for the divergence functions in Table 2. Here, $\eta_1=\frac{p}{\rho}$, $\eta_2=\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{\rho}$, and $a\vee b\triangleq\max\{a,b\}$.

engineering. For a more detailed review of these divergences, we refer readers to Yue et al. (2024, Section 4).

The following proposition shows that the setting of Theorem 2 covers all the divergences in Table 2. The proof is similar to that of Yue et al. (2024, Theorem 2), we skip the details.

Proposition 7 (Convex spectral divergence functions)**.** *All divergence functions listed in Table 2 satisfy Assumptions 2 and 3.*

Taking D as any divergence function in Table 2, we can construct the optimal solution to (RO) by Theorem 2. The key steps are (i) to determine the mapping of eigenvalues φ as a function of (τ,γ,b), and (ii) to determine the dual variable $\gamma^\star$ that solves the equation

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

Proposition 8 collects the forms of eigenvalue mapping φ and gives the suggested upper bound of $\gamma^\star$ so that we can use bisection to find it efficiently for all divergences in Table 2.

Proposition 8 (Estimator induced by convex divergences)**.** *If Assumptions 1, 4 and 5 hold, and τ is taken such that $\widehat{\lambda}_1<\sqrt{\frac{1}{\tau}}<\widehat{\lambda}_p$, then the eigenvalue mappings and upper bounds of dual variable $\gamma^\star$ corresponding to the divergence functions listed in Table 2 can be established as in columns 2-3 of Table 3.*

In the case of Kullback-Leibler, Squared Frobenius, and Weighted Frobenius, equation (10) reduces to a quadratic equation, and thus its root admits a simple expression. For the Wasserstein divergence and Symmetrized Stein divergence, the equation can only be reformulated as a cubic equation or a quartic equation. The expressions of the roots are complicated and not useful even for practical implementations. We suggest using a numerical method to find the root. Note that $\lim_{\gamma\to\infty}\varphi(\tau,\gamma,b)=$

b by Proposition 4(iii). Then, for Newton's method, b itself can be a good starting point when γ is large. For bisection, Proposition 4(i) ensures that the root always falls between $\sqrt{\frac{1}{\tau}}$ and b, which provides the bisection interval.

3.6 Consistency and Finite-sample Performance Guarantee

In this section, we demonstrate that the covariance-precision matrix estimator is consistent and possesses a finite-sample performance guarantee. To emphasize the dependence on sample size n, we let the nominal estimator and the radius in (RO) be chosen as $\widehat{\Sigma} = \widehat{\Sigma}_n$ and $\rho = \rho_n$, where $\widehat{\Sigma}_n$ is any covariance estimator constructed from n i.i.d. samples, and ρ_n is a non-negative radius that may depend on sample size n. The corresponding nonlinear shrinkage covariance matrix estimator in this case is denoted by $\Sigma_n^\star(\tau, \rho_n)$. In the following proposition, we recall that an estimator of Σ_0 is strongly consistent if it converges to Σ_0 almost surely as n tends to infinity.

Proposition 9 (Consistency)**.** *Suppose that Assumptions 3, 4 and 5 hold. If $\widehat{\Sigma}_n$ is a strongly consistent estimator of Σ_0 and ρ_n converges to 0 as n tends to infinity, then $\Sigma_n^\star(\tau, \rho_n)$ is strongly consistent for any $\tau > 0$.*

Proposition 9 states that if the nominal covariance estimator is consistent, the covariance-precision matrix estimator inherits the consistency as long as the radius ρ_n shrinks to 0 with n. A typical choice for a consistent nominal covariance estimator is the sample covariance matrix. In the next proposition, we recall the finite-sample performance guarantees of the ambiguity set $\mathcal{B}_\rho(\widehat{\Sigma}_n) = \{S \in \mathbb{S}_{++}^p : D(S, \widehat{\Sigma}_n) \leq \rho\}$ in model (RO) from Yue et al. (2024, Proposition 8) for completeness. Note that a probability distribution $\mathbb{P}$ is sub-Gaussian if there exists $\sigma^2 \geq 0$ with $\mathbb{E}_{\mathbb{P}}[\exp(z^\top \xi)] \leq \exp(\frac{1}{2}\sigma^2 \|z\|_2^2)$ for every $z \in \mathbb{R}^p$.

Proposition 10 (Finite-sample performance guarantee, (Yue et al., 2024, Proposition 8))**.** *Suppose that $\mathbb{P}$ is sub-Gaussian with covariance matrix $\Sigma_0 \in \mathbb{S}_+^p$, and let $\widehat{\Sigma}_n$ be the sample covariance matrix of n i.i.d. samples from $\mathbb{P}$. For any divergence function D from Table 2 and $\eta \in (0, 1)$, there exist $n_{\min}(\eta) = O(\log \eta^{-1})$ and $\rho_{\min}(n, \eta) = O(n^{-\frac{1}{2}}(\log \eta^{-1})^{\frac{1}{2}})$ that may depend on $\mathbb{P}$ such that $\mathbb{P}^n \left(\Sigma_0 \in \mathbb{B}_\rho(\widehat{\Sigma}_n)\right) \geq 1 - \eta$ for all $n \geq n_{\min}(\eta)$ and $\rho \geq \rho_{min}(n, \eta)$.*

Proposition 10 provides a potential guideline for us to set the radius of the ambiguity set. Let the nominal estimator be the sample covariance matrix $\widehat{\Sigma}_n$. The proposition ensures that there exist constants $c_1, c_2 > 0$ such that if we choose $\rho_n \geq c_1 n^{-\frac{1}{2}}$, then with probability at least $1 - \exp(-c_2 n)$, the optimal value of the robust problem (RO) is a upper bound on the actual estimation loss $f(\Sigma_0, \Sigma^\star(\tau, \rho_n), \Sigma^\star(\tau, \rho_n)^{-1})$. The argument of finite-sample performance guarantee has been widely discussed in the DRO literature; see the discussion in the paper of Wasserstein distributionally robust precision matrix estimation model (Nguyen et al., 2022). In a recent follow-up paper, however, Blanchet and Si (2019) point out that an $n^{-\frac{1}{2}}$-ordered radius could be too conservative and does not agree with the empirical findings in Nguyen et al. (2022). They propose a customized radius tuning scheme by minimizing the estimation loss of the precision matrix estimator induced by the radius and show that the asymptotically optimal choice of the radius is of the order n^{-1}. Inspired by their theoretical findings, in the next section, we propose a parameter tuning framework for our covariance-precision matrix estimator and show that the order of the optimal choice of radius is also more optimistic than $n^{-\frac{1}{2}}$ in our case.

4 Optimal Shrinkage Target and Intensity

In this section, we propose a parameter tuning scheme for the nonlinear shrinkage covariance matrix estimator $\Sigma^\star(\tau,\rho)$. The parameter tuning can be viewed as deciding the shrinkage target $\sqrt{\frac{1}{\tau}}I$ and shrinkage intensity ρ such that $\Sigma^\star(\tau,\rho)$ is close to the true covariance Σ_0 under some measure such as the Frobenius norm. Ledoit and Wolf (2004) used this idea to construct the optimal linear shrinkage estimator:

$$\Sigma^{LW}(\nu,t) \triangleq t\nu I + (1-t)\widehat{\Sigma},$$

where the authors choose the shrinkage target νI and intensity t of the linear shrinkage estimator by solving following problem:

$$\min_{\nu\in\mathbb{R}_+,t\in[0,1]} \mathbb{E}_n[\|\Sigma^{LW}(\nu,t)-\Sigma_0\|_F^2]. \tag{14}$$

Here $\mathbb{E}_n$ denotes the expectation with respect to the n-product measures of the data-generating distribution. To solve this problem, the authors show that it is equivalent to take two steps: (i) find the optimal target by

$$\nu^\star = \arg\min_{\nu\in\mathbb{R}_+} \|\Sigma_0-\nu I\|_F, \tag{15}$$

and (ii) find the optimal intensity by

$$t^\star = \arg\min_{t\in[0,1]} \mathbb{E}_n[\|\Sigma^{LW}(\nu^\star,t)-\Sigma_0\|_F^2]. \tag{16}$$

The optimization problems in both steps are convex quadratic programs and have a closed-form expression for the optimal solution. To choose the parameter of $\Sigma^\star(\tau,\rho)$, one may mimic the approach to determine (τ,ρ) by the following minimization problem

$$\min_{\tau\in\mathbb{R}_{++},\rho\in(0,\rho_{\max}]} \mathbb{E}_n[\|\Sigma^\star(\tau,\rho)-\Sigma_0\|_F^2] \tag{17}$$

and solve the problem in two steps. Unfortunately, it turns out that each step is a nonconvex problem in general and the optimal solution does not have a closed form like (14). This is because $\Sigma^\star(\tau,\rho)$ is nonlinear in τ and ρ (see Table 3 for specific form of φ), and the quadratic form of $\Sigma^\star(\tau,\rho)$ is hard to analyze.

In the forthcoming discussions of this section, we follow the general framework of the parameter tuning procedure of solving (14), i.e., first finding the optimal shrinkage target and then, with this fixed target, finding the optimal intensity but with some important variations. To formulate a tractable parameter-tuning problem, we propose to use the violation of the optimality of (P-Mat) evaluated at $\Sigma=\Sigma_0$, which depends on τ and ρ, as a measure of distance from the target $\sqrt{\frac{1}{\tau}}I$ and the estimator $\Sigma^\star(\tau,\rho)$ to the ground truth Σ_0. Such a concept of the violation of optimality is widely used in inverse optimization, where a decision maker aims to infer the parameters of an optimization model so that the optimal solution is equal to or at least close to the solution we have observed. A typical approach is to minimize the violation of the optimality evaluated at the observed solution by tuning the parameters. See Chan et al. (2025) for a survey of inverse optimization. We note that the choice of the optimal shrinkage target and intensity can be viewed as an inverse optimization task: we want to

choose the parameters τ and ρ properly so that the ground true optimal solution Σ_0 is nearly optimal to (P-Mat), and thus the distance between $\Sigma^\star(\tau,\rho)$ and Σ_0 is close. To simplify the derivation, the nominal covariance estimator $\widehat{\Sigma}$ is chosen as the sample covariance matrix $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^\top$ of n i.i.d. samples in this section, where we use the subscript n in $\widehat{\Sigma}_n$ to emphasize its dependence on sample size n. The corresponding estimator is denoted by $\Sigma_n^\star(\tau,\rho)$ throughout this section.

4.1 Optimal Shrinkage Target

We begin by identifying the optimal shrinkage target which minimizes the violation of the optimality of (P-Mat) evaluated at $\Sigma = \Sigma_0$. Let $\ell_\tau(\Sigma) \triangleq \log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2$ be the objective function of problem (P-Mat), where we make explicit the dependence of the objective function on the parameter τ. Proposition 1 asserts that the shrinkage target $\sqrt{\frac{1}{\tau}}I$ is the unique maximizer of (P-Mat) when the constraint is redundant, i.e., it solves $\max_{\Sigma\in\mathbb{S}_+^p}\ell_\tau(\Sigma)$. Thus to choose τ so that $\sqrt{\frac{1}{\tau}}I$ is close to Σ_0, we minimize the violation of the optimality evaluated at $\Sigma = \Sigma_0$, which is captured by the norm of the gradient $\nabla_\Sigma\ell_\tau(\Sigma_0)$:

$$\tau^\star \triangleq \arg\min_{\tau\in\mathbb{R}_+} \|\nabla_\Sigma\ell_\tau(\Sigma_0)\|_F. \tag{18}$$

Observe that $\|\nabla_\Sigma\ell_\tau(\Sigma_0)\|_F^2 = \|\Sigma_0^{-1} - \tau\Sigma_0\|_F^2$, which is a quadratic function of τ. Thus, we can easily obtain the optimal solution of the problem $\tau^\star = \frac{p}{\|\Sigma_0\|_F^2}$. The induced optimal shrinkage target is

$$\Sigma_{\text{target}} = \sqrt{\frac{1}{\tau^\star}}I = \frac{\|\Sigma_0\|_F}{\sqrt{p}}I. \tag{19}$$

Since $\left\|\frac{\|\Sigma_0\|_F}{\sqrt{p}}I\right\|_F = \|\Sigma_0\|_F$, then the target can be viewed as a normalized guess of Σ_0 with the same scale as Σ_0, i.e., the total scale is distributed evenly to each dimension. This is a reasonable prior when we only have access to the scale of Σ_0, or an estimation of it. In practice, if we approximate $\|\Sigma_0\|_F$ by $\|\widehat{\Sigma}_n\|_F$, then the approximation $\widehat{\tau}^\star \triangleq \frac{p}{\|\widehat{\Sigma}_n\|_F^2}$ satisfies $\widehat{\lambda}_1 \le \sqrt{\frac{1}{\widehat{\tau}^\star}} \le \widehat{\lambda}_p$. As discussed after Theorem 3, $\widehat{\tau}^\star$ well balances the shrinkage effects of the Stein loss and the Frobenius loss in (DRO), and corrects the spectral bias of $\widehat{\Sigma}_n$.

4.2 Optimal Shrinkage Intensity (Radius) and its Limit Behavior

With the optimal target chosen as above, we tune the radius ρ so that the estimator $\Sigma_n^\star(\tau^\star,\rho)$ is close to the true covariance Σ_0. Recall that $\Sigma^\star(\tau^\star,\rho)$ is the optimal solution of the following problem:

$$\max_{\Sigma\in\mathbb{S}_{++}^p : D(\Sigma,\widehat{\Sigma}_n)\le\rho} \log\det\Sigma - \frac{1}{2}\tau^\star\|\Sigma\|_F^2. \tag{20}$$

In the following, we first derive the optimality condition of (20), and then find the optimal radius ρ by minimizing the violation of the optimality evaluating at $\Sigma = \Sigma_0$. The Lagrangian function of (20) is

$$L(\Sigma,\gamma) \triangleq \log\det\Sigma - \frac{1}{2}\tau^\star\|\Sigma\|_F^2 - \gamma(D(\Sigma,\widehat{\Sigma}_n) - \rho),$$

where $\gamma \geq 0$ is the dual variable. Then the optimiality condition of (20) is given by the Karush-Kuhn-Tucker condition

$$\nabla_1 L(\Sigma, \gamma) = \Sigma^{-1} - \tau^\star \Sigma - \gamma \nabla_1 D(\Sigma, \widehat{\Sigma}_n) = 0, \tag{21}$$

$$\gamma(D(\Sigma, \widehat{\Sigma}_n) - \rho) = 0, \tag{22}$$

$$\gamma \geq 0, \Sigma \in \mathbb{S}^p_{++}, \tag{23}$$

where ∇_1 represents the gradient with respect to the first argument. By the construction of $\Sigma^\star_n(\tau^\star, \rho)$ and $\gamma^\star_{\widehat{\lambda}}(\rho)$, $(\Sigma = \Sigma^\star_n(\tau^\star, \rho), \gamma = \gamma^\star_{\widehat{\lambda}}(\rho))$ solves (21)-(23), where $\gamma^\star_{\widehat{\lambda}}(\rho)$ is the optimal dual solution as defined in (12). It is intuitive to see $\|\nabla_1 L(\Sigma, \gamma^\star_{\widehat{\lambda}}(\rho))\|_F$ as a measure of the violation of the optimality evaluated at Σ. This prompts us to consider the ρ that minimizes the violation evaluated at $\Sigma = \Sigma_0$, that is,

$$\rho^\star(\widehat{\Sigma}_n) \triangleq \arg \min_{\rho \in (0, \rho_{\max}]} \|\Sigma_0^{-1} - \tau^\star \Sigma_0 - \gamma^\star_{\widehat{\lambda}}(\rho) \nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)\|_F^2. \tag{24}$$

The above defines a sample-dependent optimal radius, since both $\gamma^\star_{\widehat{\Sigma}}$ and $\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)$ depend on the realization of $\widehat{\Sigma}_n$. In practice, however, the radius should be set independently of the samples to facilitate simulations in both model training and validation. To tackle the issue, we replace $\gamma^\star_{\widehat{\lambda}}(\rho)$ with $\gamma_\lambda(\rho)$, which is defined as the unique root $\gamma^\star$ of $\sum_{i=1}^d d(\varphi(\tau^\star, \gamma^\star, \lambda_i), \lambda_i) = 0$, and $\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)$ with $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]$. Consequently, we consider

$$\rho^\star_n \triangleq \arg \min_{\rho \in \mathbb{R}_{++}} \|\Sigma_0^{-1} - \tau^\star \Sigma_0 - \gamma_\lambda(\rho) \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F^2. \tag{25}$$

The replacement of $\gamma^\star_{\widehat{\lambda}}$ is justified by the fact that $\gamma^\star_{\widehat{\lambda}}(\rho) \to \gamma^\star_\lambda(\rho)$ as $n \to \infty$ with probability 1. The optimal intensity defined by (25) is still not achievable in that Σ_0 is unknown. However, the formulation enables us to derive the limiting behavior of $\rho^\star_n$ as $n \to \infty$. It reveals that an asymptotic optimal intensity is in the form $\rho^\star / n^2$, where $\rho^\star$ is a constant that can be figured out explicitly. This kind of analysis for the optimal radius was also studied by Blanchet and Si (2019). Before stating the result, we make some assumptions about the divergence functions.

Assumption 6 (Locally-quadratic divergence)**.** *For any $b > 0$, there exists positive constant $C_{b,d}$ depending on the generator d of the divergence and b such that*

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

Assumption 7 (Non-degenerate gradient of divergence function)**.** *For any sample size n, the gradient of divergence function D satisfies $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \neq 0$ and there exists strictly positive constants C_1 and C_2 such that*

$$\lim_{n \to \infty} n \|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = C_1 \tag{26}$$

and

$$\lim_{n \to \infty} n \left\langle \Sigma_0^{-1} - \tau^\star \Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \right\rangle = C_2. \tag{27}$$

Condition $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \neq 0$ means that Σ_0 is not a minimizer of $\mathbb{E}_n[D(\Sigma, \widehat{\Sigma}_n)]$ as

$$\nabla_1 \mathbb{E}_n[D(\Sigma, \widehat{\Sigma}_n)]|_{\Sigma=\Sigma_0} = \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \neq 0,$$

provided that the differential operator ∇_1 and $\mathbb{E}_n$ are interchangeable. This condition is not satisfied when $\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)$ is linear in the second argument. For example, the squared Frobenius and weighted Frobenius divergence functions do not satisfy the condition because $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] = 0$ for all n. Condition (26) requires the norm of the expected gradient $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]$ goes to zero at rate $1/n$ as $n \to \infty$. Condition (27) means that inner product between $\Sigma_0^{-1} - \tau^\star \Sigma_0$ and $\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]$ goes to zero at rate $1/n$. Here we specify the conditions in terms of the gradient instead of the divergence function itself since we apply the first-order optimality condition to the definition of $\rho_n^\star$. It can be verified that the first three divergence functions in Table 2 satisfy Assumptions 6 and 7, except Squared Frobenius and Weighted Frobenius divergence functions. With the above two assumptions, we are ready to state the asymptotic behavior of $\rho_n^\star$ in the next theorem.

Theorem 4 ($1/n^2$-order optimal radius)**.** *Under Assumptions 2, 3, 6 and 7, there exists $\rho^\star > 0$ such that the optimal intensity $\rho_n^\star$ defined as in (25) satisfies*

$$\lim_{n\to\infty} n^2 \rho_n^\star = \rho^\star, \tag{28}$$

where the limit $\rho^\star$ depends on the choice of divergence function D and the constants C_1 and C_2.

The above theorem implies that when the sample size is large, the asymptotically optimal radius tends to 0 and is of the order $\rho^\star/n^2$. We emphasize that a $1/n^2$-order radius of our model is parallel with optimal radius tuning schemes proposed in the literature. In Nguyen et al. (2022), the authors consider a distributionally robust inverse covariance estimation model with Wasserstein distance, which is a square root version of Wasserstein divergence in Table 2. Then, Blanchet and Si (2019, Theorem 1) showed that the optimal radius is of rate $1/n$ in the sense of minimizing Stein's loss of the proposed estimator. This is of the same order as $\rho_n^\star$ after taking the square.

4.3 Optimal Shrinkage Intensity of Convex Divergences

In this section, we discuss special cases of Theorem 4, where D is chosen as the Kullback-Leibler divergence, Wasserstein divergence, and Symmetrized Stein divergence as listed in Table 2. To obtain a concrete expression of the constants C_1, C_2 and $\rho^\star$, we consider a specific probability distribution of the underlying random vector ξ. The next assumption specifies this.

Assumption 8 (Normal distribution)**.** *The random vector ξ follows the normal distribution $\mathcal{N}(0, \Sigma_0)$ with $\Sigma_0 \neq \sigma I$ for any $\sigma \in \mathbb{R}_{++}$.*

The next corollary asserts that under the normal distribution assumption, Assumptions 6 and 7 hold for divergences of Kullback-Leibler, Wasserstein, and Symmetrized Stein, and give explicit expressions of the limit constant $\rho^\star$ for each type of divergence.

Corollary 1 (Asymptotic convergence rate of the optimal radius)**.** *Let $\tau^\star$ and $\rho_n^\star$ be defined as in (18) and (25) respectively. Under Assumption 8, the following assertions hold.*

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

$$\lim_{n\to\infty} n^2 \rho_n^\star = \rho_{KL}^\star \triangleq \frac{(p+1)^2 \|\Sigma_0^{-1}\|_F^4}{16\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2} \sum_{i=1}^{p} \left(1 - \tau^\star \lambda_i^2\right)^2. \tag{29}$$

(ii) If D is taken as Wasserstein divergence, then

$$\lim_{n\to\infty} n^2\rho_n^\star = \rho_W^\star \triangleq \frac{(p+1)^2p^2}{256\left(\mathrm{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\mathrm{Tr}(\Sigma_0)\right)^2}\sum_{i=1}^{p}\frac{(1-\tau^\star\lambda_i^2)^2}{\lambda_i}. \tag{30}$$

(iii) If D is taken as symmetrized Stein divergence, then

$$\lim_{n\to\infty} n^2\rho_n^\star = \rho_{SS}^\star \triangleq \frac{(p+1)^2\|\Sigma_0^{-1}\|_F^4}{32\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2}\sum_{i=1}^{p}\left(1-\tau^\star\lambda_i^2\right)^2. \tag{31}$$

To apply the optimal target and the optimal radius in practice, we may use the sample covariance $\widehat{\Sigma}_n$ and empirical eigenvalues to approximate the coefficient and constants in (19) and (29)-(31). For given sample size n, the target can be set as $\frac{\|\widehat{\Sigma}_n\|_F}{\sqrt{p}}I$ and the radius can be set as $\widehat{\rho}_n = \widehat{\rho}^\star/n^2$, where $\rho^\star$ is approximated by $\widehat{\rho}^\star$ with empirical counterpart of Σ_0 and $\lambda_1,\ldots,\lambda_p$. As n is sufficiently large, the approximations are close to their true counterparts since $\widehat{\Sigma}_n$ is a consistent estimator of Σ_0.

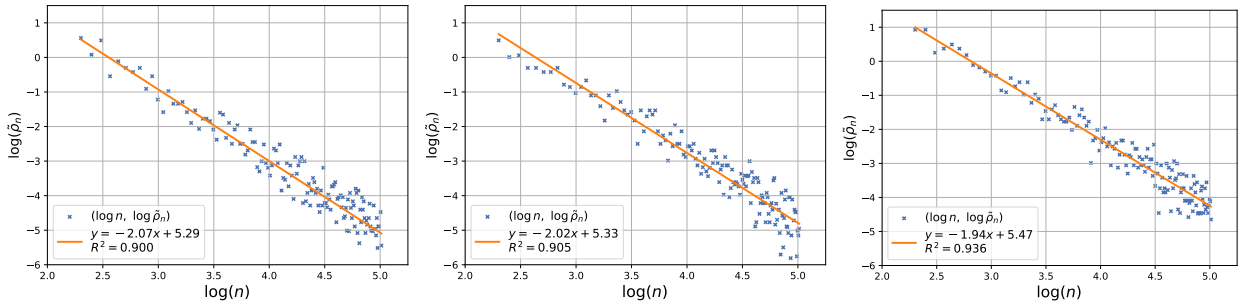
5 Numerical Experiments

In this section, we conduct numerical experiments to compare the performance of the proposed nonlinear shrinkage estimator SCOPE with the classical Ledoit-Wolf's linear shrinkage estimator (LW-L) by Ledoit and Wolf (2004), the state-of-the-art nonlinear shrinkage estimator (LW-NL) recently proposed by Ledoit and Wolf (2020a), the Wasserstein-based distributionally robust covariance matrix estimator (W-DRCOE) by Yue et al. (2024), and the Wasserstein-based distributionally robust precision matrix estimator (WISE) by Nguyen et al. (2022). As mentioned in (14), the optimal target and optimal intensity of Ledoit-Wolf's linear shrinkage have closed-form expressions and can be directly applied. The implementations of LW-L, LW-NL, W-DRCOE, and WISE are based on the open-source code provided by the authors. In the experiments, we focus on SCOPEs induced by the KL divergence (KL-SCOPE), the Wasserstein divergence (W-SCOPE), and the symmetrized Stein divergence (SS-SCOPE), all of which fall within the framework of Corollary 1.

In the first experiment, we provide an empirical validation of the order of the asymptotically optimal radius proposed in Section 4.2. After that, we compare SCOPEs with other estimators by assessing their estimation error with synthetic data. In the subsequent experiments, we use real data and synthetic data to compare the performance of the estimators in practical applications, including anomaly detection, A/B tests, and minimum variance portfolio selection. The implementation of our methods can be found at `https://github.com/searsh/SCOPE`.

5.1 Validation of Optimal Radius

In the first experiment, we validate the order of optimal radius proposed in Theorem 4 and Corollary 1, i.e., whether the optimal radius is of order n^{-2}, via grid search on radius. To this end, we set $p = 5$ and the true covariance matrix is generated via $\Sigma_0 = V^\top\Lambda V \in \mathbb{S}_{++}^p$, where $\Lambda = \mathrm{diag}(1,2,\ldots,p)$ and V is an orthogonal matrix. For fixed sample size n, we draw n samples from $\mathbb{P} = \mathcal{N}(0,\Sigma_0)$ and construct the sample covariance matrix $\widehat{\Sigma}_n$. Then, taking $\widehat{\Sigma}_n$ as the nominal matrix, we construct KL-SCOPE, W-SCOPE, and SS-SCOPE, denoted by $\Sigma_n^\star(\widehat{\tau}_n^\star,\rho_n)$, where we set $\widehat{\tau}_n^\star = p/\|\widehat{\Sigma}_n\|_F^2$. We write $\Sigma_n^\star$ for



(a) KL divergence (b) Wasserstein divergence (c) Symmetrized Stein divergence

Figure 2: Log–log regression of optimal radius $\widetilde{\rho}_n$ vs. sample size n under divergences.

short. Recall that $\tau^\star = p/\|\Sigma_0\|_F^2$ and the combined Stein-Frobenius loss of the estimator $\Sigma_n^\star$ is defined by

$$\ell(\Sigma_n^\star, \Sigma_0) \triangleq -\log\det\left((\Sigma_n^\star)^{-1}\right) + \left\langle (\Sigma_n^\star)^{-1}, \Sigma_0 \right\rangle + \frac{1}{2}\tau^\star\left(\|\Sigma_n^\star\|_F^2 - 2\langle \Sigma_n^\star, \Sigma_0\rangle\right). \tag{32}$$

For the choice of ρ_n, we do a grid search over $\rho_n \in [10^{-5}, 2\times 10^3]$, and find the best radius $\widetilde{\rho}_n$ that minimizes the average loss over 50 independent repeats of construction of $\widehat{\Sigma}_n$ and $\widehat{\Sigma}_n^\star$. We repeat the above procedure for sample size n over $\{10, 11, \ldots, 150\}$, and record the corresponding optimal radius $\widetilde{\rho}_n$. To find out the dependence of $\widetilde{\rho}_n$ on n, we conduct a log-log-linear regression between $\widetilde{\rho}_n$ and n. That is, we use a straight line $y = \alpha x + \beta$ to fit the points $(\log n, \log \widetilde{\rho}_n)$. Figure 2 visualizes the regression results. It is shown that the points are well fitted by the line ($R^2 > 0.9$), and the slopes of the lines in all three cases are close to -2. The results indicate that the dependence of $\widetilde{\rho}_n$ on n can be approximated by $\exp(\log \widetilde{\rho}_n) \approx \exp(-2\log n + \beta)$, and thus $\widetilde{\rho}_n \sim c \cdot n^{-2}$. This validates the theoretical results in Section 4.2.

5.2 Estimation Error

In this experiment, we use synthetic data to compare SCOPEs with other estimators proposed in the literature. The true covariance matrix is generated via $\Sigma_0 = V^\top \Lambda V \in \mathbb{S}_{++}^p$, where $\Lambda = \mathrm{diag}(1, 2, \ldots, p)$ and V is an orthogonal matrix. For fixed sample size n, we draw n samples from $\mathbb{P} = \mathcal{N}(0, \Sigma_0)$ and construct the the sample covariance matrix (Sample) $\widehat{\Sigma}_n$ as the nominal estimator of Σ_0. With $\widehat{\Sigma}$, we generate the Wasserstein-based SCOPE (W-SCOPE), Ledoit-Wolf's linear estimator (LW-L), Ledoit-Wolf's nonlinear estimator (LW-NL), Wasserstein-based distributionally robust covariance estimator (W-DRCOE), and Wasserstein-based distributionally robust precision matrix estimator (WISE). To make a fair comparison, the sizes of ambiguity sets of SCOPEs, W-DRCOE, and WISE are all tuned via grid search over $\rho_n \in [10^{-5}, 2\times 10^3]$ as in Section 5.1. Specifically, let $\Sigma_0' \in \mathbb{S}_{++}^p$ be the true covariance matrix used in grid search. For different sample size n, we draw n i.i.d. samples from $\mathcal{N}(0, \Sigma_0')$ repeatedly. For W-SCOPE, we generate the covariance matrix estimator from the samples repeatedly and choose the radius that minimizes the average combined Stein-Frobenius loss $\ell(\Sigma_n^\star, \Sigma_0)$. For W-DRCOE, we generate the covariance matrix estimator from the samples repeatedly and choose the radius that minimizes the average Frobenius loss. For WISE, we generate the precision matrix estimator from the samples repeatedly and choose the radius that minimizes the average Stein's loss.

The numerical experiments are conducted for $p \in \{150, 200\}$ and various sample sizes n. The estimation error is measured by the combined Stein-Frobenius loss. Moreover, to evaluate the estimators' ability to correct the spectral bias, we compute the relative error of the eigenvalues of the estimators. Let $\{\lambda_i\}_{i=1}^p$ and $\{\widehat{\lambda}_i\}_{i=1}^p$ be the eigenvalues of Σ_0 and the estimator, respectively, both in decreasing order. We define the relative error of the eigenvalue estimation by

$$\frac{1}{p}\sum_{i=1}^{p}\frac{|\widehat{\lambda}_i - \lambda_i|}{\lambda_i}.$$

Table 4-5 display the combined Stein-Frobenius loss of the estimators. The loss is recorded as N/A when the covariance matrix estimator is singular. It is shown that in data-deficient cases, i.e., when $n < p$, only LW-L, WISE, and W-SCOPE provide invertible estimators of the covariance matrix, and W-SCOPE achieves comparable or smaller loss than WISE. Table 6-7 display the relative error of the eigenvalue estimation. The result reveals that W-SCOPE achieves better relative error than other estimators except LW-NL. Moreover, when $n < p$, W-SCOPE achieves the smallest relative error among all the invertible estimators. The result of this experiment shows that W-SCOPE is especially useful when data is insufficient. It provides a choice when both the covariance matrix and the precision matrix are to be estimated.

Sample size	50	70	100	120	150	200	300
Sample	N/A	N/A	N/A	N/A	2.0×10^9	1136.87	861.87
LW-L	796.62	795.59	794.37	793.50	792.35	790.52	787.23
LW-NL	N/A	N/A	N/A	N/A	N/A	784.86	777.04
W-DRCOE	N/A	N/A	N/A	N/A	2.0×10^9	1150.21	866.78
WISE	824.16	826.39	816.29	810.30	803.58	797.10	787.99
W-SCOPE	941.90	868.45	813.88	799.02	791.75	788.27	782.86

Table 4: Combined Stein–Frobenius loss of the estimators for $p = 150$.

Sample size	50	100	120	150	170	200	300	400
Sample	N/A	N/A	N/A	N/A	N/A	3.1×10^5	1379.82	1203.59
LW-L	1119.95	1117.72	1116.86	1115.62	1114.81	1113.59	1109.91	1106.74
LW-NL	N/A	N/A	N/A	N/A	N/A	N/A	1099.36	1092.63
W-DRCOE	N/A	N/A	N/A	N/A	N/A	3.1×10^5	1392.79	1212.47
WISE	1153.17	1151.09	1145.93	1139.13	1136.12	1131.08	1115.28	1106.69
W-SCOPE	1789.84	1347.02	1252.64	1170.21	1141.70	1120.44	1105.90	1100.83

Table 5: Combined Stein–Frobenius loss of the estimators for $p = 200$.

Sample size	50	70	100	120	150	200	300
Sample	0.97	0.85	0.73	0.67	0.59	0.49	0.36
LW-L	1.79	1.75	1.64	1.57	1.48	1.38	1.18
LW-NL	0.57	0.46	0.31	0.29	0.57	0.46	0.30
W-DRCOE	0.97	0.85	0.73	0.67	0.57	0.48	0.36
WISE	1.71	2.63	1.12	1.06	0.99	0.97	0.71
W-SCOPE	0.83	0.75	0.71	0.72	0.98	0.70	0.59

Table 6: Relative error of eigenvalue estimation of the estimators for $p = 150$.

Sample size	50	100	120	150	170	200	300	400
Sample	1.07	0.83	0.77	0.69	0.65	0.59	0.45	0.36
LW-L	2.01	1.84	1.80	1.71	1.67	1.60	1.41	1.27
LW-NL	0.66	0.44	0.37	0.31	0.31	0.56	0.40	0.30
W-DRCOE	1.07	0.83	0.77	0.69	0.65	0.55	0.44	0.37
WISE	2.23	1.52	1.47	1.43	1.46	1.38	0.68	0.52
W-SCOPE	0.92	0.68	0.63	0.57	0.54	0.59	0.72	0.54

Table 7: Relative error of eigenvalue estimation of the estimators for $p = 200$.

5.3 Application in Anomaly Detection of Hyperspectral Images

Hyperspectral imagery, typically used in remote sensing, agriculture, and environmental monitoring, provides detailed data of observed items via hundreds of continuous spectral bands. Anomaly detection in hyperspectral images seeks to identify materials(pixels) different from the surroundings, spatially or spectrally. Reed–Xiaoli (RX) detector (Reed and Yu, 1990) is a well-known statistical tool that models anomaly detection as an unsupervised binary classification problem. It assumes the background pixels of a hyperspectral image follow a multivariate normal distribution $\mathcal{N}(\mu_0, \Sigma_0)$. For every pixel in the image, the RX-detector calculates the Mahalanobis distance of the pixel's spectral vector x from the mean of the background distribution, defined by

$$d_M(x) = (x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0).$$

The distance measures how far the pixel vector x is away from the background. If the Mahalanobis distance is larger than a threshold set by the statistician, the pixel is identified as an anomaly. In practice, the true covariance Σ_0 is unknown and needs to be estimated from the data(all pixels). Moreover, with the development of sensing techniques, the increase of spectral bands in hyperspectral images makes anomaly detection more difficult, and calls for a high-dimensional covariance estimator when using the RX-detector.

In this experiment, we use RX-detectors induced by several covariance/precision estimators, i.e., sample covariance matrix (Sample), Ledoit-Wolf's linear estimator (LW-L), Ledoit-Wolf's nonlinear estimator (LW-NL), Wasserstein-based SCOPE (W-SCOPE), Kullback-Leibler-based SCOPE (KL-SCOPE), Symmetrized Stein-Based SCOPE (SS-SCOPE), Wasserstein-based distributionally robust covariance matrix estimator (W-DRCOE), and Wasserstein-based distributionally robust precision matrix estimator (WISE) to conduct anomaly detection on the sensing dataset collected from Lin et al.

(2022). The dataset consists of five real hyperspectral images, namely San Diego, Pavia, HYDICE, Texas Coast, and SpecTIR, with labeled anomalies. To compare the detectors' performance under high-dimensional regimes, we choose sub-images of these five images so that the number of samples used to estimate covariance is nearly equal to the dimension of spectral vectors. Figure 3 visualizes the tested images. The first row lists the original images, the second row shows the labeled anomalies, and the red boxes in the first row highlight the sub-images we chose to conduct the detections. The sub-images focus on where the anomaly occurs, and also contain less background information.

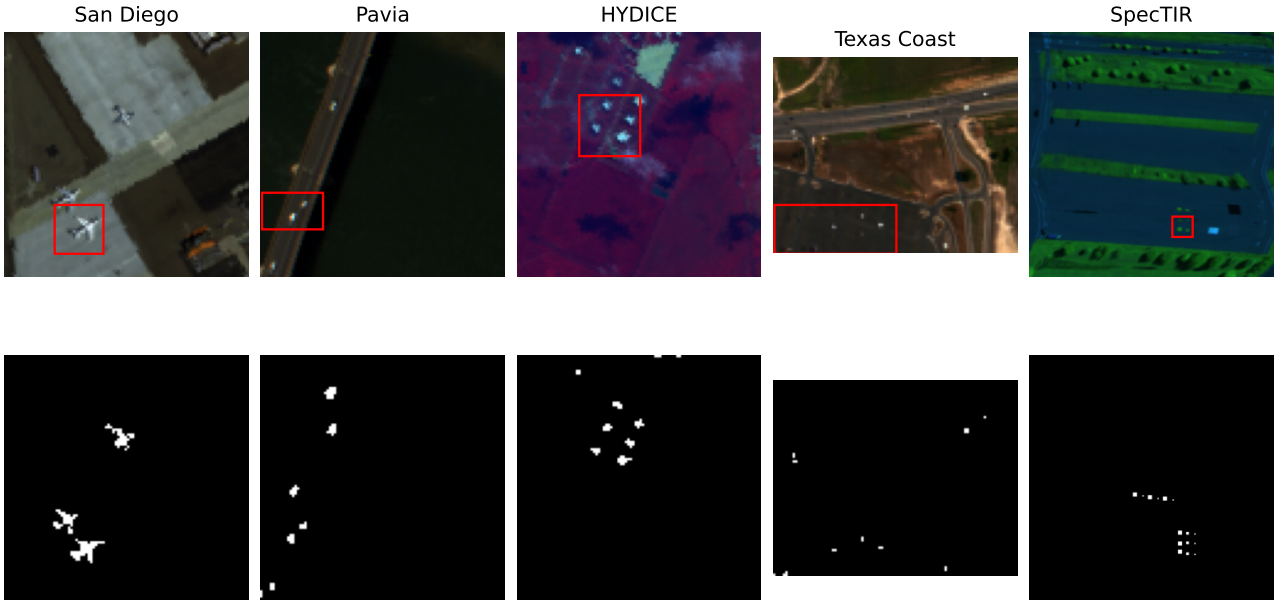


Figure 3: Hyperspectral images used in detection. The red boxes highlight the sub-images we chose to conduct the detections.

To compare the performance of different detectors, we report the area under the curve(AUC) values of these binary classifiers in Table 8. We observe that the RX-detectors induced by SCOPEs outperform others throughout all five images, and the advantage is more significant when the sample covariance performs poorly. This highlights the suitability of our estimators in high-dimensional regimes, where sample covariance fails to give a meaningful estimation of true covariance.

Estimator	San Diego	Pavia	HYDICE	Texas Coast	SpecTIR
Sample	0.7718	0.8866	0.9980	0.8891	0.9621
LW-L	0.9529	0.9567	0.9846	0.8666	0.9940
LW-NL	0.9132	0.9693	0.9971	0.9100	0.8191
W-DRCOE	0.7719	0.8867	0.9969	0.9035	0.9621
WISE	0.7718	0.8866	0.9980	0.8891	0.9621
KL-SCOPE	0.9794	**0.9893**	0.9976	**0.9101**	0.9940
W-SCOPE	0.7803	0.8955	**0.9986**	0.8891	0.9687
SS-SCOPE	**0.9760**	0.9824	0.9982	0.9006	**0.9956**

Table 8: AUC values of the detectors induced by covariance estimators under five hyperspectral images.

5.4 Application in A/B Tests

In this experiment, we consider the application of the covariance estimator in constructing a sensitive metric of A/B tests. A/B testing is an online controlled experiment where users are split into two groups—the control (A) and the treatment (B) to compare performance between a current production system (software, user interface, etc.) and a new variant. This method uses statistical hypothesis tests to determine whether differences in user engagement metrics are significant enough to infer a true user preference. In modern usage of A/B tests, there are always several relevant features (of the system) being taken into consideration, and it is essential to find a combination of these features that is sensitive to model changes. Let $\mu_A \in \mathbb{R}^p$ be the mean of feature vectors collected from group A, and $\mu_B \in \mathbb{R}^p$ be the mean of feature vectors collected from group B, both with p features. The literature aims to find a linear combination of features $\theta(x) \triangleq w^\top x$ that is sensitive to model performance improvements when switching from A to B. To avoid confusion and fix the direction of the combination coefficient w, we stipulate that $\theta(\mu_B) > \theta(\mu_A)$ when the treatment model (B) performs better than the control model (A), and denote it by $B \succ A$.

In Kharitonov et al. (2017), the authors propose a machine learning framework that learns sensitive feature combinations from historical data of multiple experiments with known preferences. Without loss of generality, we assume $B_k \succ A_k$ for all K experiments. Let $x_i^k, y_j^k \in \mathbb{R}^p, i = 1, ..., n_A, j = 1, ..., n_B$ be feature samples collected from group A_k and group B_k, respectively. Let $\widehat{\mu}_{A_k}, \widehat{\mu}_{B_k}$ and $\widehat{\Sigma}_{A_k}, \widehat{\Sigma}_{B_k}$ be the sample mean vectors and sample covariance matrices of group A_k and group B_k. The z-score of pair (A_k, B_k) under combination coefficient w is

$$z(w, A_k, B_k) = \frac{w^\top(\widehat{\mu}_{B_k} - \widehat{\mu}_{A_k})}{\sqrt{w^\top \widehat{\Sigma}_{A_k} w + w^\top \widehat{\Sigma}_{B_k} w}}. \tag{33}$$

Then the optimal combination coefficient for experiment k in the sense of maximizing the z-score of combined features is given by

$$w_k^\star = \alpha \cdot (\widehat{\Sigma}_{A_k} + \widehat{\Sigma}_{B_k} + \rho I)^{-1}(\widehat{\mu}_{B_k} - \widehat{\mu}_{A_k}) \tag{34}$$

for any scalar $\alpha > 0$. The regularization term ρI in (34) is used to deal with the singularity or ill-conditioning of $\widehat{\Sigma}_A + \widehat{\Sigma}_B$, and it can be viewed as an application of Ledoit-Wolf's linear estimator. Based on this, the optimal weights $w^\star$ calculated over experiments $1, \ldots, K$ is

$$w^\star = \frac{1}{K} \sum_{k=1}^{K} \frac{w_k^\star}{\|w_k^\star\|}. \tag{35}$$

We have carried out tests on the performance of metric $w^{\star\top} x$ when the regularization is substituted by a covariance matrix estimator, i.e., $w^\star = k \cdot (f(\widehat{\Sigma}_A) + f(\widehat{\Sigma}_B))^{-1}(\widehat{\mu}_A - \widehat{\mu}_B)$ for different choice of estimator $f(\cdot)$.

Our goal is to test whether the resulting metric $w^{\star\top} x$ can distinguish between A/B groups with significant differences and those without significant differences. Because of the limitations of open-source data, we generate synthetic data in the following ways. We consider A/B test problems with 50 features. In each round, the training data consists of 100 pairs of experiments $B_k \succ A_k, k = 1, \ldots, 100$, and the testing data consists of 200 pairs of experiments, with 60% of them being recognizable A/B pairs and the remaining being A/A pairs without significant differences. All the training control

groups A consist of n multivariate normal samples with mean 0 and covariance Σ_1, and the testing control groups A consist of 200 multivariate normal samples with the same mean and covariance. All the training treatment groups B consist of n multivariate normal samples with mean μ and covariance Σ_2, and the testing treatment groups B consist of 200 multivariate normal samples with mean μ and covariance Σ_2 [2]. The sample size n is set to different values throughout the experiments to test the performance of the metric under different sample size regimes. The optimal coefficient $w^\star$ is learned using the training data $A_k, B_k, k = 1, \ldots, 100$ by (34) and (35), with $(\widehat{\Sigma}_{A_k} + \widehat{\Sigma}_{B_k} + \rho I)^{-1}$ being replaced with $(f(\widehat{\Sigma}_{A_k}) + f(\widehat{\Sigma}_{B_k}))^{-1}$, for different choice of covariance matrix estimator $f(\cdot)$. For each pair of groups A/Y from testing data, where Y can be A or B, let $x_i, y_j, i = 1, ..., n_A, j = 1, ..., n_Y$ be samples from group A and group Y, respectively, and $\widehat{\mu}_A, \widehat{\mu}_Y$ and $\widehat{\Sigma}_A, \widehat{\Sigma}_Y$ be the corresponding sample means and sample covariances. Then we calculate the z-score of A/Y:

$$z(w^\star, A, Y) = \frac{w^{\star\top}(\widehat{\mu}_Y - \widehat{\mu}_A)}{\sqrt{w^{\star\top}\widehat{\Sigma}_A w^\star + w^{\star\top}\widehat{\Sigma}_Y w^\star}}.$$

We use the z-score as a criterion to infer whether $Y = A$ or $Y = B$. Note that a larger z-score indicates that the difference between Y and A is significant, i.e., more likely that we have $Y = B$. The inference task based on the z-score can be viewed as a binary classification problem. Thus, we use the area under the curve (AUC) to justify whether the z-score induced by different covariance estimators performs well. We repeat the above process 20 times for each choice of sample size n, and record the average AUCs in Table 9. It is shown that as the sample size increases, the performance of all estimators becomes better, and the SCOPEs perform the best when $n > p$.

Sample size	$n = 50$	$n = 60$	$n = 70$	$n = 100$	$n = 120$
Sample	0.7620	0.8228	0.8067	0.8494	0.8506
LW-L	**0.7895**	0.8193	0.8072	0.8456	0.8460
LW-NL	0.5002	0.8139	0.8015	0.8430	0.8463
W-SCOPE	0.7620	0.8246	0.8105	**0.8501**	0.8502
KL-SCOPE	0.7729	0.8225	0.8083	0.8459	0.8503
SS-SCOPE	0.7675	**0.8256**	**0.8108**	0.8494	**0.8514**
W-DRCOE	0.7137	0.8143	0.7995	0.8372	0.8388

Table 9: Average AUC of z-score classifier induced by different covariance estimators. The results are obtained from 20 independent repetitions, and the AUCs are given in the format of mean(variance).

5.5 Application in Minimum Variance Portfolio Selection

We consider the minimum variance portfolio selection problem Jagannathan and Ma (2003):

$$\min_{w \in \mathbb{R}^p_+} w^\top \Sigma_0 w, \quad \text{s.t. } w^\top \mathbb{1} = 1, \tag{36}$$

where $w \in \mathbb{R}^p_+$ are the weights allocated to p different risky assets, $\mathbb{1}$ denotes the vector of all ones, Σ_0 is the covariance matrix of the random returns of assets, and the objective $w^\top \Sigma_0 w$ represents the

[2] We choose the parameters by $\mu = \xi, \Sigma_1 = \sum_{i=1}^{100} \xi_i \xi_i^\top, \Sigma_2 = \sum_{i=101}^{200} \xi_i \xi_i^\top$, where $\xi, \xi_1, \ldots, \xi_{400}$ are all drawn from multivariate normal distribution.

variance of the portfolio return. Here, we do not allow short selling by setting $w \geq 0$. The solution of (36) corresponds to the portfolio that minimizes the risk. In practice, the true covariance Σ_0 is unknown and can only be estimated from historical return data.

In this section, we compare the performance of various covariance matrix estimators in the minimum-variance portfolio allocation application (36). We use the "48 industry portfolios" dataset from the Fama-French online library[3]. The dataset contains the monthly returns of 48 portfolio groups by industry, e.g., agriculture, machinery, or communication. We use data from January 1986 to December 1995 for parameter tuning and data from January 1996 to August 2025 for testing. For both the tuning step and testing step, we adopt the following rolling horizon procedure: First, we estimate the covariance matrix from the historical returns within a rolling estimation window of 60 months (5 years), and construct the minimum variance portfolio. We then compute the return of the portfolio in the next month. In the month after that, we shift the rolling window forward by one month, re-estimate the covariance matrix from the historical returns of the past 60 months again, and rebalance the portfolio according to the newest estimation of the covariance matrix. The re-estimation and rebalancing are conducted every month. In the radius tuning step, we search over $[10^{-3}, 5 \times 10^3]$ and choose the radius that yields the highest average portfolio return. In the testing step, we record the monthly returns of the portfolio induced by every covariance matrix estimator. Figure 4 displays the cumulative returns of all the portfolios.

To further compare the performance of the different estimators, Table 10 reports the average returns, Sharpe ratios, Sortino ratios, and cumulative returns of the portfolios induced by the estimators. It is shown that the SCOPEs achieve the highest average return and cumulative return, and maintain the Sharpe ratio and Sortino ratio comparable to or better than the other estimators. This result indicates that the proposed shrinkage estimators, through suitable spectral bias correction, effectively balance the risk and the return of the market, and thus produce portfolios with higher quality.

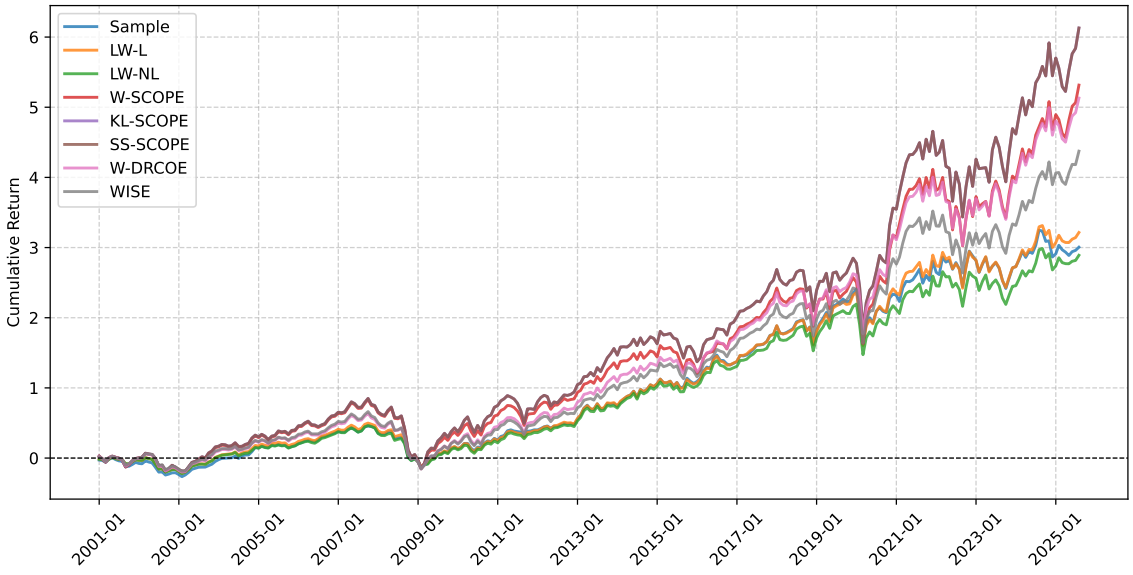


Figure 4: Cumulative returns of the portfolios induced by different estimators.

[3] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library.html

Estimator	Average Return (%)	Sharpe Ratio	Sortino Ratio	Cumulative Return (%)
Sample	5.79	0.5314	0.6593	300.55
LW-L	6.01	0.5491	0.6768	321.48
LW-NL	5.66	0.5239	0.6463	289.00
W-DRCOE	7.63	**0.5862**	**0.7596**	513.02
WISE	7.05	0.5615	0.7101	437.38
KL-SCOPE	**8.29**	0.5597	0.7532	612.88
WA-SCOPE	7.76	0.5477	0.7158	531.54
SS-SCOPE	**8.29**	0.5597	0.7533	**612.91**

Table 10: Performance summary of the portfolios induced by different estimators.

6 Concluding Remarks

In this paper, we propose SCOPE, a novel distributionally robust framework for the simultaneous estimation of covariance and precision matrices. By jointly minimizing the worst-case Frobenius loss and Stein's loss over a divergence-based ambiguity set, we derive a convex optimization formulation that yields a quasi-analytical, nonlinear shrinkage estimator. This estimator corrects spectral bias on both ends, thereby enhancing the condition number and numerical stability of the estimated matrices.

The shrinkage target and intensity are governed by two interpretable parameters, τ and ρ. We introduce a new parameter-tuning approach inspired by inverse optimization and demonstrate that the asymptotically optimal radius scales as $O(n^{-2})$. Extensive numerical experiments support our theoretical findings and show that SCOPE achieves competitive or superior empirical performance compared to state-of-the-art methods in real-world applications.

An interesting direction for future work is the development of a data-driven calibration scheme for the shrinkage radius. Although the asymptotic convergence rate $\rho_n \sim c/n^2$ serves as a useful theoretical guideline, it cannot be directly applied in empirical settings for two main reasons. First, the true distribution of the random vector ξ is generally unknown, making it impossible to derive closed-form expressions for the constants in equations (29)-(31). Second, in finite-sample regimes, the optimal radius may deviate from its asymptotic counterpart, rendering purely asymptotic tuning potentially suboptimal. Nevertheless, the asymptotic form $\rho_n \sim c/n^2$ offers a valuable starting point. Since the constant c is invariant across sample sizes, it may be estimated via regression models, which could then guide the tuning of ρ_n. Developing such a data-driven calibration method entails significant theoretical and empirical work, which we leave for future research.

References

Abbott, S. (2015). *Understanding analysis*. Springer.

Becquin, G. and S. Esmeir (2023). Semantic similarity covariance matrix shrinkage. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9977–9992.

Blanchet, J. and N. Si (2019). Optimal uncertainty size in distributionally robust inverse covariance estimation. *Operations Research Letters 47*(6), 618–621.

Bodnar, T., A. K. Gupta, and N. Parolya (2016). Direct shrinkage estimation of large dimensional precision matrix. *Journal of Multivariate Analysis 146*, 223–236. Special Issue on Statistical Models and Methods for High or Infinite Dimensional Spaces.

Boyd, S. (2004). Convex optimization. *Cambridge UP*.

Chan, T. C., R. Mahmood, and I. Y. Zhu (2025). Inverse optimization: Theory and applications. *Operations Research 73*(2), 1046–1074.

Delage, E. and Y. Ye (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research 58*(3), 595–612.

Dontchev, A. L. and R. T. Rockafellar (2009). *Implicit functions and solution mappings*, Volume 543. Springer.

Gao, R. and A. Kleywegt (2023). Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research 48*(2), 603–655.

Gupta, A. K. and D. K. Nagar (2018). *Matrix variate distributions*. Chapman and Hall/CRC.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The journal of finance 58*(4), 1651–1683.

James, W., C. Stein, et al. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 361–379. University of California Press.

Kharitonov, E., A. Drutsa, and P. Serdyukov (2017). Learning sensitive combinations of a/b test metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 651–659.

Ledoit, O. and M. Wolf (2003a). Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper* (691).

Ledoit, O. and M. Wolf (2003b). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10*(5), 603–621.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics 40*(2), 1024 – 1060.

Ledoit, O. and M. Wolf (2020a). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics 48*(5), 3043 – 3065.

Ledoit, O. and M. Wolf (2020b, 06). The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Journal of Financial Econometrics 20*(1), 187–218.

Ledoit, O. and M. Wolf (2022). Quadratic shrinkage for large covariance matrices. *Bernoulli 28*(3), 1519–1547.

Lin, S., M. Zhang, X. Cheng, K. Zhou, S. Zhao, and H. Wang (2022). Hyperspectral anomaly detection via sparse representation and collaborative representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16*, 946–961.

Liu, W. and Y. Liu (2025). Covariance matrix estimation for positively correlated assets. *arXiv preprint arXiv:2507.01545*.

Marčenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik 1*(4), 457.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

Michaud, R. O. (1989). The markowitz optimization enigma: Is 'optimized'optimal? *Financial analysts journal 45*(1), 31–42.

Mohajerin Esfahani, P. and D. Kuhn (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming 171*(1), 115–166.

Nguyen, V. A., D. Kuhn, and P. Mohajerin Esfahani (2022). Distributionally robust inverse covariance estimation: The wasserstein shrinkage estimator. *Operations research 70*(1), 490–515.

Ravikumar, P., M. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Reed, I. S. and X. Yu (1990). Adaptive multiple-band cfar detection of an optical pattern with unknown spectral distribution. *IEEE transactions on acoustics, speech, and signal processing 38*(10), 1760–1770.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton: Princeton University Press.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management science 9*(2), 277–293.

Stein, C. (1975). Estimation of a covariance matrix. In *39th Annual Meeting IMS, Atlanta, GA, 1975*.

Stein, C. (1986). Lectures on the theory of estimation of many parameters. *Journal of Soviet Mathematics 34*, 1373–1403.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Volume 48. Cambridge university press.

Wang, M., N. Mehr, A. Gaidon, and M. Schwager (2020). Game-theoretic planning for risk-aware interactive agents. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6998–7005. IEEE.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam (2012, 12). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology 75*(3), 427–450.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika 94*(1), 19–35.

Yue, M.-C., Y. Rychener, D. Kuhn, and V. A. Nguyen (2024). A geometric unification of distributionally robust covariance estimators: Shrinking the spectrum by inflating the ambiguity set. *arXiv preprint arXiv:2405.20124*.

Zhao, H. and Z. Duan (2019). Cancer genetic network inference using gaussian graphical models. *Bioinformatics and Biology Insights 13*, 1–9.

Appendix

A Side Results

Lemma 1 (Unbounded problem under singularity)**.** *If $\widehat{\Sigma}$ is singular, then the optimal value*

$$\begin{array}{ll}\min & -\log\det X + \langle X, \widehat{\Sigma}\rangle + \frac{\tau}{2}\left(\|\Sigma\|_F^2 - 2\langle \Sigma, \widehat{\Sigma}\rangle\right)\\ \text{s.t.} & \Sigma \in \mathbb{S}^p_{++},\ X \in \mathbb{S}^p_{++},\ X\Sigma = I.\end{array}$$

is unbounded below.

Proof of Lemma 1. Let $\widehat{\Sigma} = Q^\top \Lambda Q$ be the spectral decomposition of $\widehat{\Sigma}$ with $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_p = 0$ being eigenvalues of $\widehat{\Sigma}$. Let $\Sigma_\eta = \widehat{\Sigma} + \eta I$ and $X_\eta = \Sigma_\eta^{-1}$ for $\eta > 0$. Then (Σ_η, X_η) is a feasible solution. The objective function at the feasible point (Σ_η, X_η) can be evaluated by

$$\begin{aligned}
&-\log\det(\widehat{\Sigma} + \eta I)^{-1} + \langle (\widehat{\Sigma} + \eta I)^{-1}, \widehat{\Sigma}\rangle + \frac{1}{2}\tau\|\eta I\|_F^2 - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2\\
&= \sum_{i=1}^p \log(\lambda_i + \eta) + \mathrm{Tr}\left(Q^\top \Lambda Q (Q^\top(\Lambda + \eta I)Q)^{-1}\right) + \frac{1}{2}\tau\eta p - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2\\
&= \sum_{i=1}^r \log(\lambda_i + \eta) + \sum_{i=r+1}^p \log\eta + \mathrm{Tr}\left(\Lambda(\Lambda + \eta I)^{-1}\right) + \frac{1}{2}\tau\eta p - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2\\
&= \sum_{i=1}^r \log(\lambda_i + \eta) + \sum_{i=r+1}^p \log\eta + \sum_{i=1}^r \frac{\lambda_i}{\lambda_i + \eta} + \frac{1}{2}\tau\eta p - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2.
\end{aligned}$$

The function value goes to $-\infty$ as $\eta \downarrow 0$. □

B Proofs

B.1 Proof of Theorem 1

We first show that the optimal solution to (P-Mat) exists and is unique. To see the uniqueness, we note that the objective function of (P-Mat) is continuous and strongly convex, and the feasible set is convex, and thus, (P-Mat) admits at most one optimal solution. Now we show the existence. Let

$$\mathcal{L} \triangleq \left\{\Sigma \in \mathbb{S}^p_+ : D(\Sigma, \widehat{\Sigma}) \leq \rho\right\} \cap \left\{\Sigma \in \mathbb{S}^p_+ : \log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2 \geq \log\det\widehat{\Sigma} - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2\right\}.$$

and consider the problem

$$\max_{\Sigma\in\mathcal{L}}\ \log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2. \tag{37}$$

Since $D(\cdot, \widehat{\Sigma})$ is continuous, convex and coercive, the set $\{\Sigma \in \mathbb{S}^p_+ : D(\Sigma, \widehat{\Sigma}) \leq \rho\}$ is nonempty, compact and convex. Moreover, since $\log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2$ is continuous and strongly concave over $\mathbb{S}^p_{++}$, then $\mathcal{L}$ is a convex and compact set, and consequently, (37) admits a unique solution, denoted by $\widetilde{\Sigma}$. Since

$$\log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2 < \log\det\widehat{\Sigma} - \frac{1}{2}\tau\|\widehat{\Sigma}\|_F^2, \forall \Sigma \in \left\{\Sigma \in \mathbb{S}^p_+ : D(\Sigma, \widehat{\Sigma}) \leq \rho\right\} \setminus \mathcal{L},$$

then $\widetilde{\Sigma}$ is also an optimal to (P-Mat) and this solution is unique.

Next, we show that (RO), namely the problem

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

admits a unique optimal solution $(\Sigma^\star, (\Sigma^\star)^{-1})$. Let

$$\begin{aligned} F(\Sigma, X) &\triangleq \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \le \rho} \left\{ f(\Sigma, X, S) = -\log\det X + \langle X, S\rangle + \frac{1}{2}\tau\left(\|\Sigma\|_F^2 - 2\langle \Sigma, S\rangle\right)\right\} \\ &= -\log\det X + \frac{1}{2}\tau\|\Sigma\|_F^2 + \max_{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \le \rho} \langle X, S\rangle - \tau\langle \Sigma, S\rangle. \end{aligned}$$

The function is strictly convex in (Σ, X). Thus it admits at most one minimizer over $\mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$. Let Σ^* be the optimal solution to (P-Mat). In the next Proposition 11, we show that the pair $(\Sigma^\star, (\Sigma^\star)^{-1})$ is the unique optimal solution to $\min_{\Sigma, X \in \mathbb{S}^p_{++}} F(\Sigma, X)$. Since $\Sigma^\star X^\star = I$, we conclude that $(\Sigma^\star, X^\star)$ is also the unique optimal solution to (RO).

Proposition 11 (Saddle point of f and minimizer of F)**.** *If Assumptions 1 and 2 hold, then the tuple $(\Sigma = \Sigma^\star, X = (\Sigma^\star)^{-1}, S = \Sigma^\star)$ is a saddle point of function $f(\Sigma, X, S)$, i.e., for any $(\Sigma, X) \in \mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$, and $S \in \left\{S \in \mathbb{S}^p_+ : D(S, \widehat{\Sigma}) \le \rho\right\}$,*

$$f\left(\Sigma^\star, (\Sigma^\star)^{-1}, S\right) \le f\left(\Sigma^\star, (\Sigma^\star)^{-1}, \Sigma^\star\right) \le f\left(\Sigma, X, \Sigma^\star\right). \tag{38}$$

The saddle point is an optimal solution to $\min_{\Sigma, X \in \mathbb{S}^p_{++}} F(\Sigma, X)$.

Proof of Proposition 11. For fixed $\Sigma = \Sigma^\star \in \mathbb{S}^p_{++}$, a basic calculation of the first-order optimality condition shows that $(\Sigma^\star, (\Sigma^\star)^{-1})$ is the minimizer of $\min_{\Sigma, X \in \mathbb{S}^p_{++}} f(\Sigma, X, \Sigma^\star)$. Thus, the second inequality of (38) holds.

Next, we show the first inequality. To this end, we show that $\Sigma^\star$ is the optimal solution to problem

$$\max_{\Sigma \in \mathbb{S}^p_+ : D(\Sigma, \widehat{\Sigma}) \le \rho} f(\Sigma^\star, (\Sigma^\star)^{-1}, \Sigma), \tag{39}$$

which is equivalent to

$$\max_{\Sigma \in \mathbb{S}^p_+ : D(\Sigma, \widehat{\Sigma}) \le \rho} \langle (\Sigma^\star)^{-1}, \Sigma\rangle - \tau\langle \Sigma^\star, \Sigma\rangle. \tag{40}$$

Note that $D(\widehat{\Sigma}, \widehat{\Sigma}) = 0 < \rho$ by Assumptions 1 and 2(a). By (Boyd, 2004, Section 5.2.3), Slater's condition ensures that strong duality for (P-Mat) and (40) both hold. Thus $\Sigma^\star$ is an optimal solution to problem (40) if and only if $\Sigma = \Sigma^\star$ solves the following KKT system:

$$\begin{aligned} \Sigma^{-1} - \tau\Sigma - \gamma\nabla_1 D(\Sigma, \widehat{\Sigma}) &= \mathbf{0}, \\ \gamma(D(\Sigma, \widehat{\Sigma}) - \rho) &= 0, \\ D(\Sigma, \widehat{\Sigma}) &\le \rho, \\ \gamma \ge 0, \Sigma &\in \mathbb{S}^p_+. \end{aligned} \tag{41}$$

On the other hand, the strong duality of (P-Mat) means that $\Sigma^\star$ satisfies the following KKT system (41). Thus $\Sigma = \Sigma^\star$ solves (39) and the first inequality of (38) holds. □

B.2 Proof of Proposition 1

Proof of Proposition 1. Let $\ell(\Sigma) \triangleq \Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2$. Then $\nabla\ell(\Sigma) = \Sigma^{-1} - \tau\Sigma$ and $\nabla\ell\left(\sqrt{\frac{1}{\tau}}I\right) = 0$. Since $\ell(\Sigma)$ is strongly concave (see e.g. Boyd (2004, Page 74)), then $\Sigma^\star = \sqrt{\frac{1}{\tau}}I$ is the global maximizer of $\ell(\Sigma)$. The conclusion follows as $\Sigma^\star$ lies in the feasible set of (P-Mat). □

B.3 Proof of Proposition 2

Proof of Proposition 2. To ease the notation, we denote $d(\cdot, b)$ by $d_b(\cdot)$ for any $b \geq 0$ in this proof. Let $f(a) \triangleq \frac{1}{a} - \tau a - \gamma d_b'(a)$. Since $d_b(\cdot)$ is convex and differentiable over $\mathbb{R}_{++}$, then $d_b'(\cdot)$ is non-decreasing over $\mathbb{R}_{++}$. Moreover, $b \geq 0$, and $d_b'(b) = 0$ (see Remark 1). Thus $\lim_{a\downarrow 0} d_b'(a) \leq 0$. Together with the fact that $\lim_{a\downarrow 0} \frac{1}{a} = +\infty$, we have that $\lim_{a\downarrow 0} f(a) = +\infty$. Note that f is strictly decreasing over $\mathbb{R}_{++}$ and $\lim_{a\to\infty} f(a) = -\infty$, we conclude that $f(a) = 0$ admits a unique solution over $\mathbb{R}_{++}$. □

B.4 Proof of Proposition 3

The proof is divided into two steps. We first show that (P-Mat) is equivalent to the following problem

$$\begin{array}{ll} \max\limits_{s\in\mathbb{R}_+^p} & \sum\limits_{i=1}^p \log s_i - \dfrac{1}{2}\tau\sum\limits_{i=1}^p s_i^2 \\ \text{s.t.} & \sum\limits_{i=1}^p d(s_i, \widehat{\lambda}_i) \leq \rho \\ & s_1 \leq s_2 \leq \ldots s_p, \end{array} \tag{P-Vec$^\uparrow$}$$

in the sense of Proposition 12. In the second step, we show the equivalence of (P-Vec$^\uparrow$) and (P-Vec) in the sense of Proposition 13. Combining these two propositions readily proves Proposition 3.

Proposition 12 (Equivalence of (P-Mat) and (P-Vec$^\uparrow$))**.** *If Assumptions 1-3 hold, then the following assertions hold.*

(i) Problem (P-Mat) is feasible if and only if problem (P-Vec$^\uparrow$) is feasible.

(ii) The feasible set of (P-Mat) is compact if and only if the feasible set of (P-Vec$^\uparrow$) is compact.

(iii) If $s^\star$ is an optimal solution to (P-Vec$^\uparrow$), then $\widehat{V}\mathrm{diag}(s^\star)\widehat{V}^\top$ is an optimal solution to (P-Mat).

(iv) If $\Sigma^\star$ is an optimal solution to (P-Mat), then $\lambda(\Sigma^\star)$ is an optimal solution to (P-Vec$^\uparrow$).

(v) The optimal values of (P-Mat) and (P-Vec$^\uparrow$) are equal.

Proof of Proposition 12. The result is similar to (Yue et al., 2024, Proposition 9). The main difference is that the objective becomes $\ell(\Sigma) = \log\det\Sigma - \frac{1}{2}\tau\|\Sigma\|_F^2$. Note that ℓ is strongly concave, and the value of ℓ only depends on the eigenvalues of Σ. That means the nature of the problem does not change, and thus, with a slight modification, proof of (Yue et al., 2024, Proposition 9) works here. We omit the details. □

Proposition 13. *Under Assumptions 1-3, (P-Vec) admits a unique optimal solution, denoted by $s^\star$, and $s^\star$ is the unique optimal solution to (P-Vec$^\uparrow$).*

Proof. We proceed with the proof in two steps.

Step 1. We show that problem (P-Vec) admits a unique optimal solution. To this end, we show that the feasible set

$$\mathcal{S} \triangleq \{s \in \mathbb{R}^p_+ : \sum_{i=1}^p d(s_i, \widehat{\lambda}_i) \leq \rho\}$$

is convex and compact and the objective is strictly convex over the set. Convexity of the set is implied by the convexity of d. To show compactness, it suffices to show closeness and boundedness of the set. The former is guaranteed by the continuity of d under Assumption 3 (ii). To show the boundedness, we note that by Remark 1, $d(s_i, \widehat{\lambda}_i) = D(s_i I, \widehat{\lambda}_i I)/p$. Consequently, the feasible set can be written as

$$\mathcal{S} = \{s \in \mathbb{R}^p_+ : \sum_{i=1}^p D(s_i I, \widehat{\lambda}_i I)/p \leq \rho\}.$$

Since $p > 1$, then

$$S \subset \{s_i \in \mathbb{R}_+ : D(s_i I, \widehat{\lambda}_i I) \leq \rho\}.$$

Assumption 2 (iii) ensures that the latter is bounded, which implies that $\mathcal{S}$ is bounded as desired. Since the objective function is continuous and strictly convex over the feasible set, the existence of a unique optimal solution is evident.

Step 2. Let $s^\star$ be the optimal solution to problem (P-Vec). We show that $s^\star = s^{\star\uparrow}$, and thus $s^\star$ is also optimal to (P-Vec$^\uparrow$). Let $q(s) \triangleq \sum_{i=1}^p \log s_i - \tau \sum_{i=1}^p s_i^2$ be the objective function of (P-Vec). Assume for the sake of a contradiction that $s^\star \neq s^{\star\uparrow}$. Then by the feasibility of $s^\star$ and Lemma 2,

$$\sum_{i=1}^p d(s_i^{\star\uparrow}, \widehat{\lambda}_i) < \sum_{i=1}^p d(s_i^\star, \widehat{\lambda}_i) \leq \rho, \tag{42}$$

which means $s^{\star\uparrow}$ is feasible to (P-Vec). Moreover, $q(s^\star) = q(s^{\star\uparrow})$, which means $s^{\star\uparrow}$ is also optimal to (P-Vec). This is a contradiction to the fact that $s^\star$ is the unique optimal solution. Finally, we note that the feasible set of (P-Vec$^\uparrow$) is a subset of that of (P-Vec), and thus $s^\star$ is the unique optimal solution to (P-Vec$^\uparrow$). □

Lemma 2 (Yue et al. (2024, Lemma 5))**.** *If Assumptions 2-3 hold, then*

$$\sum_{i=1}^p d(s_i^\uparrow, y_i^\uparrow) \leq \sum_{i=1}^p d(s_i, y_i^\uparrow) \quad \forall x, y \in \mathbb{R}^p_+.$$

If the right-hand side is finite, then the equality holds if and only if $s = s^\uparrow$.

B.5 Proof of Proposition 4

To ease the notation, we denote $d(\cdot, b)$ by $d_b(\cdot)$ for any $b \geq 0$ in the following proof.

Lemma 3 (Derivative of d_b)**.** *Let Assumptions 2 and 3 hold and $b > 0$ be any fixed positive constant. Then $d_b'(a) < 0$ for $a \in (0, b)$ and $d_b'(a) > 0$ for $a \in (b, \infty)$.*

Proof. Let $b > 0$ be fixed. Since $d(\cdot, b)$ is convex over $\mathbb{R}_+$, then

$$0 = d(b, b) \geq d(a, b) + (b - a) d_b'(a), \quad \forall a \in \mathbb{R}_{++}.$$

Thus, for any $a \in (0, b)$, $d_b'(a) \leq -\frac{d(a,b)}{b-a} < 0$, and for any $a > b$, $d_b'(a) \geq \frac{d(a,b)}{a-b} > 0$. In both cases, we use the fact that $d(a, b) > 0$ for any $a > 0, a \neq b$. □

Proof of Proposition 4. Part (i). Let $f(a) = \frac{1}{a} - \tau a - \gamma d_b'(a)$. We prove by the fact that f is strictly decreasing over $\mathbb{R}_{++}$. If $b < \sqrt{\frac{1}{\tau}}$, we have

$$f(b) = \frac{1}{b} - \tau b > 0, f(1/\sqrt{\tau}) = -\gamma d_b'(1/\sqrt{\tau}) \leq 0$$

and thus $b < \varphi_{\tau,b}(\gamma) \leq \sqrt{\frac{1}{\tau}}$. If $b > \sqrt{\frac{1}{\tau}}$, then $f(b) < 0, f(1/\sqrt{\tau}) \geq 0$ and thus

$$\sqrt{\frac{1}{\tau}} \leq \varphi_{\tau,b}(\gamma) < b.$$

If $b = 1/\sqrt{\tau}$, then $f(1/\sqrt{\tau}) = 0$ and

$$\varphi_{\tau,b}(\gamma) = 1/\sqrt{\tau}.$$

Part (ii). Let $H(\gamma, a) = \frac{1}{a} - \tau a - \gamma d_b'(a)$. By Assumption 3(ii), $d_b(a)$ is convex and $d_b'(a)$ is continuously differentiable on $\mathbb{R}_{++}$. Thus

$$\frac{\partial H(\gamma, a)}{\partial a} = -\frac{1}{a^2} - \tau - \gamma d_b''(a) < 0, \forall a \in \mathbb{R}_{++}.$$

Since $\varphi_{\tau,b}(\gamma) > 0$ by Part (i), we can use classical implicit function theorem (see e.g. (Dontchev and Rockafellar, 2009, Section 1.2)) to assert that $\varphi_{\tau,b}$ is differentiable (and thus continuous) over $\mathbb{R}_+$ and

$$\varphi_{\tau,b}'(\gamma) = -\frac{d_b'(\varphi_{\tau,b}(\gamma))}{\frac{1}{a^2} + \tau + \gamma d_b''(\varphi_{\tau,b}(\gamma))}.$$

The sign of $\varphi_{\tau,b}'(\gamma)$ depends on the sign of $-d_b'(a)$. By (i) and Lemma 3, $\varphi_{\tau,b}(\gamma) > b$ if $b < 1/\sqrt{\tau}$ and thus $-d_b'(a) < 0$ and $\varphi_{\tau,b}(\gamma) < b$ if $b > 1/\sqrt{\tau}$ and thus $-d_b'(a) > 0$.

Part (iii). Since $\varphi_{\tau,b}(\gamma)$ is continuous at 0, then

$$\lim_{\gamma \downarrow 0} \varphi_{\tau,b}(\gamma) = \varphi_{\tau,b}(0) = 1/\sqrt{\tau}.$$

If $b = 1/\sqrt{\tau}$, then $\varphi_{\tau,b}(\gamma) = 1/\sqrt{\tau}$ for any $\gamma \geq 0$ and thus $\lim_{\gamma\to\infty} = b$. Now we assume $b \neq 1/\sqrt{\tau}$. Note that $\varphi_{\tau,b}(\gamma)$ is strictly increasing and bounded above if $b > 1/\sqrt{\tau}$ and strictly decreasing and bounded below if $b < 1/\sqrt{\tau}$. So in both cases, the limit is well defined. By definition, we have

$$d_b'(\varphi_{\tau,b}(\gamma)) = \frac{1}{\gamma}\left(\frac{1}{a} - \tau a\right)$$

for $\gamma > 0$. Taking limits on both sides and exploiting the fact that d_b' is continuous gives

$$0 = \lim_{\gamma\to\infty} d_b'\left(\varphi_{\tau,b}(\gamma)\right) = d_b'\left(\lim_{\gamma\to\infty} \varphi_{\tau,b}(\gamma)\right).$$

By Lemma 3 and the fact that $d_b(a)$ is minimized by $a = b$, we prove that $\lim_{\gamma\to\infty} \varphi_{\tau,b}(\gamma) = b$.

Proof of Part (iv). Let $a' = \phi(b) = \frac{\varphi(\tau,\gamma,b)}{b}$. Then a' is the unique solution to the equation

$$\frac{1}{a'b} - \tau a' b - \gamma d_b'(a'b) = 0.$$

By treating $\phi(b)$ as the solution to the equation above, we can use the classical implicit function theorem again to derive

$$\phi'(b) = -\frac{-\frac{1}{a'b^2} - \tau a' - \gamma a' d_b''(a'b)}{-\frac{1}{a'^2 b} - \tau b - \gamma b d_b''(a'b)} \leq 0,$$

which completes the proof. □

B.6 Proof of Proposition 5

Proof of Proposition 5. The uniqueness is established at Step 1 in the proof of Proposition 13. So we are left to show that the unique optimal solution can be represented by $s_i^\star = \varphi(\tau, \gamma^\star, \widehat{\lambda}_i), i = 1, \ldots, p$, where $\gamma^\star$ is the unique solution of the nonlinear equation $\sum_{i=1}^p d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \rho = 0$. Define the Lagrangian function of problem (P-Vec)

$$L(\gamma, s) = \sum_{i=1}^p \log s_i - \frac{1}{2}\tau \sum_{i=1}^p s_i^2 - \gamma \left(\sum_{i=1}^p d(s_i, \widehat{\lambda}_i) - \rho \right).$$

By Rockafellar (1970, Theorem 28.3), (P-Vec) can be represented as

$$\max_{s \in \mathbb{R}_+^p} \min_{\gamma \geq 0} L(\gamma, s) = \min_{\gamma \geq 0} \max_{s \in \mathbb{R}_+^p} L(\gamma, s), \tag{43}$$

where the left-hand side is the primal problem and the right-hand side is the Lagrange dual problem. Equation (43) means that strong duality holds. Let $\gamma^\star$ be an optimal solution of the dual problem. Then

$$\max_{s \in \mathbb{R}_+^p} L(\gamma^\star, s) = \gamma^\star \rho + \sum_{i=1}^p \max_{s_i > 0} \left\{ \log s_i - \frac{1}{2}\tau s_i^2 - \gamma^\star d(s_i, \widehat{\lambda}_i) \right\}.$$

The problem above is decomposable, which means that it can be solved by solving the following univariate maximization problem

$$\max_{s_i > 0} \ \log s_i - \frac{1}{2}\tau s_i^2 - \gamma^\star d(s_i, \widehat{\lambda}_i), \tag{44}$$

for $i = 1, \ldots, p$. The first-order optimality condition of the problem above

$$\frac{1}{s_i} - \tau s_i - \gamma^\star \frac{\partial d}{\partial s_i}(s_i, \widehat{\lambda}_i) = 0. \tag{45}$$

By Proposition 2, (45) always admits a unique strictly positive solution for any fixed $\tau > 0, \gamma^\star \geq 0$ and $\widehat{\lambda}_i \geq 0$. Exploiting the notation defined in (10), the optimal solution to (44) is $s_i^\star = \varphi(\tau, \gamma^\star, \widehat{\lambda}_i)$.

Next, we prove that $\sum_{i=1}^p d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) - \rho = 0$. If $\rho = \rho_{\max}$, then $\gamma^\star = 0$ solves the equation since $\varphi(\tau, 0, \widehat{\lambda}_i) = \sqrt{\frac{1}{\tau}}$ and thus $\sum_{i=1}^p d(\varphi(\tau, \gamma^\star, \widehat{\lambda}_i), \widehat{\lambda}_i) = \rho_{\max}$ by definition of $\rho_{\max}$. Let $\rho \in (0, \rho_{\max})$. It suffices to prove $\gamma^\star > 0$ since the complementarity condition in the KKT conditions will ensure the equality and the feasibility. Assume for the sake of a contradiction that $\gamma^\star = 0$. Then by (45), $\left(0, \left(\sqrt{\frac{1}{\tau}}, \ldots, \sqrt{\frac{1}{\tau}}\right)^\top\right)$ is a saddle point of (43), which contradicts the assumption that $\sum_{i=1}^p d\left(\sqrt{\frac{1}{\tau}}, \widehat{\lambda}_i\right) = \rho_{\max} > \rho$. This shows that $\gamma^\star > 0$ as desired. Moreover, Lemma 4 below ensures the uniqueness of $\gamma^\star$, which completes the proof. □

Lemma 4 (Strictly decreasing $F_{\widehat{\lambda}}$)**.** *Let the function $F_{\widehat{\lambda}} : \mathbb{R}_+ \to \mathbb{R}$ be defined by*

$$F_{\widehat{\lambda}}(\gamma) = \sum_{i=1}^p d(\varphi(\tau, \gamma, \widehat{\lambda}_i), \widehat{\lambda}_i). \tag{46}$$

If Assumptions 1-4 hold, then the function $F_{\widehat{\lambda}}$ is strictly decreasing over $\mathbb{R}_+$. If Assumption 5 holds in addition, then

$$F_{\widehat{\lambda}}(0) = \rho_{\max} \geq \rho \quad \text{and} \quad \lim_{\gamma \to \infty} F_{\widehat{\lambda}}(\gamma) = 0 \leq \rho. \tag{47}$$

Proof of Lemma 4. We first show that $F_{\widehat{\lambda}}$ is strictly decreasing. If $\widehat{\lambda}_i = 1/\sqrt{\tau}$, then by Proposition 4 (i), $d(\varphi(\tau,\gamma,\widehat{\lambda}_i),\widehat{\lambda}_i) \equiv 0$. Note that by Assumption 4, there always exists $\widehat{\lambda}_i$ such that $\widehat{\lambda}_i \neq 1/\sqrt{\tau}$. So it suffices to prove that $d(\varphi(\tau,\gamma,\widehat{\lambda}_i),\widehat{\lambda}_i)$ is strictly decreasing on γ if $\widehat{\lambda}_i \neq 1/\sqrt{\tau}$. If $\widehat{\lambda}_i > 1/\sqrt{\tau}$, then for $\gamma_1 > \gamma_2$, we have

$$\varphi(\tau,\gamma_2,\widehat{\lambda}_i) < \varphi(\tau,\gamma_1,\widehat{\lambda}_i) < \widehat{\lambda}_i$$

by Proposition 4 (i) and (ii). Then by Lemma 5 below, we have

$$d(\varphi(\tau,\gamma_2,\widehat{\lambda}_i),\widehat{\lambda}_i) > d(\varphi(\tau,\gamma_1,\widehat{\lambda}_i),\widehat{\lambda}_i),$$

which proves that $d(\varphi(\tau,\gamma,\widehat{\lambda}_i),\widehat{\lambda}_i)$ is strictly decreasing on γ if $\widehat{\lambda}_i > 1/\sqrt{\tau}$. If $\widehat{\lambda}_i < 1/\sqrt{\tau}$, then for $\gamma_1 > \gamma_2$, similarly by Proposition 4 (i) and (ii) we have

$$\varphi(\tau,\gamma_2,\widehat{\lambda}_i) > \varphi(\tau,\gamma_1,\widehat{\lambda}_i) > \widehat{\lambda}_i.$$

By Lemma 5, we have

$$d(\varphi(\tau,\gamma_2,\widehat{\lambda}_i),\widehat{\lambda}_i) > d(\varphi(\tau,\gamma_1,\widehat{\lambda}_i),\widehat{\lambda}_i).$$

Thus, we derive that F is strictly decreasing over $\mathbb{R}_+$.

Finally, by Proposition 4(iii) and Assumption 5, we have $F_{\widehat{\lambda}}(0) = \sum_{i=1}^p d(\sqrt{\frac{1}{\tau}},\widehat{\lambda}_i) = \rho_{\max} \geq \rho$ and $\lim_{\gamma\to\infty} F_{\widehat{\lambda}}(\gamma) = \sum_{i=1}^p d(\widehat{\lambda}_i,\widehat{\lambda}_i) = 0 < \rho$ and thus completes the proof. □

Lemma 5 (Monotonicity of d)**.** *If Assumptions 2 and 3 hold, then we have*

(i) if $a_1 < a_2 < b$, then $d(a_1,b) > d(a_2,b)$,

(ii) if $a_1 > a_2 > b$, then $d(a_1,b) > d(a_2,b)$.

Proof of Lemma 5. By convexity of d, for $a_1 < a_2 < b$ we have

$$d(a_1,b) \geq d(a_2,b) + (a_1 - a_2)d_b'(a_2) > d(a_2,b),$$

which proves assertion (i). Assertion (ii) can be proved similarly. □

B.7 Proof of Proposition 6

Proof of Proposition 6. In view of Lemma 4, it only remains to show that $F_{\widehat{\lambda}}$ is differentiable over $\mathbb{R}_+$. Note that $d_b(\cdot)$ is differentiable by Assumption 3(ii) and $\varphi_{\tau,b}(\cdot)$ is differentiable by Proposition 4(ii). Then by the chain rule, we conclude that $F_{\widehat{\lambda}}$ is differentiable over $\mathbb{R}_+$. □

B.8 Proof of Theorem 3

Proof of Theorem 3. We show that $\Sigma^\star(\tau,\rho)$ is continuous in ρ for each fixed τ. To facilitate reading, we recall that

$$\Sigma^\star(\tau,\rho) = \widehat{V}\Phi(\tau,\gamma^\star_{\widehat{\lambda}}(\rho),\widehat{\lambda})\widehat{V}^\top \triangleq \widehat{V}\mathrm{diag}(\varphi(\tau,\gamma^\star_{\widehat{\lambda}}(\rho),\widehat{\lambda}_1),\ldots,\varphi(\tau,\gamma^\star_{\widehat{\lambda}}(\rho),\widehat{\lambda}_p))\widehat{V}^\top, \rho \in (0,\rho_{\max}]. \quad (48)$$

By definition in (12), $\gamma^\star_{\widehat{\lambda}}(\rho)$ is continuous in ρ. On the other hand, it follows by Proposition 4 (ii) that $\varphi(\tau,\gamma,b)$ is continuous in γ. The continuity of $\Sigma^\star$ follows directly from (48). Recall that $\lim_{\rho\downarrow 0}\gamma^\star_{\widehat{\lambda}}(\rho) = \infty$ and by Proposition 4(iii), $\lim_{\gamma\to\infty}\varphi(\tau,\gamma,\widehat{\lambda}_i) = \widehat{\lambda}_i$. Then

$$\lim_{\rho\downarrow 0}\Sigma^\star(\tau,\rho) = \lim_{\gamma\to\infty}\widehat{V}\mathrm{diag}(\varphi(\tau,\gamma,\widehat{\lambda}_1),\ldots,\varphi(\tau,\gamma,\widehat{\lambda}_p))\widehat{V}^\top = \widehat{V}\mathrm{diag}(\widehat{\lambda}_1,\ldots,\widehat{\lambda}_p)\widehat{V}^\top = \widehat{\Sigma}.$$

Likewise, since $\gamma^\star_{\widehat{\lambda}}(\rho_{\max}) = 0$, and $\varphi(\tau, 0, \widehat{\lambda}_i) = \sqrt{1/\tau}$, then

$$\Sigma^\star(\tau, \rho_{\max}) = \widehat{V}\mathrm{diag}(\varphi(\tau, 0, \widehat{\lambda}_1), \ldots, \varphi(\tau, 0, \widehat{\lambda}_p))\widehat{V}^\top = \widehat{V}\mathrm{diag}(\sqrt{1/\tau}, \ldots, \sqrt{1/\tau})\widehat{V}^\top = \sqrt{\frac{1}{\tau}}I.$$

The assertions (i)-(iii) of the theorem follow directly from Proposition 4 (i)-(ii). □

B.9 Proof of Proposition 8

Proof of Proposition 8. The conclusion on the eigenvalue mappings follows directly from the definition of φ (see equation (10)) by substituting the specific forms of generators d into the equation. We omit the details of the calculations.

In what follows, we derive the upper bound, denoted by $\overline{\gamma}$, of the dual variables $\gamma^\star$. By Proposition 6, $\overline{\gamma} \geq \gamma^\star$ if and only if $F_{\widehat{\lambda}}(\overline{\gamma}) \leq \rho$, i.e.,

$$\sum_{i=1}^{p} d(\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i), \widehat{\lambda}_i) \leq \rho. \tag{49}$$

It suffices to choose $\overline{\gamma}$ such that

$$d(\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i), \widehat{\lambda}_i) \leq \frac{\rho}{p}, \quad \text{for } i = 1, ..., p. \tag{50}$$

To ease the exposition, let $a^\star \triangleq \varphi(\tau, \gamma, b)$. We also recall that by Theorem 3 (i) and (ii)

$$\widehat{\lambda}_1 \leq \varphi(\tau, \gamma, \widehat{\lambda}_i) \leq \widehat{\lambda}_p \ \ \forall \gamma > 0, \tag{51}$$

for $i = 1, ..., p$.

Upper bound of Kullback-Leibler divergence. Substituting $d(a, b) = \frac{1}{2}\left(\frac{a}{b} - 1 - \log\frac{a}{b}\right)$ into (50), we obtain

$$\frac{1}{2}\left(\frac{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)}{\widehat{\lambda}_i} - 1 - \log\frac{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)}{\widehat{\lambda}_i}\right) \leq \frac{\rho}{p}, \quad \text{for } i = 1, ..., p.$$

The inequalities above are guaranteed by

$$\frac{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)}{\widehat{\lambda}_i} - 1 \leq \frac{\rho}{p} \text{ and } -\log\frac{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)}{\widehat{\lambda}_i} \leq \frac{\rho}{p}, \quad \text{for } i = 1, ..., p. \tag{52}$$

By Proposition 4(iv), $\phi(b) \triangleq \frac{\varphi(\tau,\gamma,b)}{b}$ is non-increasing in b. Moreover, by Proposition 4(ii) and the assumption that $\widehat{\lambda}_1 < \sqrt{\frac{1}{\tau}} < \widehat{\lambda}_p$, we have $\frac{\varphi(\tau,\overline{\gamma},\widehat{\lambda}_1)}{\widehat{\lambda}_1} > 1$ and $\frac{\varphi(\tau,\overline{\gamma},\widehat{\lambda}_p)}{\widehat{\lambda}_p} < 1$. Inequalities (52) are ensured by

$$\frac{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_1)}{\widehat{\lambda}_1} - 1 \leq \frac{\rho}{p} \text{ and } \log\frac{\widehat{\lambda}_p}{\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_p)} \leq \frac{\rho}{p}. \tag{53}$$

Note that

$$\varphi(\tau, \overline{\gamma}, b) = \frac{2(2 + \overline{\gamma})b}{\left(\overline{\gamma} + \sqrt{\overline{\gamma}^2 + 8\tau(2 + \overline{\gamma})b^2}\right)} \leq \frac{(2 + \overline{\gamma})b}{\overline{\gamma}}$$

for $b > 0$. By taking $\overline{\gamma} \geq \frac{2p}{\rho}$, we have

$$\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_1) \leq \frac{(2+\overline{\gamma})\widehat{\lambda}_1}{\overline{\gamma}} \leq (\frac{\rho}{p} + 1)\widehat{\lambda}_1,$$

which gives rise to the first inequality of (53). On the other hand, since

$$\varphi(\tau, \overline{\gamma}, b) = \frac{2(2+\overline{\gamma})b}{\left(\overline{\gamma} + \sqrt{\overline{\gamma}^2 + 8\tau(2+\overline{\gamma})b^2}\right)} \geq \frac{2(2+\overline{\gamma})b}{\overline{\gamma} + \overline{\gamma} + 4\tau b^2 + 2} \geq \frac{\overline{\gamma}b}{,\overline{\gamma} + 2\tau b^2 + 1}$$

then

$$\overline{\gamma} \geq \frac{2\tau\widehat{\lambda}_p^2 + 1}{e^{\rho/p} - 1} \;\Rightarrow\; \frac{\overline{\gamma}\widehat{\lambda}_p}{\overline{\gamma} + 2\tau\widehat{\lambda}_p^2 + 1} \geq \widehat{\lambda}_p e^{-\frac{\rho}{p}}$$

which gives rise to the second inequality of (53). Combining the above results, we obtain an upper bound of $\gamma^\star$:

$$\overline{\gamma} = \max\left\{\frac{2p}{\rho}, \frac{2\tau\widehat{\lambda}_p^2 + 1}{e^{\rho/p} - 1}\right\}.$$

Upper bound of Wasserstein divergence. The generator is $d(a, b) = a + b - 2\sqrt{ab}$, and the eigenvalue mapping is the unique positive solution $a^\star$ of

$$\frac{1}{a^\star} - \tau a^\star - \gamma\left(1 - \frac{\sqrt{b}}{\sqrt{a^\star}}\right) = 0.$$

Reformulating the equation gives

$$d(a^\star, b) = \left(\sqrt{a^\star} - \sqrt{b}\right)^2 = \frac{(1 - \tau a^{\star 2})^2}{\gamma^2 a^\star}.$$

Then (50) becomes

$$\frac{(1 - \tau\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)^2)^2}{\overline{\gamma}^2\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)} \leq \frac{\rho}{p} \iff \overline{\gamma}^2 \geq \frac{p(1 - \tau\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)^2)^2}{\rho\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)}, \quad \text{for } i = 1, ..., p.$$

Note that

$$\frac{p(1 - \tau\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)^2)^2}{\rho\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)} \leq \frac{p(1 - \tau\varphi(\tau, \overline{\gamma}, \widehat{\lambda}_i)^2)^2}{\rho\varphi(\tau, \overline{\gamma}, 0)} \leq \max\left\{\frac{p}{\rho\varphi(\tau, \overline{\gamma}, 0)}, \frac{p(1 - \tau\widehat{\lambda}_p^2)^2}{\rho\varphi(\tau, \overline{\gamma}, 0)}\right\},$$

where the last inequality is by Lemma 6 and

$$\varphi(\tau, \overline{\gamma}, 0) = \frac{-\overline{\gamma} + \sqrt{\overline{\gamma}^2 + 4\tau}}{2\tau}.$$

Then it suffices to have

$$\max\left\{\frac{p}{\rho\varphi(\tau, \overline{\gamma}, 0)}, \frac{p(1 - \tau\widehat{\lambda}_p^2)^2}{\rho\varphi(\tau, \overline{\gamma}, 0)}\right\} \leq \overline{\gamma}^2.$$

Solving the above inequality, we conclude that

$$\overline{\gamma} = \max\left\{\frac{\eta_1 + \sqrt{\eta_1^2 + 16\eta_1\tau^{2.5}}}{8\tau^2}, \frac{\eta_2 + \sqrt{\eta_2^2 + 16\eta_2\tau^{2.5}}}{8\tau^2}\right\}$$

satisfies the inequality, where

$$\eta_1 = \frac{p}{\rho} \quad \text{and} \quad \eta_2 = \frac{p(1-\tau\widehat{\lambda}_p^2)^2}{\rho}.$$

Upper bound of symmetrized Stein divergence. The generator is $d(a,b) = \frac{1}{2}\left(\frac{b}{a} + \frac{a}{b} - 2\right)$, and the eigenvalue mapping is the unique positive solution $a^\star$ of

$$\frac{1}{a^\star} - \tau a^\star - \frac{\gamma}{2}\left(-\frac{b}{a^{\star 2}} + \frac{1}{b}\right) = 0.$$

Reordering the terms of the above equation gives

$$d(a^\star, b) = \frac{(a^\star - b)^2}{2a^\star b} = \frac{2a^\star b(1-\tau a^{\star 2})^2}{\gamma^2(a^\star + b)^2}.$$

Then (50) becomes

$$\frac{2\widehat{\lambda}_i\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{\overline{\gamma}^2(\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)+\widehat{\lambda}_i)^2} \le \frac{\rho}{p} \;\Leftrightarrow\; \overline{\gamma}^2 \ge \frac{2p\widehat{\lambda}_i\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{\rho(\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)+\widehat{\lambda}_i)^2}, \forall i = 1,\ldots,p.$$

Note that

$$\frac{2p\widehat{\lambda}_i\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{\rho(\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)+\widehat{\lambda}_i)^2} \le \frac{p\widehat{\lambda}_p^2(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{2\rho\widehat{\lambda}_1^4} \le \max\left\{\frac{p\widehat{\lambda}_p^2}{2\rho\widehat{\lambda}_1^4}, \frac{2p\widehat{\lambda}_p^2(1-\tau\widehat{\lambda}_p^2)^2}{4\rho\widehat{\lambda}_1^4}\right\},$$

where the last inequality is by Lemma 6. So we set

$$\overline{\gamma} \triangleq \sqrt{\max\left\{\frac{p\widehat{\lambda}_p^2}{2\rho\widehat{\lambda}_1^4}, \frac{p\widehat{\lambda}_p^2(1-\tau\widehat{\lambda}_p^2)^2}{2\rho\widehat{\lambda}_1^4}\right\}}.$$

Upper bound of squared Frobenius divergence. The generator is $d(a,b) = (a-b)^2$, and the eigenvalue mapping is the unique positive solution $a^\star$ of

$$\frac{1}{a^\star} - \tau a^\star - 2\gamma(a^\star - b) = 0.$$

Reordering the terms of the above equation gives

$$d(a^\star, b) = (a^\star - b)^2 = \frac{(1-\tau a^{\star 2})^2}{4\gamma^2 a^{\star 2}}.$$

Consequently, (50) reduces to

$$\frac{(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\overline{\gamma}^2\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2} \le \frac{\rho}{p} \iff \overline{\gamma}^2 \ge \frac{p(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2}, \quad \text{for } i = 1,...,p.$$

Note that

$$\frac{p(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2} \le \frac{p(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\varphi(\tau,\overline{\gamma},0)^2} \le \max\left\{\frac{p}{4\rho\varphi(\tau,\overline{\gamma},0)^2}, \frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho\varphi(\tau,\overline{\gamma},0)^2}\right\},$$

where the last inequality is by Lemma 6 and

$$\varphi(\tau,\overline{\gamma},0)=\sqrt{\frac{1}{\tau+2\overline{\gamma}}}.$$

Let

$$\max\left\{\frac{p}{4\rho}(\tau+2\overline{\gamma}),\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho}(\tau+2\overline{\gamma})\right\}\leq\overline{\gamma}^2.$$

We get

$$\overline{\gamma}\geq\max\left\{\frac{p}{4\rho}+\sqrt{\frac{p^2}{16\rho^2}+\tau\frac{p}{4\rho}},\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho}+\sqrt{\frac{p^2(1-\tau\widehat{\lambda}_p^2)^4}{16\rho^2}+\tau\frac{p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho}}\right\}.$$

Upper bound of weighted Frobenius divergence. The generator is $d(a,b)=\frac{(a-b)^2}{b}$, and the eigenvalue mapping is the unique positive solution $a^\star$ of

$$\frac{1}{a^\star}-\tau a^\star-\frac{2\gamma}{b}(a^\star-b)=0.$$

Reordering the terms of the above equation gives

$$d(a^\star,b)=\frac{(a^\star-b)^2}{b}=\frac{b(1-\tau a^{\star 2})^2}{4\gamma^2a^{\star 2}}.$$

Consequently (50) reduces to

$$\frac{\widehat{\lambda}_i(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\overline{\gamma}^2\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2}\leq\frac{\rho}{p}\iff\overline{\gamma}^2\geq\frac{p\widehat{\lambda}_i(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2},\quad\text{for } i=1,...,p.$$

Since

$$\frac{p\widehat{\lambda}_i(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2}\leq\frac{p\widehat{\lambda}_p(1-\tau\varphi(\tau,\overline{\gamma},\widehat{\lambda}_i)^2)^2}{4\rho\widehat{\lambda}_1^2}\leq\max\left\{\frac{p\widehat{\lambda}_p}{4\rho\widehat{\lambda}_1^2},\frac{p\widehat{\lambda}_p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho\widehat{\lambda}_1^2}\right\},$$

where the last inequality is by Lemma 6, then we can set

$$\overline{\gamma}\triangleq\sqrt{\max\left\{\frac{p\widehat{\lambda}_p}{4\rho\widehat{\lambda}_1^2},\frac{p\widehat{\lambda}_p(1-\tau\widehat{\lambda}_p^2)^2}{4\rho\widehat{\lambda}_1^2}\right\}}.$$

□

Lemma 6. *Let $\tau>0$ and $f(x)=(1-\tau x^2)^2$. If $0\leq a\leq\overline{a}$, then $f(a)\leq\max\{1,(1-\tau\overline{a}^2)^2\}$.*

Proof. It is easy to verify that $f(a)\leq 1$ for $0\leq a\leq\sqrt{\frac{2}{\tau}}$. Then it suffices to show that $f(a)\leq f(\overline{a})$ for $\sqrt{2/\tau}\leq a\leq\overline{a}$, i.e., f is increasing over $[\sqrt{2/\tau},\infty)$. The derivative of f is $f'(x)=4\tau x(\tau x^2-1)$. Solving the inequality $f'(x)\geq 0$, we conclude that $f'(x)>0$ for $x\in[\sqrt{2/\tau},\infty)$. □

B.10 Proof of Proposition 9

Proof of Proposition 9. The proof is similar to Yue et al. (2024, Proposition 7). In this proof, we write $x^\star_{i,n} = \lambda_i(\Sigma^\star_n(\tau,\rho_n))$ and $\widehat{x}_{i,n} = \lambda_i(\widehat{\Sigma})$ for $i = 1,\dots,n$ and $n \in \mathbb{N}$. By the strong consistency assumption, $\widehat{\Sigma}_n$ converges to Σ_0 almost surely as $n \to \infty$. Now fix a particular realization of the uncertainties, where $\widehat{\Sigma}_n$ converges to Σ_0 deterministically. Then it holds that $\lim_{n\to\infty} \widehat{x}_{i,n} = \lambda_i(\Sigma_0)$ for $i = 1,\dots,p$ since the eigenvalue mapping λ_i is continuous. On the other hand, by Proposition 4(ii) we know that $0 \le x^\star_{i,n} \le \max\left\{\sqrt{\frac{1}{\tau}}, b\right\}$, i.e., $\{x^\star_{i,n}\}_{n\in\mathbb{N}}$ is bounded. Let $\{x^\star_{i,n_k}\}_{k\in\mathbb{N}}$ be a convergent subsequence of $\{x^\star_{i,n}\}_{n\in\mathbb{N}}$. Then by Assumption 3, it holds that

$$d\left(x^\star_{i,n_k}, \widehat{x}_{i,n_k}\right) \le \sum_{j=1}^{p} d(x^\star_{j,n_k}, \widehat{x}_{j,n_k}) = D\left(\Sigma^\star_n(\tau,\rho_n), \widehat{\Sigma}_n\right) \le \rho_{n_k}, \quad \forall k \in \mathbb{N}, \tag{54}$$

for $i = 1,\dots,p$. Since ρ_{n_k} converges to 0 and d is continuous by Assumption 3 (ii), then (54) implies that

$$d\left(\lim_{k\to\infty} x^\star_{i,n_k}, \lambda_i(\Sigma_0)\right) = d\left(\lim_{k\to\infty} x^\star_{i,n_k}, \lim_{k\to\infty} \widehat{x}_{i,n_k}\right) = \lim_{k\to\infty} d\left(x^\star_{i,n_k}, \widehat{x}_{i,n_k}\right) = 0. \tag{55}$$

Recall that by Assumption 3, d satisfies the identity of indiscernibles. Thus (55) implies that $\lim_{k\to\infty} x^\star_{i,n_k} = \lambda_i(\Sigma_0)$. This shows that every convergent subsequence of the bounded sequence $\{x^\star_{i,n}\}_{n\in\mathbb{N}}$ must have the same limit $\lambda_i(\Sigma_0)$. By (Abbott, 2015, Exercise 2.5.5), it holds that $\lim_{n\to\infty} x^\star_{i,n} = \lambda_i(\Sigma_0)$. The derivation so far applies to every uncertainty realization where $\widehat{\Sigma}_n$ converges to Σ_0 deterministically. Since $\widehat{\Sigma}_n$ converges to Σ_0 almost surely, we conclude that $x^\star_{i,n}$ converges to $\lambda_i(\Sigma_0)$ almost surely. This implies that

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} \|\Sigma^\star_n(\tau,\rho_n) - \Sigma_0\|_F = 0\right) &\ge \mathbb{P}\left(\lim_{n\to\infty} \left(\|\Sigma^\star_n(\tau,\rho_n) - \widehat{\Sigma}_n\|_F + \|\widehat{\Sigma}_n - \Sigma_0\|_F\right) = 0\right) \\
&= \mathbb{P}\left(\lim_{n\to\infty} \left(\|x^\star_n - \widehat{x}_n\|_2 + \|\widehat{\Sigma}_n - \Sigma_0\|_F\right) = 0\right) \\
&\ge \mathbb{P}\left(\lim_{n\to\infty} \left(\|x^\star_n - \lambda(\Sigma_0)\|_2 + \|\widehat{x}_n - \lambda(\Sigma_0)\|_2 + \|\widehat{\Sigma}_n - \Sigma_0\|_F\right) = 0\right) \\
&= 1,
\end{aligned}$$

where $x^\star_n$ and $\widehat{x}_n$ are the vectors of $x^\star_{i,n}, i = 1,\dots,p$ and $\widehat{x}_{i,n}, i = 1,\dots,p$, and $\lambda(\Sigma_0)$ is the eigenvalue vector of Σ_0. Both inequalities hold because of the triangle inequality. The first equality holds by the fact that $\|\Sigma^\star_n(\tau,\rho_n) - \widehat{\Sigma}_n\|_F = \|x^\star_n - \widehat{x}_n\|_2$ since $\Sigma^\star_n(\tau,\rho_n)$ and $\widehat{\Sigma}_n$ share the same eigenvectors. The last equality holds by the fact that $x^\star_n$ converges to $\lambda(\Sigma_0)$ almost surely, $\widehat{x}_n$ converges to $\lambda(\Sigma_0)$ almost surely, and $\widehat{\Sigma}_n$ converges to Σ_0 almost surely. This shows that $\Sigma^\star_n(\tau,\rho_n)$ converges to Σ_0 almost surely and completes the proof. □

B.11 Proof of Theorem 4

Recall that $\gamma_\lambda(\rho)$ is defined as the unique solution to $F_\lambda(\gamma) = \rho$, where F_λ is a strictly decreasing function mapping from $[0,+\infty)$ to $(0,\rho_{\max}]$. Then $\rho^\star_n$ can be represented by $\rho^\star_n = F_\lambda(\gamma^\star_n)$, where

$$\gamma^\star_n \triangleq \arg\min_{\gamma\in[0,+\infty)} \|\Sigma_0^{-1} - \tau^\star\Sigma_0 - \gamma\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F^2. \tag{56}$$

To show the rate of $\rho^\star_n$ as $n \to \infty$, we first analyze the rate of $F_\lambda(\gamma)$ as $\gamma \to \infty$.

Proposition 14 (Limit of $\gamma^2 F_\lambda(\gamma)$)**.** *Let Assumptions 1- 3 and 6 hold, and $\tau > 0$ and $\lambda \in \mathbb{R}_+^p$ be such that $\lambda_i \neq 1/\sqrt{\tau}$ for some $i = 1, \ldots, p$. Then*

$$\lim_{\gamma\to\infty} \gamma^2 F_\lambda(\gamma) = \sum_{i=1}^{p} C_{\lambda_i,d} \left(\frac{1/\lambda_i - \tau\lambda_i}{d''_{\lambda_i}(\lambda_i)} \right)^2, \tag{57}$$

where $C_{\lambda_i,d}$ is the constant defined as in Assumption 6.

Proof of Proposition 14. Recall that $F_\lambda(\gamma) = \sum_{i=1}^p d(\varphi(\tau,\gamma,\lambda_i),\lambda_i)$. It suffices to prove that

$$\lim_{\gamma\to\infty} \gamma^2 d(\varphi(\tau,\gamma,b),b) = C_{b,d} \left(\frac{1/b - \tau b}{d''_b(b)} \right)^2 \tag{58}$$

for any fixed $\tau > 0$ and $b > 0$. Let $f(a) = \frac{1}{a} - \tau a - \gamma d'_b(a)$. Recall that $\varphi(\tau,\gamma,b)$ is defined as the unique positive solution of $f(a) = 0$. By Taylor expansion of f at point b to the first order, we obtain

$$f(a) = \frac{1}{b} - \tau b + \left(-\frac{1}{b^2} - \tau - \gamma d''_b(b) \right)(a-b) + o(|a-b|).$$

Let $\varphi_{\tau,b}(\gamma) \triangleq \varphi(\tau,\gamma,b)$, and set $f(\varphi_{\tau,b}(\gamma)) = 0$, i.e.,

$$\varphi_{\tau,b}(\gamma) - b = \frac{1/b - \tau b}{1/b^2 + \tau + \gamma d''_b(b)} + o(|\varphi_{\tau,b}(\gamma) - b|).$$

Then

$$(\varphi_{\tau,b}(\gamma) - b)^2 = \left[\frac{1/b - \tau b}{1/b^2 + \tau + \gamma d''_b(b)} \right]^2 + o(|\varphi_{\tau,b}(\gamma) - b|^2).$$

Consequently

$$\lim_{\gamma\to\infty} \gamma^2 (\varphi_{\tau,b}(\gamma) - b)^2 = \lim_{\gamma\to\infty} \left[\frac{1/b - \tau b}{1/\gamma b^2 + \tau/\gamma + d''_b(b)} \right]^2 = \left(\frac{1/b - \tau b}{d''_b(b)} \right)^2.$$

Together with Assumption 6, we have

$$\lim_{\gamma\to\infty} \gamma^2 d(\varphi(\tau,\gamma,b),b) = \lim_{\gamma\to\infty} \frac{d(\varphi(\tau,\gamma,b),b)}{(\varphi_{\tau,b}(\gamma) - b)^2} \cdot \gamma^2 (\varphi_{\tau,b}(\gamma) - b)^2 = C_{b,d} \left(\frac{1/b - \tau b}{d''_b(b)} \right)^2.$$

The proof is complete. □

Now we are ready to prove Theorem 4.

Proof of Theorem 4. Since (56) is a convex quadratic minimization problem of γ, we can figure out the optimal solution

$$\gamma_n^\star = \frac{2 \left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2} \Sigma_0, \mathbb{E}_n \nabla_1 D(\Sigma_0, \widehat{\Sigma}_n) \right\rangle}{\|\mathbb{E}_n \nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)\|_F^2}.$$

By Assumption 7,

$$\lim_{n\to\infty} \gamma_n^\star / n = \lim_{n\to\infty} \frac{2n \left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2} \Sigma_0, \mathbb{E}_n \nabla_1 D(\Sigma_0, \widehat{\Sigma}_n) \right\rangle}{n^2 \|\mathbb{E}_n \nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)\|_F^2} = \frac{2C_2}{C_1^2}. \tag{59}$$

A combination of (57) and (59) yields

$$\lim_{n\to\infty} n^2 \rho_n^\star = \lim_{n\to\infty} (\gamma_n^\star)^2 F_\lambda(\gamma_n^\star) \cdot \lim_{n\to\infty} n^2/(\gamma_n^\star)^2 = \rho^\star \triangleq \frac{C_1^4}{4C_2^2} \sum_{i=1}^{p} C_{\lambda_i,d} \left(\frac{1/\lambda_i - \tau\lambda_i}{d''_{\lambda_i}(\lambda_i)} \right)^2.$$

The proof is complete. □

B.12 Proof of Corollary 1

To prove Corollary 1, we first verify that some divergence functions satisfy Assumption 6-7 under Assumption 8 and figure out constants $C_{b,d}, C_1$ and C_2. Then Corollary 1 follows directly from Theorem 4.

B.12.1 Proof of equation (29)

Proposition 15 (Locally quadratic KL divergence)**.** *Let* $d(a,b) = \frac{1}{2}(\frac{a}{b} - 1 - \log\frac{a}{b})$ *be defined on* $\mathbb{R}_{++} \times \mathbb{R}_{++}$*. Then for any* $b > 0$*,*

$$\lim_{a\to b} \frac{d(a,b)}{(a-b)^2} = \frac{1}{4b^2}. \tag{60}$$

Proof of Proposition 15. Let $y = \frac{a}{b} - 1$. Then

$$\lim_{y\to 0} \frac{d(a,b)}{(a-b)^2} = \lim_{y\to 0} \frac{y - \log(y+1)}{2b^2y^2} = \lim_{y\to 0} \frac{y - \left(y - \frac{y^2}{2} + \frac{y^3}{3} - \dots\right)}{2b^2y^2} = \lim_{y\to 0} \frac{\frac{y^2}{2} + o(y^2)}{2b^2y^2} = \frac{1}{4b^2}.$$

□

Proposition 16 (Non-degenerate gradient of KL divergence)**.** *Let*

$$D(\Sigma_1, \Sigma_2) \triangleq \frac{1}{2}\big(\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) - p + \log\det(\Sigma_2\Sigma_1^{-1})\big) \tag{61}$$

be defined on $\mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$*. Then*

$$\nabla_1 D(\Sigma_1, \Sigma_2) = \frac{1}{2}\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right). \tag{62}$$

Under Assumption 8,

$$\lim_{n\to\infty} n\|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = \frac{p+1}{2}\|\Sigma_0^{-1}\|_F > 0 \tag{63}$$

and hence

$$\lim_{n\to\infty} n\left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\right\rangle = \frac{p+1}{2}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right) > 0. \tag{64}$$

Proof of Proposition 16. Note that under Assumption 8, $(n\widehat{\Sigma}_n)^{-1}$ follows Inverse-Wishart distribution and thus

$$\mathbb{E}_n[(n\widehat{\Sigma}_n)^{-1}] = \frac{1}{n-p-1}\Sigma_0^{-1}.$$

Consequently we have

$$\mathbb{E}_n[\nabla_1 D(\Sigma_0 - \widehat{\Sigma}_n)] = \frac{1}{2}\mathbb{E}_n[\widehat{\Sigma}_n^{-1} - \Sigma_0^{-1}] = \frac{p+1}{2(n-p-1)}\Sigma_0^{-1}. \tag{65}$$

A combination of the two equations above yields

$$\lim_{n\to\infty} n\|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = \lim_{n\to\infty} \frac{n(p+1)}{2(n-p-1)}\|\Sigma_0^{-1}\|_F = \frac{p+1}{2}\|\Sigma_0^{-1}\|_F,$$

which is (63). Likewise, we can use (65) to establish (64), i.e.,

$$\lim_{n\to\infty} n\left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\right\rangle = \lim_{n\to\infty} \frac{n(p+1)}{2(n-p-1)}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)$$
$$= \frac{p+1}{2}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right) > 0,$$

where the last inequality is due to the fact that

$$\|\Sigma_0\|_F^2\|\Sigma_0^{-1}\|_F^2 = \left(\sum_{i=1}^p \lambda_i^2\right)\left(\sum_{i=1}^p \frac{1}{\lambda_i^2}\right) > p^2.$$

Note that the above inequality never becomes equality because $\lambda_1, \ldots, \lambda_p$ are not identical. □

Proof of equation (29). All the assumptions of Theorem 4 are fulfilled and the constants are $C_{b,d} = \frac{1}{4b^2}$, $C_1 = \frac{p+1}{2}\|\Sigma_0^{-1}\|_F$ and $C_2 = \frac{p+1}{2}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)$. So the limit is

$$\lim_{n\to\infty} n^2\rho_n^\star = \frac{C_1^4}{4C_2^2}\sum_{i=1}^p C_{\lambda_i,d}\left(\frac{1/\lambda_i - \tau^\star\lambda_i}{d''_{\lambda_i}(\lambda_i)}\right)^2 = \frac{(p+1)^2\|\Sigma_0^{-1}\|_F^4}{16\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2}\sum_{i=1}^p \left(1-\tau^\star{\lambda_i}^2\right)^2.$$

□

B.12.2 Proof of equation (30)

Proposition 17 (Locally-quadratic Wasserstein divergence)**.** *Let* $d(a,b) = a + b - 2\sqrt{ab}$ *defined on* $\mathbb{R}_+ \times \mathbb{R}_+$. *Then for any* $b > 0$, *we have*

$$\lim_{a\to b}\frac{d(a,b)}{(a-b)^2} = \frac{1}{4b}.$$

Proof of Proposition 17. Note that

$$d(a,b) = a + b - 2\sqrt{ab} = (\sqrt{a} - \sqrt{b})^2 = \left(\frac{a-b}{\sqrt{a}+\sqrt{b}}\right)^2.$$

So

$$\lim_{a\to b}\frac{d(a,b)}{(a-b)^2} = \lim_{a\to b}\frac{1}{(\sqrt{a}+\sqrt{b})^2} = \frac{1}{4b}.$$

This completes the proof. □

Proposition 18 (Non-degenerate gradient of Wasserstein divergence)**.** *Let*

$$D(\Sigma_1, \Sigma_2) = \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}) \tag{66}$$

be defined on $\mathbb{S}_+^p \times \mathbb{S}_+^p$ *and* $\nabla_1 D(\Sigma_1, \Sigma_2) = I - \Sigma_2(\Sigma_1\Sigma_2)^{-1/2}$. *Under Assumption 8, we have*

$$\lim_{n\to\infty} n\|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = \frac{(p+1)\sqrt{p}}{8}$$

and

$$\lim_{n\to\infty} n\left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\right\rangle = \frac{p+1}{8}\left(\mathrm{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\mathrm{Tr}(\Sigma_0)\right) > 0.$$

Proof of Proposition 18. For $\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n) = I - \widehat{\Sigma}_n(\Sigma_0\widehat{\Sigma}_n)^{-1/2}$, we have

$$\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] = \mathbb{E}_n[I - \widehat{\Sigma}_n(\Sigma_0\widehat{\Sigma}_n)^{-1/2}] = \mathbb{E}_n[I - \Sigma_0^{-1}(\Sigma_0\widehat{\Sigma}_n)^{1/2}].$$

The following proof is based on the expansion of $\mathbb{E}_n[(\Sigma_0\widehat{\Sigma}_n)^{1/2}]$ given in Lemma 7 (iv). For the first limit, we have

$$\begin{aligned}
\|\mathbb{E}_n[I - \Sigma_0^{-1}(\Sigma_0\widehat{\Sigma}_n)^{1/2}]\|_F^2 &= \operatorname{Tr}\left((I - \Sigma_0^{-1}\mathbb{E}_n[(\Sigma_0\Sigma_n)^{1/2}])^2\right) \\
&= \operatorname{Tr}\left(\left(\frac{p+1}{8n}I + \Sigma_0^{-1}\mathbb{E}_n[R(\widehat{\Sigma}_n)]\right)^2\right) \\
&= \frac{(p+1)^2 p}{64n^2} + o(1/n^2). \qquad (67)
\end{aligned}$$

Then

$$\lim_{n\to\infty} n\|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = \sqrt{\lim_{n\to\infty} n^2\|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F^2} = \frac{(p+1)\sqrt{p}}{8}.$$

For the second limit, we have

$$\begin{aligned}
&\left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[I - \Sigma_0^{-1}(\Sigma_0\widehat{\Sigma}_n)^{1/2}] \right\rangle \\
&= \frac{p+1}{8n}\left(\operatorname{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\operatorname{Tr}(\Sigma_0)\right) + \left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[R(\widehat{\Sigma}_n)] \right\rangle \\
&= \frac{p+1}{8n}\left(\operatorname{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\operatorname{Tr}(\Sigma_0)\right) + o(1/n). \qquad (68)
\end{aligned}$$

Then

$$\lim_{n\to\infty} n\left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \right\rangle = \frac{p+1}{8}\left(\operatorname{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\operatorname{Tr}(\Sigma_0)\right) > 0,$$

where the last inequality is by Lemma 8. □

Proof of equation (30). All the assumptions of Theorem 4 are fulfilled and the constants are $C_{b,d} = \frac{1}{4b}$, $C_1 = \frac{(p+1)\sqrt{p}}{8}$ and $C_2 = \frac{p+1}{8}\left(\operatorname{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\operatorname{Tr}(\Sigma_0)\right)$. So the limit is

$$\lim_{n\to\infty} n^2\rho_n^\star = \frac{C_1^4}{4C_2^2}\sum_{i=1}^p C_{\lambda_i,d}\left(\frac{1/\lambda_i - \tau^\star\lambda_i}{d''_{\lambda_i}(\lambda_i)}\right)^2 = \frac{(p+1)^2p^2}{256\left(\operatorname{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\operatorname{Tr}(\Sigma_0)\right)^2}\sum_{i=1}^p \frac{(1-\tau^\star\lambda_i^2)^2}{\lambda_i}.$$

□

Lemma 7 (Expectation of functions of sample covariance matrix)**.** *Let $\xi_1, \dots, \xi_n \in \mathbb{R}^p$ be i.i.d. normal random vectors with mean 0 and covariance Σ_0, and let $\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \xi_i\xi_i^\top$ be the sample covariance matrix. Then the following holds.*

(i) $\mathbb{E}_n[\widehat{\Sigma}_n^2] = \frac{1}{n}\Sigma_0^2 + \frac{1}{n}\operatorname{Tr}(\Sigma_0)\Sigma_0 + \Sigma_0^2$,

(ii) $\mathbb{E}_n[\widehat{\Sigma}_n\Sigma_0^{-1}\widehat{\Sigma}_n] = \frac{1+p+n}{n}\Sigma_0$,

(iii) $\mathbb{E}_n[\|\Sigma_0 - \widehat{\Sigma}_n\|_F^2] = \mathbb{E}_n\left[\mathrm{Tr}\left((\widehat{\Sigma}_n - \Sigma_0)^2\right)\right] = \frac{1}{n}(\mathrm{Tr}(\Sigma_0^2) + \mathrm{Tr}(\Sigma_0)^2)$,

(iv) $\mathbb{E}_n[(\Sigma_0\widehat{\Sigma}_n)^{1/2}] = \Sigma_0 - \frac{p+1}{8n}\Sigma_0 + \mathbb{E}_n[R(\widehat{\Sigma}_n)]$, *where the remainder satisfies* $\|\mathbb{E}_n R(\widehat{\Sigma}_n)\|_F = o(1/n)$.

Proof of Lemma 7. Let $S \sim \mathcal{W}_p(n, \Sigma)$ be a random matrix following Wishart distribution and A be a constant symmetric matrix, then by Gupta and Nagar (2018, Theorem 3.3.15 (ii)), $\mathbb{E}_n[SAS] = n\Sigma A\Sigma + n\mathrm{Tr}(\Sigma A)\Sigma + n^2\Sigma A\Sigma$. By the fact that $n\widehat{\Sigma}_n \sim \mathcal{W}_p(n, \Sigma_0)$, we prove the following results.

Part (i). $\mathbb{E}_n[\widehat{\Sigma}_n^2] = \frac{1}{n^2}\mathbb{E}_n[(n\widehat{\Sigma}_n)^2] = \frac{1}{n^2}(n\Sigma_0^2 + n\mathrm{Tr}(\Sigma_0)\Sigma_0 + n^2\Sigma_0^2) = \frac{1}{n}\Sigma_0^2 + \frac{1}{n}\mathrm{Tr}(\Sigma_0)\Sigma_0 + \Sigma_0^2$.

Part (ii).

$$\begin{aligned}
\mathbb{E}_n[\widehat{\Sigma}_n\Sigma_0^{-1}\widehat{\Sigma}_n] &= \frac{1}{n^2}\mathbb{E}_n[n\widehat{\Sigma}_n\Sigma_0^{-1}n\widehat{\Sigma}_n] \\
&= \frac{1}{n^2}(n\Sigma_0\Sigma_0^{-1}\Sigma_0 + n\mathrm{Tr}(\Sigma_0\Sigma_0^{-1})\Sigma_0 + n^2\Sigma_0\Sigma_0^{-1}\Sigma_0) \\
&= \frac{1}{n^2}(n\Sigma_0 + np\Sigma_0 + n^2\Sigma_0) \\
&= \frac{1+p+n}{n}\Sigma_0.
\end{aligned}$$

Part (iii).

$$\begin{aligned}
\mathbb{E}_n[\|\Sigma_0 - \widehat{\Sigma}_n\|_F^2] &= \mathbb{E}_n\left[\mathrm{Tr}\left((\widehat{\Sigma}_n - \Sigma_0)^2\right)\right] \\
&= \mathbb{E}_n[\mathrm{Tr}(\Sigma_0^2) + \mathrm{Tr}(\widehat{\Sigma}_n^2) - 2\mathrm{Tr}(\Sigma_0\widehat{\Sigma}_n)] \\
&= \mathrm{Tr}(\Sigma_0^2) + \mathrm{Tr}(\mathbb{E}_n[\widehat{\Sigma}_n^2]) - \frac{2}{n}\mathrm{Tr}(\Sigma_0\mathbb{E}_n[n\widehat{\Sigma}_n]) \\
&= \mathrm{Tr}\left(\frac{1}{n}\Sigma_0^2 + \frac{1}{n}\mathrm{Tr}(\Sigma_0)\Sigma_0 + \Sigma_0^2\right) - \mathrm{Tr}(\Sigma_0^2) \\
&= \frac{1}{n}(\mathrm{Tr}(\Sigma_0^2) + \mathrm{Tr}(\Sigma_0)^2).
\end{aligned}$$

Part (iv). Define matrix-valued function $f(X) = (\Sigma_0 X)^{\frac{1}{2}}$. Taylor's expansion (in the Fréchet differentiable sense) at $X = \Sigma_0$ gives

$$f(X) = \Sigma_0 + \frac{1}{2}(X - \Sigma_0) - \frac{1}{8}(X - \Sigma_0)\Sigma_0^{-1}(X - \Sigma_0) + R(X),$$

where $\|R(X)\|_F = o(\|\Sigma_0 - \widehat{\Sigma}_n\|_F^2)$. Evaluating $\mathbb{E}_n[f(X)]$ at $X = \widehat{\Sigma}_n$ gives

$$\mathbb{E}_n[(\Sigma_0\widehat{\Sigma}_n)^{\frac{1}{2}}] = \mathbb{E}_n[f(\widehat{\Sigma}_n)] = \Sigma_0 - \frac{1}{8}\mathbb{E}_n\left[(\widehat{\Sigma}_n - \Sigma_0)\Sigma_0^{-1}(\widehat{\Sigma}_n - \Sigma_0)\right] + \mathbb{E}_n[R(\widehat{\Sigma}_n)].$$

By Part (ii), we have $\mathbb{E}_n\left[(\widehat{\Sigma}_n - \Sigma_0)\Sigma_0^{-1}(\widehat{\Sigma}_n - \Sigma_0)\right] = \mathbb{E}_n[\widehat{\Sigma}_n\Sigma_0^{-1}\widehat{\Sigma}_n - 2\widehat{\Sigma}_n + \Sigma_0] = \frac{p+1}{n}\Sigma_0$. Then

$$\mathbb{E}_n[(\Sigma_0\widehat{\Sigma}_n)^{1/2}] = \Sigma_0 - \frac{p+1}{8n}\Sigma_0 + \mathbb{E}_n[R(\widehat{\Sigma}_n)],$$

where

$$\|\mathbb{E}_n[R(\widehat{\Sigma}_n)]\|_F \le \mathbb{E}_n[\|R(\widehat{\Sigma}_n)\|_F] = \mathbb{E}_n[o(\|\Sigma_0 - \widehat{\Sigma}_n\|_F^2)] = o(\mathbb{E}_n[\|\Sigma_0 - \widehat{\Sigma}_n\|_F^2]) = o(1/n).$$

This completes the proof. □

Lemma 8 (Trace-inverse inequality). *For any $\Sigma_0 \in \mathbb{S}^p_{++}$, it holds that*

$$\mathrm{Tr}(\Sigma_0^{-1}) - \frac{p}{\|\Sigma_0\|_F^2}\mathrm{Tr}(\Sigma_0) \geq 0,$$

where the equality holds if and only if all the eigenvalues of Σ_0 are identical.

Proof of Lemma 8. Let $\lambda_1, \ldots, \lambda_p$ be eigenvalues of Σ_0. Then, it suffices to prove

$$\sum_{i=1}^p \frac{1}{\lambda_i} \geq \frac{p}{\sum_{i=1}^p \lambda_i^2} \sum_{i=1}^p \lambda_i. \tag{69}$$

By the arithmetic mean-harmonic mean, we have $\sum_{i=1}^p \frac{1}{\lambda_i} \geq \frac{p^2}{\sum_{i=1}^p \lambda_i}$, where the inequality becomes equality if and only if all the eigenvalues are identical, i.e., $\lambda_1 = \lambda_2 = \ldots = \lambda_p$. By the Cauchy-Schwarz inequality, we have

$$\left(\sum_{i=1}^p \lambda_i \cdot 1\right)^2 \leq p \sum_{i=1}^p \lambda_i^2 \;\Rightarrow\; \frac{p}{\left(\sum_{i=1}^p \lambda_i\right)^2} \geq \frac{1}{\sum_{i=1}^p \lambda_i^2} \;\Rightarrow\; \frac{p^2}{\sum_{i=1}^p \lambda_i} \geq \frac{p}{\sum_{i=1}^p \lambda_i^2}\left(\sum_{i=1}^p \lambda_i\right),$$

where the inequalities become equalities if and only if $\lambda_1 = \lambda_2 = \ldots = \lambda_p$. Combining the above two inequalities completes the proof. □

B.12.3 Proof of equation (31)

Proposition 19 (Locally-quadratic Symmetrized Stein divergence). *Let $d(a,b) = \frac{1}{2}(\frac{b}{a} + \frac{a}{b} - 2)$ defined on $\mathbb{R}_{++} \times \mathbb{R}_{++}$. Then for any $b > 0$, we have*

$$\lim_{a \to b} \frac{d(a,b)}{(a-b)^2} = \frac{1}{2b^2}.$$

Proof of Proposition 19. Note that

$$d(a,b) = \frac{b^2 + a^2 - 2ab}{2ab} = \frac{(a-b)^2}{2ab}.$$

So

$$\lim_{a \to b} \frac{d(a,b)}{(a-b)^2} = \frac{1}{2b^2}.$$

□

Proposition 20 (Non-degenerate gradient of Symmetrized Stein divergence). *Let*

$$D(\Sigma_1, \Sigma_2) = \frac{1}{2}\left(\mathrm{Tr}(\Sigma_1\Sigma_2^{-1} + \Sigma_2\Sigma_1^{-1}) - 2p\right) \tag{70}$$

be defined on $\mathbb{S}^p_{++} \times \mathbb{S}^p_{++}$ and $\nabla_1 D(\Sigma_1, \Sigma_2) = \frac{1}{2}\left(\Sigma_2^{-1} - \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}\right)$. Under Assumption 8, we have

$$\lim_{n \to \infty} n \|\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)]\|_F = \frac{p+1}{2}\|\Sigma_0^{-1}\|_F$$

and

$$\lim_{n \to \infty} n \left\langle \Sigma_0^{-1} - \frac{p}{\|\Sigma_0\|_F^2}\Sigma_0, \mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] \right\rangle = \frac{p+1}{2}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right) > 0.$$

Proof of Proposition 20. For $\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n) = \frac{1}{2}\left(\widehat{\Sigma}_n^{-1} - \Sigma_0^{-1}\widehat{\Sigma}_n\Sigma_0^{-1}\right)$, we have

$$\mathbb{E}_n[\nabla_1 D(\Sigma_0, \widehat{\Sigma}_n)] = \frac{1}{2}\left(\frac{n}{n-p-1}\Sigma_0 - \Sigma_0^{-1}\right) = \frac{p+1}{2(n-p-1)}\Sigma_0^{-1}.$$

We are in the same case as in Proposition 16. The result can be proven similarly, and thus we omit the proof. □

Proof of equation (31). All the assumptions of Theorem 4 are fulfilled and the constants are $C_{b,d} = \frac{1}{2b^2}$, $C_1 = \frac{p+1}{2}\|\Sigma_0^{-1}\|_F$ and $C_2 = \frac{p+1}{2}\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)$. So the limit is

$$\lim_{n\to\infty} n^2\rho_n^\star = \frac{C_1^4}{4C_2^2}\sum_{i=1}^{p} C_{\lambda_i,d}\left(\frac{1/\lambda_i - \tau^\star\lambda_i}{d''_{\lambda_i}(\lambda_i)}\right)^2 = \frac{(p+1)^2\|\Sigma_0^{-1}\|_F^4}{32\left(\|\Sigma_0^{-1}\|_F^2 - \frac{p^2}{\|\Sigma_0\|_F^2}\right)^2}\sum_{i=1}^{p}\left(1-\tau^\star{\lambda_i}^2\right)^2.$$

□