

A Geometric Unification of Distributionally Robust Covariance Estimators: Shrinking the Spectrum by Inflating the Ambiguity Set

MAN-CHUNG YUE, YVES RYCHENER, DANIEL KUHN, VIET ANH NGUYEN

ABSTRACT. The state-of-the-art methods for estimating high-dimensional covariance matrices all shrink the eigenvalues of the sample covariance matrix towards a data-insensitive shrinkage target. The underlying shrinkage transformation is either chosen *heuristically*—without compelling theoretical justification—or *optimally* in view of restrictive distributional assumptions. In this paper, we propose a principled approach to construct covariance estimators without imposing restrictive assumptions. That is, we study distributionally robust covariance estimation problems that minimize the worst-case Frobenius error with respect to all data distributions close to a nominal distribution, where the proximity of distributions is measured via a divergence on the space of covariance matrices. We identify conditions on this divergence under which the resulting minimizers represent shrinkage estimators. We show that the corresponding shrinkage transformations are intimately related to the geometrical properties of the underlying divergence. We also prove that our robust estimators are efficiently computable and asymptotically consistent and that they enjoy finite-sample performance guarantees. We exemplify our general methodology by synthesizing explicit estimators induced by the Kullback-Leibler, Fisher-Rao, and Wasserstein divergences. Numerical experiments based on synthetic and real data show that our robust estimators are competitive with state-of-the-art estimators.

1. INTRODUCTION

The covariance matrix Σ_0 of a random vector $\xi \in \mathbb{R}^p$ is a fundamental summary statistic that captures the dispersion of ξ . Together with the mean vector μ_0 , it characterizes a unique member of the family of Gaussian distributions, which occupies the central stage in statistics and probability theory. Hence, any probabilistic model involving Gaussian distributions requires an estimate of Σ_0 as an input. For example, Gaussian distributions are ubiquitous in finance (*e.g.*, in portfolio theory [41]), in statistical learning (*e.g.*, in linear and quadratic discriminant analysis [20, § 4.3]) or control and signal processing (*e.g.*, in Kalman filtering [25]). In addition, Σ_0 is intimately related to the correlation matrix, including the Pearson correlation coefficients [48], and it permeates medical statistics [60] and correlation network analysis [13, 40] etc.

If the distribution \mathbb{P} of ξ is *known*, then the mean vector $\mu_0 = \mathbb{E}_{\mathbb{P}}[\xi]$ and the covariance matrix $\Sigma_0 = \mathbb{E}_{\mathbb{P}}[(\xi - \mu_0)(\xi - \mu_0)^\top]$ can be obtained by evaluating the relevant integrals with respect to \mathbb{P} —either analytically or via numerical integration quadratures. If \mathbb{P} is *unknown*, however, one typically has to estimate μ_0 and Σ_0 from n independent samples $\hat{\xi}_1, \dots, \hat{\xi}_n \sim \mathbb{P}$. Arguably the simplest estimators for μ_0 and Σ_0 are the sample mean $\hat{\mu}_{\text{SA}} = \frac{1}{n} \sum_{i=1}^n \hat{\xi}_i$ and the sample covariance matrix $\hat{\Sigma}_{\text{SA}} = \frac{1}{n-1} \sum_{i=1}^n (\hat{\xi}_i - \hat{\mu}_{\text{SA}})(\hat{\xi}_i - \hat{\mu}_{\text{SA}})^\top$, respectively. An elementary

Date: November 23, 2025.

The authors are with the University of Hong Kong (mcyue@hku.hk), the Ecole Polytechnique Fédérale de Lausanne (yves.rychener, daniel.kuhn@epfl.ch), and the Chinese University of Hong Kong (nguyen@se.cuhk.edu.hk).

calculation shows that $\hat{\Sigma}_{\text{SA}}$ is unbiased. Up to scaling, $\hat{\Sigma}_{\text{SA}}$ further coincides with the maximum likelihood estimator for Σ_0 provided that \mathbb{P} constitutes a normal distribution. In 1975, much to the surprise of statisticians, Charles Stein showed that one can strictly reduce the mean squared error of $\hat{\Sigma}_{\text{SA}}$ by shrinking it towards a constant matrix independent of the data [23, 57]. Even though it improves the mean squared error, Stein’s shrinkage transformation suffers from two major shortcomings, that is, it may alter the order of the estimator’s eigenvalues and may even render some eigenvalues negative [51]. Nonetheless, since Stein’s surprising discovery, the study of shrinkage estimators embodies an important research area in statistics.

Note also that $\hat{\Sigma}_{\text{SA}}$ is ill-conditioned if $p \lesssim n$ and even singular if $p > n$ [63]. Indeed, as $\hat{\Sigma}_{\text{SA}}$ is unbiased and as the maximum eigenvalue function is convex on the space of symmetric matrices, Jensen’s inequality ensures that the largest eigenvalue of $\hat{\Sigma}_{\text{SA}}$ exceeds, in expectation, the largest eigenvalue of Σ_0 . Similarly, the smallest eigenvalue of $\hat{\Sigma}_{\text{SA}}$ undershoots, in expectation, the smallest eigenvalue of Σ_0 . Hence, the condition number of $\hat{\Sigma}_{\text{SA}}$, defined as the ratio of its largest to its smallest eigenvalue, tends to exceed the condition number of Σ_0 . This effect is most pronounced if Σ_0 is (approximately) proportional to the identity matrix I_p and is exacerbated with increasing dimension p . A simple and effective method to improve the condition number is to construct a *linear* shrinkage estimator by forming a convex combination of $\hat{\Sigma}_{\text{SA}}$ and a data-insensitive shrinkage target such as $\frac{1}{p} \text{Tr}[\hat{\Sigma}_{\text{SA}}]I_p$ [32]. Other popular shrinkage targets include the constant correlation model [31], that is, a modified sample covariance matrix under which all pairwise correlations are equalized, the single index model [30], that is, the sum of a rank-one and a diagonal matrix representing systematic and idiosyncratic risk factors as in Sharpe’s single index model [56], and the diagonal matrix model [61], that is, the diagonal matrix that contains all sample eigenvalues on its main diagonal. The shrinkage weight of $\hat{\Sigma}_{\text{SA}}$ is usually tuned to minimize the Frobenius risk, that is, the expected squared Frobenius norm distance between the estimator and Σ_0 . Linear shrinkage estimators can be computed highly efficiently, improve the condition number of the sample covariance matrix, and are guaranteed to have full rank even if $p > n$.

In the remainder of the paper, we focus on covariance estimators that depend on the samples only indirectly through the sample covariance matrix. This assumption is unrestrictive. Indeed, it is satisfied by all commonly used covariance estimators. Moreover, it comes at no loss of generality if \mathbb{P} is a normal distribution, in which case $\hat{\Sigma}_{\text{SA}}$ constitutes a sufficient statistic for Σ_0 . Without prior information about the eigenvectors of Σ_0 , it is natural to restrict attention to rotation equivariant estimators. Rotation equivariance means that evaluating the estimator $\hat{\Sigma}$ on the rotated dataset $\{\widehat{R\xi}_i\}_{i=1}^N$ is equivalent to evaluating the rotated estimator $R\hat{\Sigma}R^\top$ on the original dataset $\{\widehat{\xi}_i\}_{i=1}^n$ for any rotation matrix R . One can show that any rotation equivariant estimator $\hat{\Sigma}$ commutes with the sample covariance matrix $\hat{\Sigma}_{\text{SA}}$, that is, $\hat{\Sigma}_{\text{SA}}$ and $\hat{\Sigma}$ share the same eigenvectors, and the spectrum of $\hat{\Sigma}$ can be viewed as a transformation of the spectrum of $\hat{\Sigma}_{\text{SA}}$ [49, Lemma 5.3]. Such spectral transformations are referred to as shrinkage transformations. Note that the linear shrinkage estimators discussed above are rotation equivariant only if the shrinkage target commutes with $\hat{\Sigma}_{\text{SA}}$.

If \mathbb{P} is governed by a spiked covariance model, that is, if \mathbb{P} is Gaussian, p and n tend to infinity at an asymptotically constant ratio and Σ_0 constitutes a fixed-rank perturbation of the identity matrix, then one can use results from random matrix theory to construct the best rotation equivariant estimators in closed form for a broad range of different loss functions [12]. Nonlinear

shrinkage estimators that are asymptotically optimal with respect to the Frobenius loss can also be constructed in the absence of any normality assumptions, and they can significantly improve on linear shrinkage estimators if the eigenvalue spectrum of Σ_0 is dispersed [33, 35]. Similarly, one can construct optimal shrinkage estimators for the *inverse* covariance matrix Σ_0^{-1} , which is usually termed the precision matrix; see [8, 36]. However, the available statistical guarantees for all shrinkage estimators described above are *asymptotic* and depend on assumptions about the structure of \mathbb{P} and/or the convergence properties of the spectral distribution of $\hat{\Sigma}_{\text{SA}}$, which may be difficult to check in practice.

In this paper, we propose a flexible and principled approach to estimate the covariance matrix Σ_0 by using ideas from distributionally robust optimization (DRO). Specifically, our approach generates a rich family of covariance matrix estimators corresponding to different ambiguity sets that can encode prior distributional information. All emerging estimators are rotation equivariant and thus represent nonlinear shrinkage estimators. In addition, they all improve the condition number of the sample covariance matrix, are invertible, and preserve the order of the sample eigenvalues. They also offer finite sample guarantees on the prediction loss and are asymptotically consistent. These appealing properties are not enforced ad hoc but emerge naturally from the solution of a principled distributionally robust estimation model. We emphasize that our results do not rely on any restrictive assumptions such as the requirement that \mathbb{P} is Gaussian or that the spectral distribution of $\hat{\Sigma}_{\text{SA}}$ converges to a well-defined limit as p and n tend to infinity at a constant ratio.

To develop the distributionally robust estimation model to be studied in this paper, we first express the unknown true covariance matrix Σ_0 as the minimizer of a stochastic optimization problem involving the unknown probability distribution \mathbb{P} . Specifically, adopting the standard assumption that $\mu_0 = \mathbb{E}_{\mathbb{P}}[\xi] = 0$ [32, 33, 34, 36] and noting that the squared Frobenius norm is strictly convex, we obtain

$$\{\Sigma_0\} = \underset{X \in \mathbb{S}_+^p}{\text{Argmin}} \|X - \Sigma_0\|_F^2 = \underset{X \in \mathbb{S}_+^p}{\text{Argmin}} \text{Tr}[X^2] - 2 \text{Tr}[X\Sigma_0] = \underset{X \in \mathbb{S}_+^p}{\text{Argmin}} \text{Tr}[X^2] - 2 \text{Tr}[X\mathbb{E}_{\mathbb{P}}[\xi\xi^\top]].$$

If we could solve the stochastic optimization problem on the right-hand side of the above expression, we could precisely recover the ideal estimator $X^* = \Sigma_0$. This is impossible, however, because the distribution \mathbb{P} needed to evaluate the stochastic optimization problem's objective function is unknown. Nevertheless, replacing \mathbb{P} with a nominal distribution $\hat{\mathbb{P}}$ constructed from the n training samples yields the nominal estimation model

$$\min_{X \in \mathbb{S}_+^p} \text{Tr}[X^2] - 2\mathbb{E}_{\hat{\mathbb{P}}}[\xi^\top X \xi], \quad (1)$$

which requires no unavailable inputs. An elementary calculation shows that (1) is uniquely solved by $\hat{\Sigma} = \mathbb{E}_{\hat{\mathbb{P}}}[\xi\xi^\top]$, which is the covariance matrix of ξ under the nominal distribution $\hat{\mathbb{P}}$, provided that $\hat{\mu} = \mathbb{E}_{\hat{\mathbb{P}}}[\xi] = 0$. Of course, characterizing $\hat{\Sigma}$ as a minimizer of (1) has no conceptual or computational benefits because we have to compute the integral $\mathbb{E}_{\hat{\mathbb{P}}}[\xi\xi^\top]$ already to evaluate the objective function of (1). Nevertheless, the nominal estimation problem (1) is useful because it allows us to construct a broad range of nonlinear shrinkage estimators in a principled and systematic manner by robustifying the prediction loss.

Any nominal distribution $\hat{\mathbb{P}}$ constructed from a finite dataset must invariably differ from the true data-generating distribution \mathbb{P} . Estimation errors in $\hat{\mathbb{P}}$ are conveniently captured by an

ambiguity set of the form

$$\mathbb{U}_\varepsilon(\hat{\mathbb{P}}) = \left\{ \mathbb{Q} : \mathbb{Q} \sim (0, \Sigma), D(\Sigma, \hat{\Sigma}) \leq \varepsilon \right\}, \quad (2)$$

where $\mathbb{Q} \sim (0, \Sigma)$ indicates that ξ has mean 0 and covariance matrix Σ under \mathbb{Q} , and D represents a divergence on the space of positive semidefinite matrices. Divergences are general distance-like functions that are non-negative and satisfy the identity of indiscernibles (that is, they satisfy $D(\Sigma, \hat{\Sigma}) = 0$ if and only if $\Sigma = \hat{\Sigma}$). However, divergences may fail to be symmetric and may violate the triangle inequality. Intuitively, $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$ can be viewed as a divergence ball of radius $\varepsilon \geq 0$ around $\hat{\mathbb{P}}$ in the space of probability distributions. Robustifying the nominal estimation problem (1) against all distributions in $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$ yields the following DRO problem.

$$\min_{X \in \mathbb{S}_+^p} \sup_{\mathbb{Q} \in \mathbb{U}_\varepsilon(\hat{\mathbb{P}})} \text{Tr}[X^2] - 2\mathbb{E}_{\mathbb{Q}}[\xi^\top X \xi]. \quad (3)$$

Problem (3) seeks an estimator X that minimizes the worst-case expected prediction loss across all distributions in $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$. Note that if $\varepsilon = 0$, then the DRO problem (3) collapses to the nominal estimation problem (1) because the divergence D satisfies the identity of indiscernibles, which ensures that $\mathbb{U}_0(\hat{\mathbb{P}}) = \{\hat{\mathbb{P}}\}$. Hence, (3) embeds (1) into a family of estimation models parametrized by D and ε . Moreover, DRO models naturally bridge optimization and statistics in that they offer an intuitive way to derive generalization bounds. Indeed, if ε is tuned to ensure that $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$ contains the data-generating distribution \mathbb{P} with high confidence $1 - \beta$, then the optimal value of the DRO problem (3) provides a $(1 - \beta)$ -upper confidence bound on the prediction loss of its unique minimizer X^\star under \mathbb{P} [42]. Stronger generalization bounds that do not require \mathbb{P} to belong to $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$ are provided in [7, 15]. Even if the ambiguity set does not contain \mathbb{P} , DRO models tend to yield high-quality solutions because there is a deep connection between robustification and regularization [16, 53, 54]. This connection may also explain the empirical success of DRO in statistical estimation [6, 27, 59].

The flexibility to choose the divergence D underlying the ambiguity set $\mathbb{U}_\varepsilon(\hat{\mathbb{P}})$ is both a blessing and a curse. On the one hand, D can encode prior distributional information and thus lead to better estimators. On the other hand, the family of divergences is vast. Hence, the choice of a suitable instance could overwhelm the modeler. Given the statistical estimation task at hand, it makes sense to restrict attention to divergences that admit a statistical interpretation. Many popular divergences on the space of covariance matrices are obtained by restricting a divergence on the space of probability distributions to the family of normal distributions. For example, the Kullback-Leibler divergence, the 2-Wasserstein distance, or the Fisher-Rao distance between zero-mean normal distributions all admit closed-form formulas in terms of the distributions' covariance matrices. These 'Gaussian' divergences are popular because they are conducive to tractable DRO models in risk management [17, 44], ethical machine learning [10, 66], likelihood evaluation [46, 47], Kalman filtering [71, 55] and control [58] etc. In addition, the shrinkage estimator for the *inverse* covariance matrix proposed in [43] also leverages a 'Gaussian' divergence. Nonetheless, the approach proposed in this paper does *not* rely on the assumption that \mathbb{P} is Gaussian.

The main contributions of this paper can be summarized as follows.

- We propose a rich family of distributionally robust covariance matrix estimators. Each estimator is defined as a solution of (3) for a particular ambiguity set of the form (2). Here, the nominal covariance matrix $\hat{\Sigma}$ characterizes the *center*, the divergence D determines the

geometry, and the radius ε determines the *size* of the ambiguity set. We demonstrate that all such estimators are well-defined, unique and efficiently computable under few structural assumptions on D and mild regularity conditions on $\hat{\Sigma}$ and ε .

- We prove that our distributionally robust covariance matrix estimators constitute nonlinear shrinkage estimators, that is, they have the same eigenbasis as $\hat{\Sigma}$, and their eigenvalues are obtained by shrinking the spectrum of $\hat{\Sigma}$ towards 0 by using a nonlinear shrinkage transformation depending on D and a shrinkage intensity depending on ε . We further prove that these estimators improve the condition number of $\hat{\Sigma}$.
- We identify various divergences commonly used in statistics, machine learning and information theory that satisfy the requisite regularity conditions. To this end, we invoke a generalization of Sion's classical minimax theorem from Euclidean spaces to Riemannian manifolds. We also exemplify our framework by deriving explicit analytical formulas for the distributionally robust covariance estimators induced by the Kullback-Leibler divergence, the 2-Wasserstein distance and the Fisher-Rao distance.
- We prove that, if ε scales with the sample size n as $\mathcal{O}(n^{-\frac{1}{2}})$, then the proposed estimators are strongly consistent and enjoy finite-sample performance guarantees at a fixed confidence level. Numerical experiments based on synthetic as well as real data for portfolio optimization and binary classification tasks suggest that our robust estimators are competitive with state-of-the-art estimators from the literature.

The first robustness interpretation of a shrinkage estimator was discovered in the context of *inverse* covariance matrix estimation [43]. Specifically, it was shown that a particular nonlinear shrinkage estimator can be obtained by robustifying the maximum likelihood estimator for Σ_0^{-1} across all Gaussian distributions of the training samples within a prescribed Wasserstein ball. This result critically relies on the restrictive assumption that the unknown data-generating distribution, the nominal distribution as well as all other distributions in the Wasserstein ball are Gaussian. In addition, this result has not been extended to more general ambiguity sets based on other divergences beyond the 2-Wasserstein distance, thus limiting the modeler's flexibility.

In this paper we show that a broad spectrum of shrinkage estimators for Σ_0 can be obtained from a versatile DRO model that does not rely on restrictive normality assumptions. That is, we seek the most general conditions on the DRO model under which a shrinkage effect emerges. In addition, we uncover a deep connection between the geometry of the ambiguity set, which is determined by the choice of the divergence D , and the nonlinear shrinkage transformation of the corresponding distributionally robust estimator.

Notation. We use $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ as a shorthand for the extended real line. The space of p -dimensional real vectors and its subsets of (entry-wise) non-negative and positive vectors are denoted by \mathbb{R}^p , \mathbb{R}_+^p , and \mathbb{R}_{++}^p , respectively. Similarly, the space of symmetric matrices in $\mathbb{R}^{p \times p}$, as well as its subsets of positive semidefinite and positive definite matrices, are denoted by \mathbb{S}^p , \mathbb{S}_+^p , and \mathbb{S}_{++}^p , respectively. The group of orthogonal matrices in $\mathbb{R}^{p \times p}$ is denoted by \mathcal{O}_p , and I_p stands for the identity matrix in $\mathbb{R}^{p \times p}$. For any $x \in \mathbb{R}^p$, we use x^\downarrow and x^\uparrow to denote the vectors obtained by rearranging the entries of x in non-increasing and non-decreasing order, respectively. The trace of a matrix $S \in \mathbb{S}^p$ is defined as $\text{Tr}[S] = \sum_{i=1}^p S_{ii}$. Finally, $\|M\| = \sup_{\|v\|_2=1} \|Mv\|_2$ and $\|M\|_F = \text{Tr}[M^\top M]^{\frac{1}{2}}$ stand for the spectral norm and the Frobenius norm of M , respectively.

2. OVERVIEW OF MAIN RESULTS

The distributionally robust estimation problem (3) perturbs—and thereby hopefully improves—the nominal estimator $\widehat{\Sigma}$ in view of the divergence D . We now derive a simple reformulation of (3) as a standard minimization problem, and we informally outline the main properties of the corresponding optimal solution, which will be established rigorously in the remainder of the paper. From now on, the nominal covariance matrix $\widehat{\Sigma}$ can be viewed as any naïve initial estimator for the covariance matrix Σ_0 . The construction of $\widehat{\Sigma}$ from the samples $\widehat{\xi}_1, \dots, \widehat{\xi}_n$ is immaterial for most of our discussion. As the loss function underlying problem (3) is quadratic in ξ and as $\mathbb{E}_{\mathbb{Q}}[\xi] = 0$, its expected value depends on \mathbb{Q} only indirectly through the covariance matrix $\Sigma = \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top]$. Thus, the DRO problem (3) is equivalent to the robust covariance estimation problem

$$\min_{X \in \mathbb{S}_+^p} \max_{\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})} \text{Tr}[X^2] - 2 \text{Tr}[\Sigma X] \quad (4)$$

with uncertainty set

$$\mathcal{B}_\varepsilon(\widehat{\Sigma}) = \left\{ \Sigma \in \mathbb{S}_+^p : D(\Sigma, \widehat{\Sigma}) \leq \varepsilon \right\}. \quad (5)$$

We stress that the divergence function D may fail to be symmetric, that is, $D(X, Y)$ may differ from $D(Y, X)$. It is therefore important to remember the convention that $\widehat{\Sigma}$ is the *second* argument of D in the definition of $\mathcal{B}_\varepsilon(\widehat{\Sigma})$. Note also that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ grows with the size parameter ε and collapses to the singleton $\{\widehat{\Sigma}\}$ for $\varepsilon = 0$. The robust estimation problem (4) constitutes a zero-sum game between the statistician, who moves first and chooses the estimator X , and nature, who moves second and chooses the covariance matrix Σ . The following dual estimation problem is obtained by interchanging the order of minimization and maximization in (4).

$$\max_{\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})} \min_{X \in \mathbb{S}_+^p} \text{Tr}[X^2] - 2 \text{Tr}[\Sigma X] \quad (6)$$

From now on, we denote by X^* and Σ^* the optimal solutions of the primal and dual estimation problems (4) and (6), respectively. In Section 3.1 below, we will identify few conditions on D and $\widehat{\Sigma}$ under which X^* and Σ^* are indeed guaranteed to exist and to be unique. If the uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is convex and compact, then strong duality prevails (that is, (4) and (6) share the same optimal value) by Sion's classical minimax theorem. As several popular divergence functions are non-convex in their first argument and thus induce a non-convex uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$; however, we will invoke a generalized minimax theorem that guarantees strong duality under significantly more general conditions. Whenever strong duality holds, (X^*, Σ^*) constitutes a Nash equilibrium of the zero-sum game between the statistician and nature [52, Lemma 36.2].

A cursory glance at its first-order optimality condition reveals that the inner minimization problem in (6) is solved by $X = \Sigma$. Hence, the inner minimum evaluates to $-\text{Tr}[\Sigma^2] = -\|\Sigma\|_{\text{F}}^2$. Eliminating the factor -1 further shows that Σ^* solves the maximization problem (6) if and only if it solves the minimization problem

$$\min_{\Sigma \in \mathbb{S}_+^p} \left\{ \|\Sigma\|_{\text{F}}^2 : D(\Sigma, \widehat{\Sigma}) \leq \varepsilon \right\}. \quad (\text{P}_{\text{Mat}})$$

Thus, nature's Nash strategy Σ^* can be computed by solving (P_{Mat}) instead of (6). By the defining properties of Nash strategies, the statistician's Nash strategy X^* must be a best response to Σ^* , that is, X^* must solve the inner minimization problem in (6) for $\Sigma = \Sigma^*$. However, the unique optimal solution of this minimization problem is Σ^* . In summary, this reasoning

implies that if strong duality holds, then the Nash strategies X^* and Σ^* of the statistician and nature coincide and are both given by the unique minimizer of problem (P_{Mat}) .

Problem (P_{Mat}) is reminiscent of a ridge regression problem [21, 64], which seeks an estimator that minimizes a weighted sum of a least squares fidelity term and a Frobenius norm regularization term. Indeed, problem (P_{Mat}) seeks a covariance matrix Σ with minimum Frobenius norm and a fidelity error of at most ε , where the fidelity of Σ with respect to the nominal covariance estimator $\hat{\Sigma}$ is measured by the divergence $D(\Sigma, \hat{\Sigma})$.

Divergence function	$D(\Sigma, \hat{\Sigma})$	Domain
Kullback-Leibler / Stein [28]	$\frac{1}{2} (\text{Tr}[\hat{\Sigma}^{-1}\Sigma] - p + \log \det(\hat{\Sigma}\Sigma^{-1}))$	$\mathbb{S}_{++}^p \times \mathbb{S}_{++}^p$
Wasserstein [18]	$\text{Tr}[\Sigma + \hat{\Sigma} - 2(\Sigma\hat{\Sigma})^{\frac{1}{2}}]$	$\mathbb{S}_+^p \times \mathbb{S}_+^p$
Fisher-Rao [3]	$\left\ \log(\hat{\Sigma}^{-\frac{1}{2}}\Sigma\hat{\Sigma}^{-\frac{1}{2}}) \right\ _{\text{F}}^2$	$\mathbb{S}_{++}^p \times \mathbb{S}_{++}^p$
Inverse Stein [28]	$\frac{1}{2} (\text{Tr}[\Sigma^{-1}\hat{\Sigma}] - p + \log \det(\Sigma\hat{\Sigma}^{-1}))$	$\mathbb{S}_{++}^p \times \mathbb{S}_{++}^p$
Symmetrized Stein / Jeffreys divergence [24]	$\frac{1}{2} (\text{Tr}[\Sigma\hat{\Sigma}^{-1} + \hat{\Sigma}\Sigma^{-1}] - 2p)$	$\mathbb{S}_{++}^p \times \mathbb{S}_{++}^p$
Quadratic / Squared Frobenius	$\text{Tr}[(\Sigma - \hat{\Sigma})^2]$	$\mathbb{S}_+^p \times \mathbb{S}_+^p$
Weighted quadratic	$\text{Tr}[(\Sigma - \hat{\Sigma})^2\hat{\Sigma}^{-1}]$	$\mathbb{S}_+^p \times \mathbb{S}_{++}^p$

TABLE 1. Popular divergence functions and their domains. We adopt the convention from convex analysis that each divergence evaluates to $+\infty$ outside of its domain.

We now informally state our key result, which applies, among others, to all divergence functions of Table 1.

Theorem 1 (Distributionally robust estimator (informal)). *If D is any divergence function from Table 1, the nominal covariance matrix $\hat{\Sigma}$ satisfies a regularity condition, and $\varepsilon > 0$ is not too large, then the distributionally robust estimator X^* exists, is unique, and can be computed efficiently via the following procedure.*

- (1) *Compute the eigenvalues and the eigenvectors of the nominal covariance matrix $\hat{\Sigma}$.*
- (2) *Construct the inverse shrinkage intensity γ^* by solving a univariate nonlinear equation that depends only on the spectrum of $\hat{\Sigma}$.*
- (3) *Shrink the eigenvalues of $\hat{\Sigma}$ by applying a nonlinear transformation that depends only on γ^* .*
- (4) *Construct X^* by combining the eigenvectors found in step (1) with the eigenvalues found in step (3).*

The estimator X^ constructed in this manner preserves the eigenvectors of $\hat{\Sigma}$, shrinks the eigenvalues of $\hat{\Sigma}$, and reduces the condition number of $\hat{\Sigma}$. Thus, it represents a nonlinear shrinkage estimator.*

Theorem 1 reveals that a wide range of nonlinear shrinkage estimators admit a robustness interpretation in the sense that they correspond to solutions of the distributionally robust estimation problem (3) for different divergence functions. This insight is of interest from a statistical point of view because it relates nonlinear shrinkage estimators to distributional ambiguity sets,

which can be used to derive new generalization bounds. Theorem 1 also implies that the distributionally robust estimation problem (3) can be solved efficiently by diagonalizing $\widehat{\Sigma}$ and solving a univariate nonlinear equation, both of which are computationally cheap.

3. DISTRIBUTIONALLY ROBUST COVARIANCE SHRINKAGE ESTIMATORS

This section formally introduces our distributionally robust estimation framework. Specifically, Section 3.1 details all technical assumptions needed throughout the paper, Section 3.2 formally states the main result, and Section 3.3 describes several desirable properties of the emerging distributionally robust estimators.

3.1. Assumptions

The uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is non-convex for some choices of the divergence function D . In these cases, we cannot use Sion's minimax theorem to establish strong duality between the primal and dual estimation problems (4) and (6), respectively. Instead, we will have to invoke a more nuanced minimax theorem. For now, we assume that such a minimax theorem is readily available.

Assumption 1 (Minimax property). *The minimum of the primal estimation problem (4) coincides with the maximum of the dual estimation problem (6).*

We will later see that Assumption 1 is satisfied for all divergence functions listed in Table 1. In addition, we require D to constitute a spectral divergence in the sense of the following assumption.

Assumption 2 (Spectral divergence). *The divergence function $D : \mathbb{S}_+^p \times \mathbb{S}_+^p \rightarrow \overline{\mathbb{R}}$ is non-negative, and satisfies the identity of indiscernibles, that is, for any $(X, Y) \in \text{dom}(D)$ we have $D(X, Y) = 0$ if and only if $X = Y$. In addition, D satisfies the following structural conditions.*

- (a) (Orthogonal equivariance) *For any $X, Y \in \mathbb{S}_+^p$ and $V \in \mathcal{O}_p$ we have that $D(X, Y) = D(VXV^\top, VYV^\top)$.*
- (b) (Spectrality) *There exists a function $d : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \overline{\mathbb{R}}$ such that*

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

and $d(a, b)$ is continuous¹ in a for every $b > 0$. In the following, we refer to d as the generator of D .

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

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

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

Assumptions 2(a) and 2(b) imply that if X and Y are simultaneously diagonalizable, then the divergence of X with respect to Y depends only on the spectra of X and Y and the generator d .

¹By convention, a continuous extended real-valued function must tend to ∞ when approaching the boundary of its domain.

Specifically, we have

$$D(X, Y) = D(V \text{Diag}(x) V^\top, V \text{Diag}(y) V^\top) = D(\text{Diag}(x), \text{Diag}(y)) = \sum_{i=1}^p d(x_i, y_i), \quad (7)$$

where the entries of the vectors x and y represent the eigenvalues and where the columns of the orthonormal matrix V represent the (common) eigenvectors of X and Y , respectively. Note that the last two equalities in (7) readily follow from 2(a) and 2(b). Assumption 2(b) further implies that if D is a spectral divergence on \mathbb{S}_+^p , then its generator d is a spectral divergence on \mathbb{R}_+ . Indeed, restricting x and y to multiples of the vector of all ones reveals via Assumption 2(b) that $\text{dom}(d) = \{(a, b) \in \mathbb{R}_+^2 : (aI_d, bI_d) \in \text{dom}(D)\}$ and that d inherits continuity, non-negativity and the identity of indiscernibles from D . Orthogonal equivariance, spectrality, and the rearrangement inequality are trivially satisfied in the one-dimensional case. Finally, we point out that Assumption 2(c) is reminiscent of the Hardy-Littlewood-Polyak rearrangement inequality [19], which asserts that $(x^\uparrow)^\top y^\downarrow \leq x^\top y \leq (x^\uparrow)^\top y^\uparrow$ for any vectors $x, y \in \mathbb{R}^p$.

Our results also require the following assumptions about the eigenvalues $\hat{x}_1, \dots, \hat{x}_p$ of the nominal covariance matrix $\hat{\Sigma}$ as well as about the radius ε of the uncertainty set $\mathcal{B}_\varepsilon(\hat{\Sigma})$.

Assumption 3 (Regularity of input parameters). *The following hold.*

- (a) For any $i = 1, \dots, p$ we have $(\hat{x}_i, \hat{x}_i) \in \text{dom}(d)$.
- (b) The radius ε of the uncertainty set satisfies $0 < \varepsilon < \bar{\varepsilon}$, where $\bar{\varepsilon} = \sum_{i=1}^p d(0, \hat{x}_i)$.

Together with Assumptions 2(a) and 2(b), Assumption 3(a) ensures that the nominal covariance matrix $\hat{\Sigma}$ is feasible in problem (P_{Mat}) . Indeed, inserting $X = Y = \hat{\Sigma}$ into (7) implies that $D(\hat{\Sigma}, \hat{\Sigma}) = 0$. This implies that $(\hat{\Sigma}, \hat{\Sigma}) \in \text{dom}(D)$ and, more importantly, that the feasible region of problem (P_{Mat}) is non-empty. This assumption is not entirely innocent because some divergence functions from Table 1 have domain $\mathbb{S}_{++}^p \times \mathbb{S}_{++}^p$. In all these cases, Assumption 3(a) requires that $\hat{\Sigma}$ has full rank and, if $\hat{\Sigma}$ is the sample covariance matrix, that the sample size n is at least as large as the dimension p . We emphasize that Assumption 3(a) does *not* generally imply that $n \geq p$. For instance, if $(0, 0) \in \text{dom}(d)$, then Assumption 3(a) holds even if $n < p$. This situation arises if D is the Wasserstein or the quadratic divergence. Conversely, Assumption 3(a) may fail to hold even when $n > p$. This happens, for example, if $(0, 0) \notin \text{dom}(d)$ and the nominal covariance matrix $\hat{\Sigma}$ is singular even though $n > p$. Assumption 3(b) ensures that the radius $\varepsilon > 0$ is small enough for the feasible region of the reformulated dual estimation problem (P_{Mat}) not to contain 0. Otherwise, problem (P_{Mat}) would trivially be solved by the nonsensical estimator $X^* = 0$.

Assumption 4 (Smoothness and convexity of the generator d). *For any $b > 0$, the function $d(\cdot, b)$ is twice continuously differentiable throughout \mathbb{R}_{++} and convex on the interval $[0, b]$.*

Assumption 4 implies that the domain of $d(\cdot, b)$ contains \mathbb{R}_{++} for every $b > 0$. Hence, $d(a, b)$ can evaluate to $+\infty$ only at $a = 0$, which means that the domain of $d(\cdot, b)$ is either \mathbb{R}_+ or \mathbb{R}_{++} . We emphasize that the convexity of $d(\cdot, b)$ on the interval $[0, b]$ does *not* imply that problem (P_{Mat}) is convex. However, we will see below that this restricted convexity assumption helps us to reduce problem (P_{Mat}) to a convex program.

3.2. Construction of the Distributionally Robust Estimator

We need the following notation to restate Theorem 1 rigorously. We denote the i -th smallest eigenvalue of a symmetric matrix $S \in \mathbb{S}^p$ by $\lambda_i(S)$, and we use $\lambda(S) = (\lambda_1(S), \dots, \lambda_p(S))$ as a shorthand for the spectrum of S . We also reserve the symbols $\hat{x}_i = \lambda_i(\hat{\Sigma})$ and \hat{v}_i for the non-negative eigenvalues and the corresponding orthonormal eigenvectors of the nominal covariance matrix $\hat{\Sigma}$. In addition, we use $\hat{x} = \lambda(\hat{\Sigma})$ and $\hat{V} = (\hat{v}_1, \dots, \hat{v}_p)$ to denote the nominal spectrum and the orthogonal matrix of the nominal eigenvectors, respectively. The nominal covariance matrix thus admits the spectral decomposition $\hat{\Sigma} = \hat{V} \text{Diag}(\hat{x}) \hat{V}^\top$. We also define the auxiliary function $s : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ corresponding to a divergence function with generator d via

$$s(\gamma, b) = \begin{cases} \text{the unique solution } a^* \geq 0 \text{ of the equation } 0 = 2a^* + \gamma \frac{\partial d}{\partial a}(a^*, b) & \text{if } b > 0 \text{ and } \gamma > 0, \\ 0 & \text{if } b = 0 \text{ or } \gamma = 0. \end{cases} \quad (8)$$

In the remainder of the paper, we refer to s as the *eigenvalue map*. We will see below that it is well-defined under Assumption 4, which implies that the nonlinear equation in (8) has a unique solution whenever $b > 0$. We will also prove that $s(\gamma, b) \leq b$ for every $\gamma, b \geq 0$, which means that it can be viewed as a shrinkage transformation that maps any input eigenvalue $b \geq 0$ to a smaller output eigenvalue $s(\gamma, b)$ for every fixed γ . Given these conventions, we are now ready to restate Theorem 1 formally.

Theorem 1 (Distributionally robust estimator (formal)). *If Assumptions 1–4 hold, then the distributionally robust estimator X^* exists and is unique. If, additionally, γ^* is the unique positive root of the equation*

$$\sum_{i=1}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon = 0,$$

then the distributionally robust estimator admits the spectral decomposition $X^ = \hat{V} \text{Diag}(x^*) \hat{V}^\top$ with eigenvalues $x_i^* = s(\gamma^*, \hat{x}_i)$, $i = 1, \dots, p$, where $0 < x_i^* < \hat{x}_i$ whenever $\hat{x}_i > 0$ and $x_i^* = 0$ whenever $\hat{x}_i = 0$.*

Theorem 1 provides a quasi-closed form expression for the optimal covariance estimator X^* that solves the robust estimation problem (4) as well as its dual reformulation (P_{Mat}) . In particular, it shows that X^* has the same eigenvectors as $\hat{\Sigma}$ and that all positive eigenvalues of X^* can be computed by solving a nonlinear equation parametrized by γ^* . Remarkably, this nonlinear equation admits a closed-form solution for all divergences listed in Table 1. In addition, we will see that γ^* can be computed efficiently by bisection. All of this implies that the complexity of computing X^* is largely determined by the complexity of diagonalizing $\hat{\Sigma}$. In addition, we will see that $x_i^* = s(\gamma^*, \hat{x}_i)$ decreases with γ^* . Thus, X^* and γ^* are naturally interpreted as a nonlinear shrinkage estimator and inverse shrinkage intensity, respectively.

We now outline the high-level structure of the proof of Theorem 1; see Figure 1 for a visualization. The proof is divided into three steps that give rise to three propositions. Proposition 1 below first shows that there is a one-to-one relationship between the minimizers of the robust estimation problem (4) and problem (P_{Mat}) .

Proposition 1 (Dual characterization of X^*). *If Assumption 1 holds, then the primal and dual robust estimation problems (4) and (6) are equivalent to problem (P_{Mat}) in the following sense.*

- (i) *If Σ^* solves (P_{Mat}) , then $X^* = \Sigma^*$ solves (4), and Σ^* solves (6).*

(ii) If X^* solves (4) and Σ^* solves (6), then X^* coincides with Σ^* and solves (P_{Mat}) .

The proof of Proposition 1 follows immediately from the discussion in Section 2 and is thus omitted. Next, we show that problem (P_{Mat}) , which optimizes over all matrices in the positive semidefinite cone \mathbb{S}_+^p , is equivalent to problem (P_{Vec}) below, which optimizes over all vectors in the non-negative orthant \mathbb{R}_+^p :

$$\min_{x \in \mathbb{R}_+^p} \left\{ \|x\|_2^2 : \sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon \right\}. \quad (P_{\text{Vec}})$$

We henceforth use x^* to denote the unique minimizer of problem (P_{Vec}) if it exists.

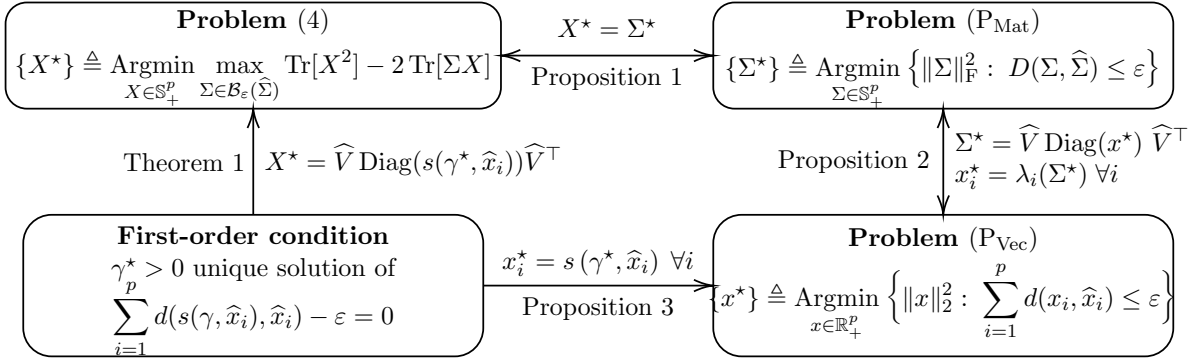


FIGURE 1. Structure of the proof of Theorem 1. An arc indicates that the solution to the problem at the arc's tail can be used to construct a solution for the problem at the arc's head.

Proposition 2 (Equivalence of (P_{Mat}) and (P_{Vec})). *If Assumption 2 holds, then problem (P_{Mat}) is equivalent to problem (P_{Vec}) in the following sense.*

- (i) Problem (P_{Mat}) is feasible if and only if problem (P_{Vec}) is feasible.
- (ii) If x^* solves (P_{Vec}) , then $\widehat{V} \text{Diag}(x^*) \widehat{V}^\top$ solves (P_{Mat}) .
- (iii) If Σ^* solves (P_{Mat}) , then $\lambda(\Sigma^*)$ solves (P_{Vec}) .
- (iv) (P_{Mat}) and (P_{Vec}) share the same optimal value.

In the third and last step, we solve problem (P_{Vec}) in quasi-analytical form. To this end, we denote the Lagrange multiplier associated with the divergence constraint $\sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon$ by γ^* . The following proposition characterizes the unique solution of problem (P_{Vec}) through an explicit function of γ^* and shows that (P_{Vec}) can be computed by solving a single nonlinear equation.

Proposition 3 (Solution of (P_{Vec})). *If Assumptions 2, 3 and 4 hold, then problem (P_{Vec}) admits a unique optimal solution x^* with components $x_i^* = s(\gamma^*, \hat{x}_i)$, $i = 1, \dots, p$, where γ^* is the unique positive root of the equation $\sum_{i=1}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon = 0$. We also have $0 < x_i^* < \hat{x}_i$ whenever $\hat{x}_i > 0$ and $x_i^* = 0$ whenever $\hat{x}_i = 0$.*

In summary, Proposition 3 provides a simple characterization of γ^* and shows how one can use γ^* to construct a unique solution x^* for problem (P_{Vec}) . Proposition 2 reveals how x^* can be used to construct a unique solution Σ^* for problem (P_{Mat}) , and Proposition 1 guarantees that Σ^* is uniquely optimal in the robust estimation problem (4). Taken together, Propositions 1, 2 and 3 therefore prove Theorem 1.

3.3. Properties of the Distributionally Robust Estimator

We now highlight several desirable characteristics of the distributionally robust covariance estimator X^* .

3.3.1. Efficient Computation

We have seen that X^* can be constructed from x^* , which can be constructed from γ^* . In addition, we have seen that the Lagrange multiplier γ^* is the unique positive root of the equation $F(\gamma) = 0$, where the function $F : \mathbb{R}_+ \rightarrow \mathbb{R}$ is defined through $F(\gamma) = \sum_{i=1}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon$. The following proposition suggests that this root-finding problem can be solved highly efficiently by bisection or Newton's method.

Proposition 4 (Structural properties of F). *If Assumptions 2, 3 and 4 hold, then the function F is differentiable and strictly decreasing on \mathbb{R}_{++} . In addition, we have $\lim_{\gamma \rightarrow 0} F(\gamma) > 0$ and $\lim_{\gamma \rightarrow \infty} F(\gamma) < 0$.*

Suppose now that we have access to an *a priori* upper bound $\bar{\gamma} > 0$ on the Lagrange multiplier γ^* . Note that $\bar{\gamma}$ is guaranteed to exist under the assumptions of the proposition. Section 4 shows that $\bar{\gamma}$ can be constructed explicitly for several popular divergence functions. The structural properties of F established in Proposition 4 allow us to estimate the number of function evaluations needed to compute γ^* . For example, γ^* can be computed via bisection to within an absolute error of $\delta > 0$ using $\log_2(\bar{\gamma}/\delta)$ function evaluations. Under additional mild conditions, γ^* can also be computed via Newton's method to within an absolute error of $\delta > 0$ using merely $O(\log_2 \log_2(\bar{\gamma}/\delta))$ function and derivative evaluations [11, Theorem 2.4.3].

3.3.2. Shrinkage Properties

Proposition 3 asserts that if Assumptions 2, 3 and 4 hold, then the optimal solution x^* of problem (P_{Vec}) is unique and can thus be seen as a function $x^*(\varepsilon)$ of the radius $\varepsilon \in (0, \bar{\varepsilon})$ of the uncertainty set, where $\bar{\varepsilon}$ is defined as in Assumption 3(b). In fact, $x^*(\varepsilon)$ can naturally be extended to a function on $[0, \bar{\varepsilon}]$. As d satisfies the identity of indiscernibles, we can define $x^*(0) = \hat{x}$ as the unique solution of problem (P_{Vec}) for $\varepsilon = 0$. In addition, we may define $x^*(\bar{\varepsilon}) = 0$. One can then show that each component of $x^*(\varepsilon)$ monotonically decreases to 0 on $[0, \bar{\varepsilon}]$. By Theorem 1, the distributionally robust estimator $X^* = \widehat{V} \text{Diag}(x^*) \widehat{V}^\top$ inherits the eigenbasis from the nominal covariance matrix $\widehat{\Sigma}$. Hence, X^* and $\widehat{\Sigma}$ commute, and X^* is rotation equivariant. In summary, these insights imply that X^* essentially shrinks the eigenvalues of $\widehat{\Sigma}$ towards zero.

Proposition 5 (Shrinkage estimator). *If Assumptions 2, 3 and 4 hold, then $x_i^*(\varepsilon)$ is non-increasing on $[0, \bar{\varepsilon}]$ and satisfies $\lim_{\varepsilon \uparrow \bar{\varepsilon}} x_i^*(\varepsilon) = 0$ for every $i = 1, \dots, p$. If additionally Assumption 1 holds, then X^* constitutes a shrinkage estimator, that is, it has the same eigenvectors as $\widehat{\Sigma}$ and satisfies $0 \preceq X^* \preceq \widehat{\Sigma}$.*

Proposition 5 asserts that the eigenvalues of X^* are bounded above by the corresponding nominal eigenvalues. This shrinkage property persists across a remarkably broad class of estimators. The shrinkage effects of robustification were first discovered in a distributionally robust *inverse* covariance estimation problem with a Wasserstein ambiguity set [43]. The results presented here are significantly more general. Indeed, they reveal that a broad class of divergence functions gives rise to diverse shrinkage estimators.

3.3.3. Improvement of the Condition Number

The condition number $\kappa(X)$ of a positive definite matrix $X \in \mathbb{S}_{++}^p$ is defined as the ratio of its largest to its smallest eigenvalue. It is well known that unless $n \gg p$, the sample covariance matrix $\widehat{\Sigma}_{\text{SA}}$ tends to be ill-conditioned, that is, $\kappa(\widehat{\Sigma}_{\text{SA}}) \gg 1$ [63]. Therefore, most shrinkage estimators are designed to improve the condition number of an ill-conditioned baseline estimator $\widehat{\Sigma}$. Below we will show that the distributionally robust estimator X^* is also guaranteed to improve the condition number of $\widehat{\Sigma}$ whenever the generator d of the divergence D satisfies a second-order differential inequality.

Assumption 5 (Differential inequality). *The generator d of the divergence function D is twice continuously differentiable on \mathbb{R}_{++}^2 and satisfies the following differential inequality for all $a, b \in \mathbb{R}_{++}$ with $a < b$.*

$$a \frac{\partial^2}{\partial a^2} d(a, b) + b \frac{\partial^2}{\partial a \partial b} d(a, b) \geq \frac{\partial}{\partial a} d(a, b)$$

Assumption 5 may be difficult to check. In Theorem 2 below, we will show, however, that it is satisfied by all divergence functions of Table 1. We can now prove that robustification improves the condition number.

Proposition 6 (Improved condition number). *If Assumptions 1–5 hold and $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, then $\kappa(X^*) \leq \kappa(\widehat{\Sigma})$.*

The proof of Proposition 6 exploits a generalized monotonicity property of the eigenvalue map $s(\gamma, b)$.

Lemma 1 (Generalized monotonicity property of the eigenvalue map s). *If Assumptions 2, 4 and 5 hold, then we have $s(\gamma, b_2)/s(\gamma, b_1) \leq b_2/b_1$ for all $\gamma > 0$ and $b_1, b_2 \in \mathbb{R}_{++}$ with $b_2 \geq b_1$.*

Recall from Theorem 1 that $x_i^* = s(\gamma^*, \hat{x}_i)$ for all $i = 1, \dots, p$ and that $\gamma^* > 0$. Therefore, Proposition 6 follows immediately from Lemma 1.

3.3.4. Statistical Guarantees

We finally show that the distributionally robust estimator is consistent and enjoys a finite-sample performance guarantee. To this end, we make the dependence on n explicit, that is, we let X_n^* be the unique solution of (4), where the nominal estimator is any covariance estimator $\widehat{\Sigma}_n$ constructed from n i.i.d. training samples, and where the radius is set to a non-negative number ε_n that may depend on $n \in \mathbb{N}$. We say a covariance estimator is strongly consistent if it converges almost surely to Σ_0 for a fixed p as n tends to infinity.

Proposition 7 (Consistency). *Suppose that Assumptions 1–4 hold and that d is continuous on $\mathbb{R}_+ \times \mathbb{R}_{++}$. If $\widehat{\Sigma}_n$ is a strongly consistent estimator and ε_n converges to 0 as n grows, then X_n^* is strongly consistent.*

Proposition 7 is intuitive because the uncertainty set is assumed to shrink with n , and the nominal covariance matrix at its center is assumed to be consistent. As the uncertainty set is defined as a generic divergence ball, however, the proof is perhaps surprisingly tedious. The standard example of a consistent nominal covariance estimator $\widehat{\Sigma}_n$ is the sample covariance matrix. Note that Proposition 7 analyzes the asymptotics of X_n^* as n tends to infinity for a fixed p , which is referred to as the *low-dimensional regime* in statistics.

Next, we establish finite-sample performance guarantees, that is, we show that the uncertainty set of radius $\varepsilon_n \propto n^{-\frac{1}{2}}$ around the sample covariance matrix constitutes a confidence region for Σ_0 . In the following we say that the probability distribution \mathbb{P} is sub-Gaussian if there exists a variance proxy $\sigma^2 \geq 0$ with $\mathbb{E}_{\mathbb{P}}[\exp(z^\top \xi)] \leq \exp(\frac{1}{2}\sigma^2\|z\|_2^2)$ for every $z \in \mathbb{R}^p$. As both sides of this inequality are differentiable and coincide at $z = 0$, one can show that any sub-Gaussian distribution \mathbb{P} must have mean 0.

Proposition 8 (Finite-sample performance guarantee). *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 corresponding to n i.i.d. samples from \mathbb{P} . For any divergence function D from Table 1 there exist functions $n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ and $\varepsilon_{\min}(p, n, \eta) = \mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}})$, which may depend on \mathbb{P} only through the variance proxy σ^2 and the smallest eigenvalue $\lambda_1(\Sigma_0)$ of Σ_0 , such that $\mathbb{P}^n[\Sigma_0 \in \mathcal{B}_\varepsilon(\widehat{\Sigma}_n)] \geq 1 - \eta$ for every $n \geq n_{\min}(p, \eta)$ and $\varepsilon \geq \varepsilon_{\min}(p, n, \eta)$.*

Proposition 8 implies that if $n \geq n_{\min}(p, \eta)$ and $\varepsilon \geq \varepsilon_{\min}(p, n, \eta)$, then the optimal value of the robust covariance estimation problem (4) provides a $(1 - \eta)$ -upper confidence bound on the actual estimation error with respect to the true covariance matrix Σ_0 . Explicit formulas for $n_{\min}(p, \eta)$ and $\varepsilon_{\min}(p, n, \eta)$ tailored to different divergence functions can be found in the proof of Proposition 8 in the appendix. The finite-sample guarantee of Proposition 8 directly yields an asymptotic guarantee in a high-dimensional regime where p grows with n . Specifically, it implies that the population covariance Σ_0 remains within the uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma}_n)$ with constant confidence $1 - \eta$ as the dimension p scales like $n^{1/3}$. This stands in contrast to standard high-dimensional performance guarantees, which permit the dimension to grow linearly with n .

4. A ZOO OF NEW COVARIANCE SHRINKAGE ESTIMATORS

In this section, we first show that the assumptions of Theorem 1 are satisfied by a broad spectrum of divergence functions commonly used in statistics, information theory, and machine learning. Next, we explicitly construct the shrinkage estimators corresponding to three popular divergence functions.

Theorem 2 (Validation of assumptions). *All divergences in Table 1 satisfy Assumptions 1, 2, 4 and 5.*

We emphasize that the uncertainty sets corresponding to the Fisher-Rao and inverse Stein divergences fail to be convex, in which case one cannot use standard minimax results to prove Assumption 1. However, perhaps surprisingly, in Appendix C.2, we show that the uncertainty sets corresponding to these divergences are geodesically convex with respect to a particular Riemannian geometry on the space of positive definite matrices. This in turn allows us to apply a Riemannian minimax theorem (see Theorem 3 in Appendix C.3) and prove the desired minimax property even for robust estimation problems based on the Fisher-Rao and inverse Stein divergences.

To showcase the richness of our framework, we now focus on three popular divergence functions and analyze the corresponding robust covariance estimators. Specifically, we will derive the optimal solutions of problem (P_{Vec}) in quasi-closed form for the Kullback-Leibler, Wasserstein, and Fisher-Rao divergences. In doing so, we develop a general recipe for the other divergence functions listed in Table 1.

4.1. The Kullback-Leibler Covariance Shrinkage Estimator

Table 1 defines the Kullback-Leibler (KL) divergence between two matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}_{++}^p$ as

$$D_{\text{KL}}(\Sigma_1, \Sigma_2) = \frac{1}{2} \left(\text{Tr}[\Sigma_2^{-1}\Sigma_1] - p + \log \det(\Sigma_2\Sigma_1^{-1}) \right).$$

This KL divergence between matrices is intimately related to the KL divergence between distributions.

Definition 1 (KL divergence). *If \mathbb{P}_1 and \mathbb{P}_2 are two probability distributions on \mathbb{R}^p , and \mathbb{P}_1 is absolutely continuous with respect to \mathbb{P}_2 , then the KL divergence from \mathbb{P}_1 to \mathbb{P}_2 is $\text{KL}(\mathbb{P}_1 \parallel \mathbb{P}_2) = \int_{\mathbb{R}^p} \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_2}(x)\right) d\mathbb{P}_2(x)$.*

The following lemma shows that the KL divergence between two non-degenerate zero-mean Gaussian distributions coincides with the KL divergence between their positive definite covariance matrices.

Lemma 2 (KL divergence between Gaussian distributions [28]). *The KL divergence from $\mathbb{P}_1 = \mathcal{N}(0, \Sigma_1)$ to $\mathbb{P}_2 = \mathcal{N}(0, \Sigma_2)$ with $\Sigma_1, \Sigma_2 \in \mathbb{S}_{++}^p$ is given by $\text{KL}(\mathbb{P}_1 \parallel \mathbb{P}_2) = D_{\text{KL}}(\Sigma_1, \Sigma_2)$.*

Lemma 2 justifies our terminology of referring to D_{KL} as the KL divergence and suggests that D_{KL} inherits many properties of the KL divergence between distributions. For example, it is easy to verify that D_{KL} satisfies the identity of indiscernibles but fails to be symmetric. Indeed, for any $\Sigma \in \mathbb{S}_{++}^p$ we have $D_{\text{KL}}(\Sigma, 2\Sigma) = \frac{p}{2}(1 - \log(2)) \approx 0.15p$, whereas $D_{\text{KL}}(2\Sigma, \Sigma) = \frac{p}{2}(\log(2) - \frac{1}{2}) \approx 0.1p$. An elementary calculation further reveals that the generator d corresponding to the KL divergence D_{KL} can be expressed as

$$d(a, b) = \frac{1}{2} \left(\frac{a}{b} - 1 - \log \left(\frac{a}{b} \right) \right).$$

The following corollary of Theorem 1 characterizes the eigenvalue map and the inverse shrinkage intensity corresponding to the KL divergence, which determines the KL covariance shrinkage estimator.

Corollary 1 (KL covariance shrinkage estimator). *If D is the KL divergence, $\hat{\Sigma} \in \mathbb{S}_{++}^p$ and $\varepsilon > 0$, then problem (4) is uniquely solved by the KL covariance shrinkage estimator $X^* = \hat{V} \text{Diag}(x^*) \hat{V}^\top$ with shrunk eigenvalues $x_i^* = s(\gamma^*, \hat{x}_i)$, $i = 1, \dots, p$. The underlying eigenvalue map is given by*

$$s(\gamma, b) = \frac{-\gamma + \sqrt{\gamma^2 + 16b^2\gamma}}{8b}, \quad (9a)$$

and the inverse shrinkage intensity $\gamma^* \in (0, \gamma_{\text{KL}}]$ is the unique positive solution of the nonlinear equation

$$2\varepsilon + p + \sum_{i=1}^p \left[-\frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} + \log \frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} \right] = 0, \quad (9b)$$

where

$$\gamma_{\text{KL}} = \frac{4\hat{x}_p^2 \exp(-4\varepsilon/p)}{1 - \exp(-2\varepsilon/p)} > 0.$$

4.2. The Wasserstein Covariance Shrinkage Estimator

Table 1 defines the Wasserstein divergence between two matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}_{++}^p$ as

$$D_{\text{W}}(\Sigma_1, \Sigma_2) = \text{Tr}[\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{\frac{1}{2}}].$$

In the following, we will show that the Wasserstein distance between matrices is closely related to the squared 2-Wasserstein distance between distributions, where the transportation cost is defined via the 2-norm.

Definition 2 (Wasserstein distance). *The 2-Wasserstein distance between two probability distributions \mathbb{P}_1 and \mathbb{P}_2 on \mathbb{R}^p with finite second moments is defined as*

$$W_2(\mathbb{P}_1, \mathbb{P}_2) = \left(\inf_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|x_1 - x_2\|_2^2 d\pi(x_1, x_2) \right)^{\frac{1}{2}},$$

where $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of probability distributions on $\mathbb{R}^p \times \mathbb{R}^p$ with marginals \mathbb{P}_1 and \mathbb{P}_2 , respectively.

One can show that Wasserstein distance W_2 is a metric on the space of probability distributions with finite second moments [65, § 6]. However, the *squared* Wasserstein distance W_2^2 is only a divergence as it fails to satisfy the triangle inequality. The following lemma shows that the squared 2-Wasserstein distance between two zero-mean Gaussian distributions matches the Wasserstein divergence between their covariance matrices.

Lemma 3 (Squared Wasserstein distance between Gaussian distributions [18]). *The squared 2-Wasserstein distance between $\mathbb{P}_1 = \mathcal{N}(0, \Sigma_1)$ and $\mathbb{P}_2 = \mathcal{N}(0, \Sigma_2)$ evaluates to $W_2(\mathbb{P}_1, \mathbb{P}_2)^2 = D_W(\Sigma_1, \Sigma_2)$.*

Lemma 3 justifies our terminology of referring to D_W as the Wasserstein divergence and suggests that D_W inherits many properties from the Wasserstein distance between distributions. Note that D_W remains well-defined even if Σ_1 or Σ_2 are rank-deficient. The generator d of the Wasserstein divergence D_W is given by

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

The following corollary of Theorem 1 characterizes the eigenvalue map and inverse shrinkage intensity corresponding to the Wasserstein divergence, which determines the Wasserstein covariance shrinkage estimator.

Corollary 2 (Wasserstein covariance shrinkage estimator). *If D is the Wasserstein divergence, $\widehat{\Sigma} \in \mathbb{S}_+^p$ and $\varepsilon \in (0, \text{Tr}[\widehat{\Sigma}])$, then problem (4) is uniquely solved by the Wasserstein covariance shrinkage estimator $X^* = \widehat{V} \text{Diag}(x^*) \widehat{V}^\top$ with eigenvalues $x_i^* = s(\gamma^*, \hat{x}_i)$, $i = 1, \dots, p$. The underlying eigenvalue map is given by*

$$s(\gamma, b) = \left(\left\{ \frac{\gamma}{4} \left(\sqrt{b} + \sqrt{b + \frac{2}{27}\gamma} \right) \right\}^{\frac{1}{3}} - \frac{\gamma}{6} \left\{ \frac{\gamma}{4} \left(\sqrt{b} + \sqrt{b + \frac{2}{27}\gamma} \right) \right\}^{-\frac{1}{3}} \right)^2 \quad (10a)$$

and the inverse shrinkage intensity $\gamma^* \in (0, \gamma_W]$ is the unique positive solution of the nonlinear equation

$$\varepsilon - \sum_{i=1}^p \left(\sqrt{\hat{x}_i} - \sqrt{s(\gamma^*, \hat{x}_i)} \right)^2 = 0, \quad (10b)$$

where $\gamma_W = 2\sqrt{p\hat{x}_p^3/\varepsilon} > 0$.

The requirement that ε be strictly smaller than $\text{Tr}[\widehat{\Sigma}]$ is equivalent to Assumption 3(b). It is needed to prevent problem (P_{Vec}) from admitting the trivial solution $x^* = 0$. To see this, note that the condition $\varepsilon \geq \text{Tr}[\widehat{\Sigma}]$ is equivalent to $\sum_{i=1}^p d(0, \hat{x}_i) \leq \varepsilon$, which in turn implies that 0

is feasible and even optimal in (P_{Vec}) . In this case, the trivial (and essentially nonsensical) estimator $X^* = 0$ would be optimal in problem (4).

4.3. The Fisher-Rao Covariance Shrinkage Estimator

Table 1 defines the Fisher-Rao divergence between two matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}_{++}^p$ as

$$D_{\text{FR}}(\Sigma_1, \Sigma_2) = \|\log(\Sigma_2^{-\frac{1}{2}} \Sigma_1 \Sigma_2^{-\frac{1}{2}})\|_{\text{F}}^2.$$

The Fisher-Rao divergence can be interpreted as the Fisher-Rao distance on a particular statistical manifold.

Definition 3 (Fisher-Rao distance). *Consider a family of probability density functions $\{f_\theta(\xi)\}_{\theta \in \Theta}$ whose parameter θ ranges over a Riemannian manifold Θ with metric*

$$I_\theta = \int_{\Xi} f_\theta(\xi) \nabla_\theta \log(f_\theta(\xi)) \nabla_\theta \log(f_\theta(\xi))^\top d\xi.$$

The geodesic distance $\text{FR}(\theta_1, \theta_2)$ on Θ induced by this metric is referred to as the Fisher-Rao distance.

Note that I_θ represents the Fisher information matrix corresponding to the parameter θ . Next, we show that the squared Fisher-Rao distance between two non-degenerate zero-mean Gaussian probability density functions is proportional to the Fisher-Rao divergence between their positive definite covariance matrices.

Lemma 4 (Fisher-Rao distance between positive definite covariance matrices [3]). *Let $\{f_\theta(\xi)\}_{\theta \in \Theta}$ be the family of all non-degenerate zero-mean Gaussian probability density functions encoded by their covariance matrices $\theta = \Sigma$, which range over the Riemannian manifold $\Theta = \mathbb{S}_{++}^p$ equipped with the Fisher-Rao distance. If $\theta_1 = \Sigma_1$ and $\theta_2 = \Sigma_2$ belong to \mathbb{S}_{++}^p , then $\text{FR}(\theta_1, \theta_2)^2 = \frac{1}{2} D_{\text{FR}}(\Sigma_1, \Sigma_2)$.*

Lemma 4 justifies our terminology of referring to D_{FR} as the Fisher-Rao divergence. As D_{FR} is proportional to the *squared* Fisher-Rao distance FR^2 , it fails to satisfy the triangle inequality and is indeed only a divergence. Moreover, Example 1 in Appendix C.1 reveals that D_{FR} is neither convex nor quasi-convex. However, it is geodesically convex. The generator d corresponding to D_{FR} can be expressed as

$$d(a, b) = (\log(a/b))^2.$$

The following corollary of Theorem 1 characterizes the eigenvalue map and inverse shrinkage intensity corresponding to the Fisher-Rao divergence, which characterizes the Fisher-Rao covariance estimator.

Corollary 3 (Fisher-Rao covariance shrinkage estimator). *If D is the Fisher-Rao divergence, $\widehat{\Sigma} \in \mathbb{S}_{++}^p$ and $\varepsilon > 0$, then problem (4) is uniquely solved by the Fisher-Rao covariance shrinkage estimator $X^* = \widehat{V} \text{Diag}(x^*) \widehat{V}^\top$ with eigenvalues $x_i^* = s(\gamma^*, \hat{x}_i)$, $i = 1, \dots, p$. The underlying eigenvalue map is given by*

$$s(\gamma, b) = b \exp\left(-\frac{1}{2} W_0(2b^2/\gamma)\right), \quad (11a)$$

and W_0 denotes the principal branch of the Lambert-W function. In addition, the inverse shrinkage intensity $\gamma^* \in (0, \gamma_{\text{FR}}]$ with $\gamma_{\text{FR}} = \|\widehat{\Sigma}\|_{\text{F}}^2/\sqrt{\varepsilon} > 0$ is the unique positive solution of the

nonlinear equation

$$\sum_{i=1}^p W_0^2(2\hat{x}_i^2/\gamma) = 4\varepsilon. \quad (11b)$$

4.4. Other Covariance Shrinkage Estimators

Theorem 2 ensures that all divergence functions from Table 1 satisfy Assumptions 1, 2, 4 and 5 and thus induce via Theorem 1 a distributionally robust covariance shrinkage estimator. The generators and eigenvalue maps corresponding to all these divergences can be derived by using similar techniques as in Corollaries 1, 2, and 3. Details are omitted for brevity. All generators and eigenvalue maps are provided in Table 2.

Divergence	$d(a, b)$	$\text{dom}(d)$	$s(\gamma, b)$ for $b > 0$
Kullback-Leibler/Stein	$\frac{1}{2} \left(\frac{a}{b} - 1 - \log \frac{a}{b} \right)$	$\mathbb{R}_{++} \times \mathbb{R}_{++}$	$\frac{-\gamma + \sqrt{\gamma^2 + 16b^2\gamma}}{8b}$
Wasserstein	$a + b - 2\sqrt{ab}$	$\mathbb{R}_+ \times \mathbb{R}_+$	$\left(\left(\frac{\gamma}{4} \left(\sqrt{b} + \sqrt{b + \frac{2}{27}\gamma} \right) \right)^{\frac{1}{3}} - \frac{\gamma}{6} \left(\frac{\gamma}{4} \left(\sqrt{b} + \sqrt{b + \frac{2}{27}\gamma} \right) \right)^{-\frac{1}{3}} \right)^2$
Fisher-Rao	$(\log \frac{a}{b})^2$	$\mathbb{R}_{++} \times \mathbb{R}_{++}$	$b \exp(-\frac{1}{2} W_0(2b^2/\gamma))$
Inverse Stein	$\frac{1}{2} \left(\frac{b}{a} - 1 - \log \frac{b}{a} \right)$	$\mathbb{R}_{++} \times \mathbb{R}_{++}$	$\frac{3^{1/3} (\sqrt{3\gamma^2(27b^2+\gamma)} + 9\gamma b)^{2/3} - 3^{2/3}\gamma}{6(\sqrt{3\gamma^2(27b^2+\gamma)} + 9\gamma b)^{1/3}}$
Symmetrized Stein/ Jeffreys divergence	$\frac{1}{2} \left(\frac{b}{a} + \frac{a}{b} - 2 \right)$	$\mathbb{R}_{++} \times \mathbb{R}_{++}$	$\frac{1}{12} \left(\frac{\gamma^2}{b(216\gamma b^4 + 12\sqrt{3(108(\gamma^2 b^8 - 3(\gamma b)^4) - \gamma^3)})} + \frac{(216\gamma b^4 + 12\sqrt{3(108(\gamma^2 b^8 - 3(\gamma b)^4) - \gamma^3)})}{b} - \frac{\gamma}{b} \right)$
Quadratic/ Squared Frobenius	$(a - b)^2$	$\mathbb{R}_+ \times \mathbb{R}_+$	$\frac{b}{\gamma + b}$
Weighted quadratic	$\frac{(a-b)^2}{b}$	$\mathbb{R}_+ \times \mathbb{R}_{++}$	$\frac{\gamma b}{\gamma + b}$

TABLE 2. Generators and eigenvalue maps of the divergences from Table 1.

5. NUMERICAL EXPERIMENTS

We now compare our distributionally robust covariance estimators against the linear shrinkage estimator with shrinkage target $\frac{1}{n} \text{Tr}[\hat{\Sigma}]I_n$ [32] as well as a state-of-the-art nonlinear shrinkage estimator proposed by Ledoit and Wolf [35], henceforth referred to as the *NLLW* estimator. The performance of the linear shrinkage estimator depends on the choice of the mixing parameter $\alpha \in [0, 1]$, which we calibrate via cross-validation.

We first study the dependence of our estimators on the radius ε of the uncertainty set, and we numerically validate the asymptotic consistency and finite-sample guarantees of Propositions 7 and 8, respectively. Using synthetic data, we then assess the Frobenius risk of our estimators as a function of the sample size. Using real data, we further test the performance of minimum variance portfolios constructed from our estimators. In addition, we illustrate the use of covariance estimators in the context of linear and quadratic discriminant analysis. The code for all experiments as well as an implementation of our methods can be found on GitHub.²

5.1. Dependence on the Radius of the Uncertainty Set

We first study the decay of the eigenvalues and the condition number of the Kullback-Leibler, Wasserstein, and Fisher-Rao covariance shrinkage estimators with the radius ε of the uncertainty set. To this end, we set $p = 3$ and consider a nominal covariance matrix with eigenvalue spectrum

²https://github.com/yvesrychener/covariance_DR0

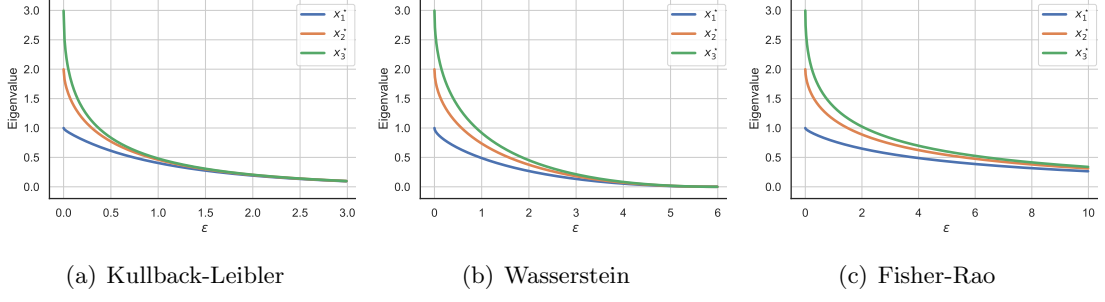


FIGURE 2. Eigenvalues of three different distributionally robust covariance estimators as a function of the radius ε for $\lambda(\hat{\Sigma}) = [1, 2, 3]$.

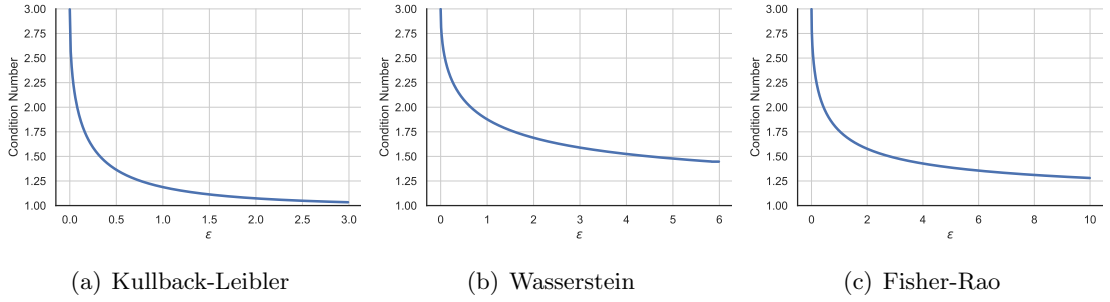
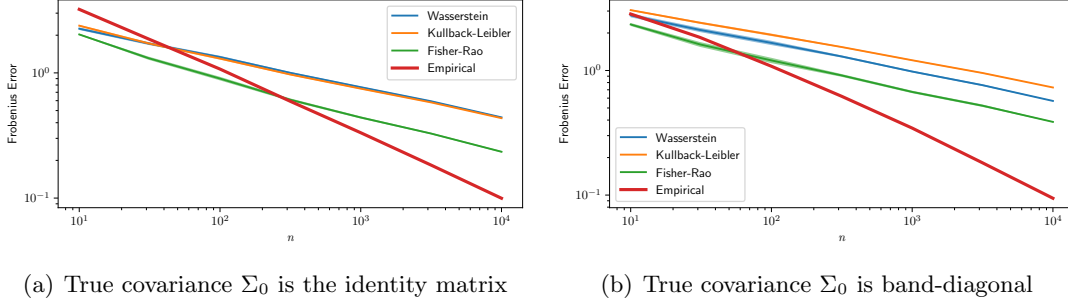


FIGURE 3. Condition number of three different distributionally robust covariance estimators as a function of the radius ε for $\lambda(\hat{\Sigma}) = [1, 2, 3]$.

$\lambda(\hat{\Sigma}) = [1, 2, 3]$. Figure 2 visualizes the eigenvalues of X^* as a function of ε . In agreement with Proposition 5, we observe that X^* shrinks the eigenvalues of the underlying nominal estimator $\hat{\Sigma}$ towards 0 as ε grows. Recall from Assumption 3(b) and the subsequent discussion that $X^* = 0$ whenever $\varepsilon \geq \sum_{i=1}^p d(0, \hat{x}_i)$. As the generator of the Wasserstein divergence satisfies $d(0, b) = b$, the eigenvalues of the Wasserstein covariance shrinkage estimator thus vanish for any $\varepsilon \geq \text{Tr}[\hat{\Sigma}]$. In contrast, the eigenvalues of the Kullback-Leibler and Fisher-Rao covariance shrinkage estimators remain strictly positive for all ε . We further observe that, for small values of ε , the Wasserstein and Fisher-Rao covariance shrinkage estimators primarily shrink the large eigenvalues of $\hat{\Sigma}$ and keep the small ones constant. Figure 3 visualizes the condition number $\kappa(X^*)$ as a function of ε . As predicted by Proposition 6, $\kappa(X^*)$ is at most as large as $\kappa(\hat{\Sigma})$. Note also that $\kappa(X^*)$ is undefined for $\varepsilon \geq \sum_{i=1}^p d(0, \hat{x}_i)$. Figure 3 indicates that the condition number of X^* decreases monotonically as ε tends to $\sum_{i=1}^p d(0, \hat{x}_i)$.

5.2. Consistency and Finite-Sample Performance

To validate both the asymptotic consistency and the finite-sample guarantees established in Propositions 7 and 8, we examine the behavior of the estimation error as n tends to infinity both in the low-dimensional regime with fixed p and the high-dimensional regime where the ratio p/n remains constant. In both cases, we evaluate our estimators under two scenarios: (i) when the true covariance matrix is $\Sigma_0 = I_p$, and (ii) when Σ_0 is a banded $p \times p$ matrix with ones on the diagonal and 0.5 on the immediate off-diagonals above and below.

FIGURE 4. Consistency of X_n^* and $\hat{\Sigma}_n$ in the low-dimensional regime when p is fixed.

5.2.1. Consistency (Low-Dimensional Regime)

Assume that $p = 10$ is fixed and that $\hat{\Sigma}_n$ is the sample covariance matrix constructed from n independent samples drawn from the distribution $\mathbb{P} = \mathcal{N}(0, \Sigma_0)$. According to Proposition 8, finite-sample guarantees require uncertainty set radii of order $\mathcal{O}(n^{-1/2})$. This motivates us to set $\varepsilon_n = 5n^{-1/2}$. Proposition 7 asserts that the distributionally robust estimator X_n^* converges almost surely to Σ_0 as n tends to infinity, given that the sample covariance matrix is consistent and ε_n tends to zero. To empirically verify this result, we plot the Frobenius distance between X_n^* and Σ_0 as a function of n . Figure 4 displays the mean Frobenius losses (solid lines) along with one-standard-deviation bands (shaded regions), computed over 10 independent datasets of size n . The results reveal that the Frobenius errors of the Wasserstein, Kullback-Leibler and Fisher-Rao estimators all approximate straight lines with negative slopes on a log-log scale, indicating polynomial decay in n . This observed behavior is consistent with the theoretical convergence guarantee of Proposition 7. However, the empirical covariance matrix converges faster than all tested distributionally robust estimators.

5.2.2. Finite-Sample Performance (High-Dimensional Regime)

We adopt the same experimental setup as in Section 5.2.1 but now focus on a high-dimensional regime where the dimension $p_n = 0.8n$ grows linearly with n . Proposition 8 states that if $\varepsilon_n = \mathcal{O}(p_n^{3/2}n^{-1/2}) = \mathcal{O}(n)$ and n is sufficiently large, then the true covariance matrix Σ_0 lies within the uncertainty set $\mathcal{B}_{\varepsilon_n}(\hat{\Sigma}_n)$ with constant confidence. By construction, the distributionally robust estimator X_n^* , which essentially minimizes the worst-case Frobenius error over all $\Sigma \in \mathcal{B}_{\varepsilon_n}(\hat{\Sigma}_n)$, is expected to exhibit a small Frobenius error. We now empirically investigate this hypothesis. Specifically, for each n we use a ternary search algorithm to determine a radius $\hat{\varepsilon}_n$ such that the corresponding distributionally robust estimator X_n^* minimizes the average Frobenius distance to Σ_0 over 10 independent datasets of size n . Figure 5 shows the empirically optimal radius $\hat{\varepsilon}_n$ as a function of n for the banded covariance matrix Σ_0 (the results are qualitatively similar when Σ_0 is the identity matrix). We observe that $\hat{\varepsilon}_n$ grows approximately linearly with n , consistent with the theoretical scaling of ε_n from Proposition 8 when $p_n = 0.8n$. This observation suggests that, in the standard high-dimensional regime where $p_n = \mathcal{O}(n)$, one may simply approximate the optimal radius by $cp_n^{3/2}n^{-1/2}$ and select the tuning constant $c > 0$ via cross-validation, for example. Figure 6 plots the normalized Frobenius loss $\|X_n^* - \Sigma_0\|_F / \|\Sigma_0\|_F$ as a function of n for the distributionally robust estimator corresponding to $\hat{\varepsilon}_n$. The normalization by $\|\Sigma_0\|_F$ accounts for increasing dimension, allowing for meaningful comparison across

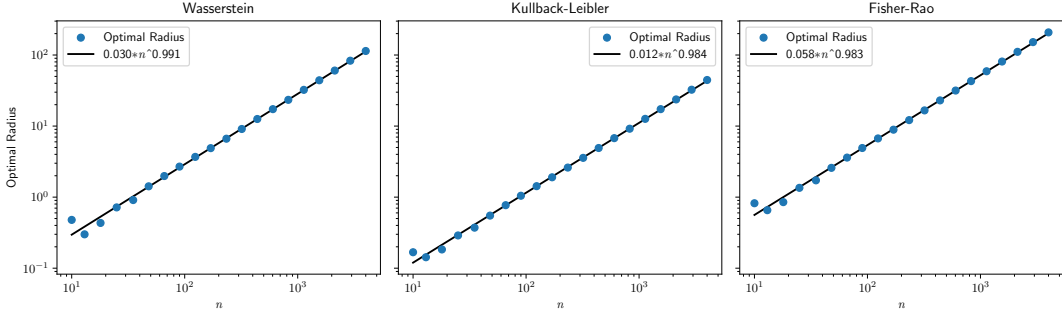


FIGURE 5. Optimal radius in the high-dimensional regime with a least-squares fit in log-log space. The plot shows $\hat{\varepsilon}_n$ as a function of n , along with a fitted curve of the form cn^α .

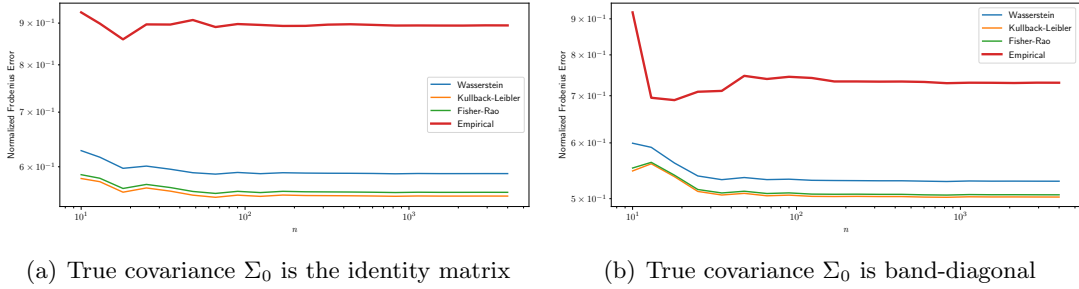


FIGURE 6. Normalized Frobenius error of the distributionally robust estimator based on the empirically optimal radius $\hat{\varepsilon}_n$ in the high-dimensional regime, plotted as a function of n .

different values of n . We find that the Wasserstein, Kullback-Leibler and Fisher-Rao estimators all achieve significantly smaller relative Frobenius error than the empirical covariance matrix across all values of n . This suggests that robustification is beneficial in high-dimensional regimes where $p_n = \mathcal{O}(n)$, even though the relative Frobenius loss may not decrease with n .

5.3. Frobenius Error

In the next experiment, we use synthetic data to analyze the Frobenius risk of different covariance estimators. Specifically, we construct a diagonal covariance matrix $\Sigma_0 \in \mathbb{S}_{++}^{100}$ with 90 eigenvalues equal to 1 and 10 ‘spiking’ eigenvalues equal to $M \in \{10, 100, 500\}$. Thus, we have $\kappa(\Sigma_0) = M$. Next, we let $\hat{\Sigma}$ be the sample covariance matrix constructed from $n \in \{100, 200, 500\}$ independent samples from $\mathbb{P} = \mathcal{N}(0, \Sigma_0)$. This experimental setup captures the small to medium sample size regime with $n \gtrsim p$, in which we expect $\hat{\Sigma}$ to provide a poor approximation for Σ_0 . We thus compare $\hat{\Sigma}$ against the Kullback-Leibler, Wasserstein, and Fisher-Rao covariance shrinkage estimators as well as against the linear shrinkage estimator with shrinkage target $\frac{1}{p} \text{Tr}[\hat{\Sigma}]I_p$ and against the NLLW estimator. Figure 7 visualizes the Frobenius loss of all estimators as a function of the underlying hyperparameters, that is, the radius ε of the uncertainty set for the distributionally robust estimators and the mixing weight α for the linear shrinkage estimator. The NLLW estimator and the sample covariance matrix involve no hyperparameters and are thus visualized as horizontal lines. Figure 7 shows both the means (solid lines) as well as the areas within one standard deviation of the means (shaded areas) of the Frobenius loss based on 10 independent training sets for all possible combinations of M and n . As ε tends to 0, all distributionally robust estimators approach the sample covariance

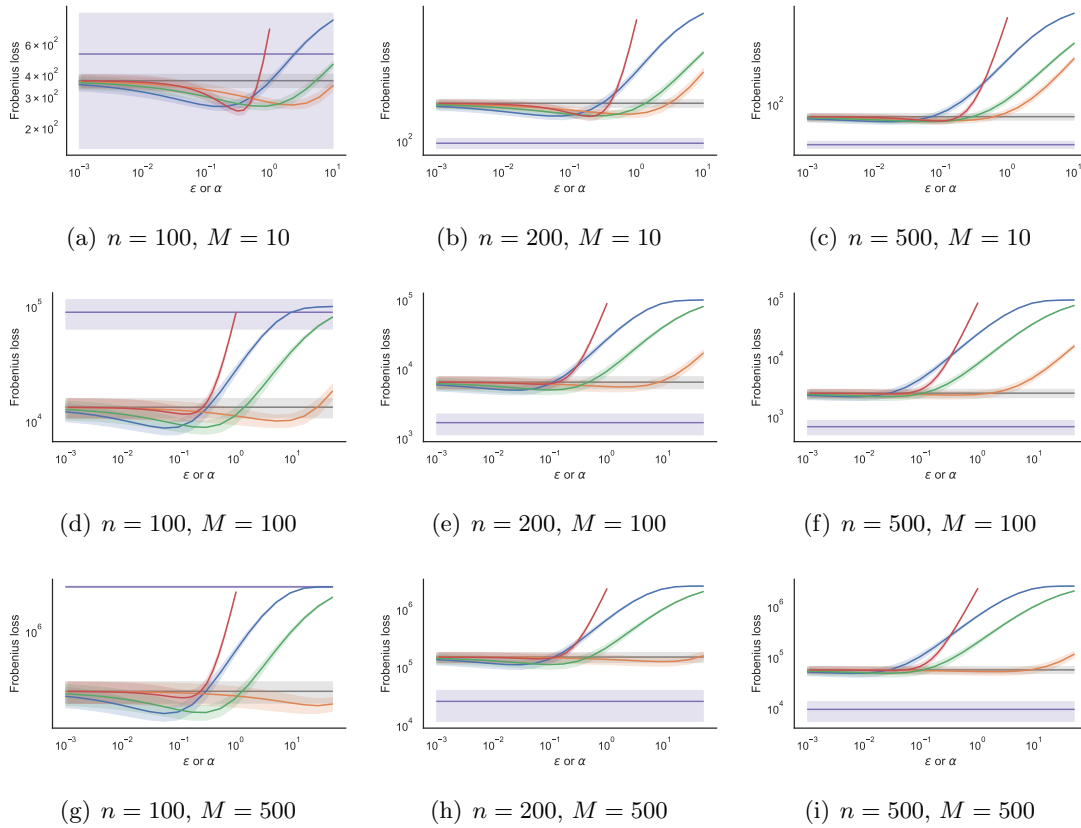


FIGURE 7. Frobenius loss of the Kullback-Leibler (blue), Wasserstein (orange), and Fisher-Rao (green) covariance shrinkage estimators and of the linear shrinkage estimator (red) as a function of the underlying hyperparameter (radius ε or mixing weight α) for different spike sizes M and sample sizes n . The sample covariance matrix (gray) and the NLLW estimator (purple) involve no hyperparameters; thus, their Frobenius error is constant.

matrix. Thus, they overfit the data and display a high variance. As ε tends to $\sum_{i=1}^p d(0, \hat{x}_i)$, on the other hand, all distributionally robust estimators collapse to 0 and thus display a high bias. We thus face a classical bias-variance trade-off. Figure 7 reveals that the Frobenius loss of the distributionally robust estimators is minimal at intermediate values of ε . We observe that the linear shrinkage estimator is competitive with the distributionally robust estimators for well-conditioned covariance matrices (small M , top row). As the covariance matrix becomes more ill-conditioned (large M , middle and bottom rows), the linear shrinkage estimator is dominated by the distributionally robust estimators, which attain a significantly smaller Frobenius loss. The advantage of the distributionally robust estimators relative to the nominal sample covariance matrix diminishes with increasing sample size n . The NLLW estimator is designed to be asymptotically optimal and, therefore, dominates the other estimators for large sample sizes. However, it is suboptimal if training samples are scarce.

The insights of this synthetic experiment can be summarized as follows. Linear shrinkage estimators are suitable for well-conditioned covariance matrices and small sample sizes, while the NLLW estimator is preferable for large sample sizes, irrespective of the condition number. The distributionally robust estimators perform better when the covariance matrix is ill-conditioned and training samples are scarce.

5.4. Minimum Variance Portfolio Selection

We consider the problem of constructing the minimum variance portfolio of p risky assets by solving the convex program $\min_{w \in \mathbb{R}^p} \{w^\top \Sigma_0 w : w^\top \mathbf{1} = 1\}$ [22], where $\mathbf{1}$ denotes the p -dimensional vector of ones, and $\Sigma_0 \in \mathbb{S}_{++}^p$ stands for the covariance matrix of the asset returns over the investment horizon. The unique optimal solution of this problem is given by $w^* = \Sigma_0^{-1} \mathbf{1} / \mathbf{1}^\top \Sigma_0^{-1} \mathbf{1}$. In practice, however, the distribution of the asset returns is unknown, and thus the covariance matrix Σ_0 needs to be estimated from historical data. If the chosen covariance estimator $\hat{\Sigma}$ is invertible, then it is natural to use $\hat{w}^* = \hat{\Sigma}^{-1} \mathbf{1} / \mathbf{1}^\top \hat{\Sigma}^{-1} \mathbf{1}$ as an estimator for the minimum variance portfolio. This approach seems reasonable, provided that the asset return distribution is stationary over the (past) estimation window and the (future) investment horizon.

In the next experiment, we assess the minimum variance portfolios induced by several covariance estimators on the “48 Industry Portfolios” dataset from the Fama-French online library,³ which contains monthly returns of 48 portfolios grouped by industry. Specifically, we adopt the following rolling horizon procedure from January 1974 to December 2022. First, we estimate Σ_0 from the historical asset returns within a rolling estimation window of 50 months and construct the corresponding minimum variance portfolio. We then compute the returns of this portfolio over the k months immediately after the estimation window. Finally, the covariance estimators are recalibrated based on a new estimation window shifted ahead by k months, and the procedure starts afresh. Some covariance estimators involve a hyperparameter, which we calibrate via leave-one-out cross-validation on the 50 return samples in each estimation window. To this end, we assume that the mixing weight α of the linear shrinkage estimator with shrinkage target $\frac{1}{p} \text{Tr}[\hat{\Sigma}] I_p$ ranges from 10^{-5} to 1, whereas the radius ε of the uncertainty set ranges from 10^{-5} to 10^2 for the Kullback-Leibler shrinkage estimator, from 10^{-10} to 10^4 for the Fisher-Rao covariance shrinkage estimators and from 10^{-10} to 10^8 for the Wasserstein covariance shrinkage estimator. We discretize these parameter ranges into 50 logarithmically spaced candidate values and select the one that induces the smallest portfolio variance. Given the selected hyperparameter, the covariance estimator corresponding to the current estimation window is computed using all 50 data points. In the following, we measure the quality of a given covariance estimator by Sharpe ratio and the mean and the standard deviation of the portfolio returns generated by the above rolling horizon procedure over the backtesting period.

Figure 8 displays the Sharpe ratios, means, and standard deviations corresponding to different covariance estimators as a function of the length k of an updating period. All shrinkage estimators produce lower standard deviations and higher Sharpe ratios than the sample covariance matrix. Even though the mean portfolio returns of the sample covariance matrix are—on average—similar to those of the shrinkage estimators, they change rapidly with k , which is troubling for investors who need to select k before seeing the results of the backtest. The distributionally robust estimators proposed in this paper outperform the other shrinkage estimators in terms of mean returns and Sharpe ratios for most choices of k , and the Wasserstein covariance shrinkage estimator results in the globally highest Sharpe ratio. However, the Kullback-Leibler and Fisher-Rao covariance shrinkage estimators result in slightly higher means and standard deviations.

³https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

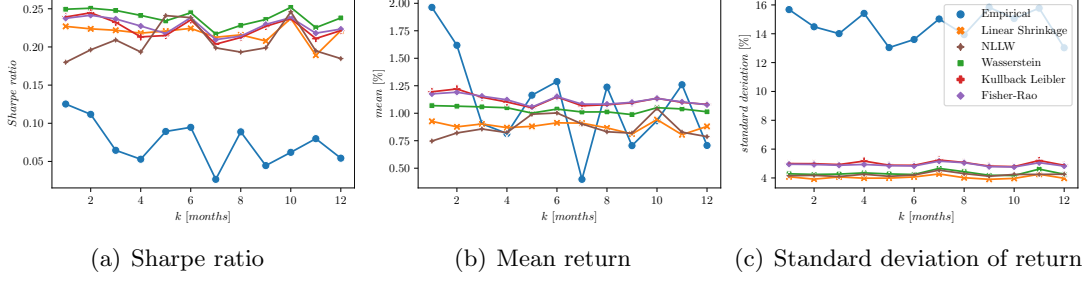


FIGURE 8. Sharpe ratios, means, and standard deviations induced by different covariance estimators on the “48 Industry Portfolios” depending on the length k of an updating period.

5.5. Linear and Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) seeks to predict a label $y \in \{0, 1\}$ from a feature vector $z \in \mathbb{R}^p$ under the assumption that $z|y \sim \mathcal{N}(\mu_y, \Sigma_y)$ for every $y \in \{0, 1\}$. If the mean μ_y , the covariance matrix Σ_y as well as the marginal class probability p_y are known for all $y \in \{0, 1\}$, then the Bayes-optimal classifier predicts y as a solution of $\min_{y \in \{0, 1\}} (z - \mu_y) \Sigma_y^{-1} (z - \mu_y) + \log \det(\Sigma_y) - 2 \log(p_y)$. Linear discriminant analysis (LDA) operates under the additional assumption that $\Sigma_0 = \Sigma_1$. The decision boundaries of the resulting LDA and QDA classifiers are thus given by linear hyperplanes and quadratic hypersurfaces, respectively [20].

In the last experiment, we use LDA and QDA to address the breast cancer detection [68] and banknote authentication [39] problems from the UCI Machine Learning Repository. As the distribution governing y and z is unobservable, we replace the unknown class probabilities p_y and class means μ_y by the empirical frequencies and sample average estimators, respectively, and we use different shrinkage estimators for the unknown covariance matrices Σ_y . All tested shrinkage estimators use the debiased empirical covariance matrix as the nominal estimator. QDA constructs a separate covariance estimator for each class y that only uses class- y samples, whereas LDA pools all samples to construct a single joint covariance estimator.

We use 50% of each dataset for training and the rest for testing. The hyperparameters ε (for the distributionally robust shrinkage estimators) and α (for the linear shrinkage estimator) are selected by the holdout method with a validation set comprising 20% of the training data. The quality of a covariance estimator is then measured by the accuracy (*i.e.*, the proportion of correct predictions) of the resulting LDA and QDA classifiers. Table 3 reports the means and standard errors of the accuracy achieved by different covariance estimators. We observe that shrinking the empirical covariance estimator can improve the performance of LDA and QDA, and that nonlinear shrinkage methods outperform the linear shrinkage method across all experiments. The Kullback-Leibler covariance shrinkage estimator consistently performs well. QDA based on the NLLW estimator attains the highest accuracy on the banknote authentication dataset but performs poorly on the breast cancer dataset. On the other hand, the distributionally robust covariance estimators are consistently on par with or better than the empirical and the linear shrinkage estimator. Note that the best-performing distributionally robust shrinkage estimator changes with the dataset. This highlights the usefulness of our approach, which results in a zoo of complementary covariance shrinkage estimators.

TABLE 3. Mean (standard error) of the LDA and QDA accuracy based on 100 independent permutations of the underlying dataset

	Dataset	Empirical	Linear	NLLW	Wasserstein	Kullback-Leibler	Fisher-Rao
LDA	Banknote	0.9751(0.0005)	0.9754(0.0005)	0.9510(0.0011)	0.9761(0.0005)	0.9763(0.0005)	0.9759(0.0005)
	Cancer	0.9520(0.0011)	0.9365(0.0015)	0.8902(0.0015)	0.9520(0.0011)	0.8874(0.0043)	0.9515(0.0013)
QDA	Banknote	0.9854(0.0005)	0.9839(0.0005)	0.9877(0.0004)	0.9854(0.0005)	0.9853(0.0005)	0.9854(0.0005)
	Cancer	0.9418(0.0012)	0.8945(0.0027)	0.6320(0.0052)	0.9418(0.0012)	0.9451(0.0013)	0.9414(0.0016)

Acknowledgments. This research was supported by the Hong Kong Research Grants Council under the GRF project 15304422, by the Swiss National Science Foundation under the NCCR Automation, grant agreement 51NF40_180545, and by CUHK through the ‘Improvement on Competitiveness in Hiring New Faculties Funding Scheme’ and the CUHK Direct Grant with project number 4055191.

REFERENCES

- [1] S. ABBOTT, *Understanding Analysis*, Springer, 2015.
- [2] P.-A. ABSIL, R. MAHONY, AND R. SEPULCHRE, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2009.
- [3] C. ATKINSON AND A. F. MITCHELL, *Rao’s distance measure*, Sankhyā: The Indian Journal of Statistics, Series A, (1981), pp. 345–365.
- [4] R. BHATIA, *Matrix Analysis*, Springer, 1997.
- [5] R. BHATIA, *Positive Definite Matrices*, Princeton University Press, 2007.
- [6] J. BLANCHET, K. MURTHY, AND V. A. NGUYEN, *Statistical analysis of Wasserstein distributionally robust estimators*, in *Emerging Optimization Methods and Modeling Techniques with Applications*, INFORMS, 2021, pp. 227–254.
- [7] J. BLANCHET, K. MURTHY, AND N. SI, *Confidence regions in Wasserstein distributionally robust estimation*, Biometrika, 109 (2021), pp. 295–315.
- [8] T. BODNAR, A. K. GUPTA, AND N. PAROLYA, *Direct shrinkage estimation of large dimensional precision matrix*, Journal of Multivariate Analysis, 146 (2016), pp. 223–236.
- [9] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, 2004.
- [10] N. BUI, D. NGUYEN, M.-C. YUE, AND V. A. NGUYEN, *Coverage-validity-aware algorithmic recourse*, Operations Research, (2025).
- [11] J. E. DENNIS JR AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, 1996.
- [12] D. DONOHO, M. GAVISH, AND I. JOHNSTONE, *Optimal shrinkage of eigenvalues in the spiked covariance model*, The Annals of Statistics, 46 (2018), pp. 1742–1778.
- [13] V. M. EGUILUZ, D. R. CHIALVO, G. A. CECCHI, M. BALIKI, AND A. V. APKARIAN, *Scale-free brain functional networks*, Physical Review Letters, 94 (2005), p. 018102.
- [14] O. P. FERREIRA, M. S. LOUZEIRO, AND L. PRUDENTE, *Gradient method for optimization on Riemannian manifolds with lower bounded curvature*, SIAM Journal on Optimization, 29 (2019), pp. 2517–2541.
- [15] R. GAO, *Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality*, Operations Research, 71 (2023), pp. 2291–2306.
- [16] R. GAO, X. CHEN, AND A. J. KLEYWEGT, *Wasserstein distributionally robust optimization and variation regularization*, Operations Research, 72 (2024), pp. 1177–1191.
- [17] L. E. GHAOUI, M. OKS, AND F. OUSTRY, *Worst-case value-at-risk and robust portfolio optimization: A conic programming approach*, Operations Research, 51 (2003), pp. 543–556.
- [18] C. R. GIVENS AND R. M. SHORTT, *A class of Wasserstein metrics for probability distributions.*, Michigan Mathematical Journal, 31 (1984), pp. 231–240.
- [19] G. H. HARDY, J. E. LITTLEWOOD, AND G. PÓLYA, *Inequalities*, Cambridge University Press, 1952.
- [20] T. HASTIE, R. TIBSHIRANI, AND J. H. FRIEDMAN, *The Elements of Statistical Learning*, Springer, 2009.

- [21] A. E. HOERL AND R. W. KENNARD, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.
- [22] R. JAGANNATHAN AND T. MA, *Risk reduction in large portfolios: Why imposing the wrong constraints helps*, The Journal of Finance, 58 (2003), pp. 1651–1683.
- [23] W. JAMES AND C. STEIN, *Estimation with quadratic loss*, in Breakthroughs in Statistics, Springer, 1992, pp. 443–460.
- [24] H. JEFFREYS, *An invariant form for the prior probability in estimation problems*, Proceedings of the Royal Society of London A, 186 (1946), pp. 453–461.
- [25] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering, 82 (1960), pp. 35–45.
- [26] H. KOMIYA, *Elementary proof for Sion’s minimax theorem*, Kodai Mathematical Journal, 11 (1988), pp. 5–7.
- [27] D. KUHN, P. MOHAJERIN ESFAHANI, V. A. NGUYEN, AND S. SHAFIEEZADEH-ABADEH, *Wasserstein distributionally robust optimization: Theory and applications in machine learning*, in Operations Research & Management Science in the Age of Analytics, INFORMS, 2019, pp. 130–166.
- [28] S. KULLBACK, *Information Theory and Statistics*, Courier Corporation, 1997.
- [29] S. LANG, *Fundamentals of Differential Geometry*, Springer, 2012.
- [30] O. LEDOIT AND M. WOLF, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of Empirical Finance, 10 (2003), pp. 603–621.
- [31] ———, *Honey, I shrunk the sample covariance matrix*, The Journal of Portfolio Management, 30 (2004), pp. 110–119.
- [32] ———, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365 – 411.
- [33] ———, *Nonlinear shrinkage estimation of large-dimensional covariance matrices*, The Annals of Statistics, 40 (2012), pp. 1024–1060.
- [34] ———, *Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks*, The Review of Financial Studies, 30 (2017), pp. 4349–4388.
- [35] ———, *Analytical nonlinear shrinkage of large-dimensional covariance matrices*, The Annals of Statistics, 48 (2020), pp. 3043–3065.
- [36] ———, *Quadratic shrinkage for large covariance matrices*, Bernoulli, 28 (2022), pp. 1519–1547.
- [37] J. M. LEE, *Riemannian Manifolds: An Introduction to Curvature*, Springer, 2006.
- [38] ———, *Introduction to Smooth Manifolds*, Springer, 2013.
- [39] V. LOHWEG, *Banknote authentication dataset*. UCI Machine Learning Repository, 2013. <https://doi.org/10.24432/C55P57>.
- [40] R. N. MANTEGNA, *Hierarchical structure in financial markets*, The European Physical Journal B, 11 (1999), pp. 193–197.
- [41] H. MARKOWITZ, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.
- [42] P. MOHAJERIN ESFAHANI AND D. KUHN, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Mathematical Programming, 171 (2018), pp. 115–166.
- [43] V. A. NGUYEN, D. KUHN, AND P. MOHAJERIN ESFAHANI, *Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator*, Operations Research, 70 (2022), pp. 490–515.
- [44] V. A. NGUYEN, S. SHAFIEEZADEH-ABADEH, D. FILIPOVIĆ, AND D. KUHN, *Mean-covariance robust risk measurement*, arXiv preprint arXiv:2112.09959, (2021).
- [45] V. A. NGUYEN, S. SHAFIEEZADEH-ABADEH, D. KUHN, AND P. MOHAJERIN ESFAHANI, *Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization*, Mathematics of Operations Research, 48 (2023), pp. 1–37.
- [46] V. A. NGUYEN, S. SHAFIEEZADEH-ABADEH, M.-C. YUE, D. KUHN, AND W. WIESEMANN, *Calculating optimistic likelihoods using (geodesically) convex optimization*, in Advances in Neural Information Processing Systems, 2019, pp. 13920–13931.
- [47] V. A. NGUYEN, S. SHAFIEEZADEH ABABEH, M.-C. YUE, D. KUHN, AND W. WIESEMANN, *Optimistic distributionally robust optimization for nonparametric likelihood approximation*, in Advances in Neural Information Processing Systems, 2019, pp. 13942–13953.

- [48] K. PEARSON, *Note on regression and inheritance in the case of two parents*, Proceedings of the Royal Society of London, 58 (1895), pp. 240–242.
- [49] M. PERLMAN, *STAT 542: Multivariate Statistical Analysis*, Lecture Notes, University of Washington, Seattle, (2007).
- [50] K. B. PETERSEN, *The Matrix Cookbook*, 2012.
- [51] B. RAJARATNAM AND D. VINCENZI, *A theoretical study of Stein’s covariance estimator*, Biometrika, 103 (2016), pp. 653–666.
- [52] R. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, 1997.
- [53] S. SHAFIEEZADEH-ABADEH, L. AOLARITEI, F. DÖRFLER, AND D. KUHN, *New perspectives on regularization and computation in optimal transport-based distributionally robust optimization*, arXiv preprint arxiv.2303.03900, (2023).
- [54] S. SHAFIEEZADEH-ABADEH, D. KUHN, AND P. MOHAJERIN ESFAHANI, *Regularization via mass transportation*, Journal of Machine Learning Research, 20 (2019), pp. 1–68.
- [55] S. SHAFIEEZADEH-ABADEH, V. A. NGUYEN, D. KUHN, AND P. MOHAJERIN ESFAHANI, *Wasserstein distributionally robust Kalman filtering*, in Advances in Neural Information Processing Systems, 2018, pp. 8483–8492.
- [56] W. F. SHARPE, *A simplified model for portfolio analysis*, Management Science, 9 (1963), pp. 277–293.
- [57] C. STEIN, *Estimation of a covariance matrix*, Rietz Lecture, in 39th Annual Meeting IMS, Institute of Mathematical Statistics, 1975.
- [58] B. TAŞKESEN, D. IANCU, Ç. KOÇYIĞIT, AND D. KUHN, *Distributionally robust linear quadratic control*, in Advances in Neural Information Processing Systems, 2023, pp. 18613–18632.
- [59] B. TASKESSEN, M.-C. YUE, J. BLANCHET, D. KUHN, AND V. A. NGUYEN, *Sequential domain adaptation by synthesizing distributionally robust experts*, in International Conference on Machine Learning, 2021, pp. 10162–10172.
- [60] R. TAYLOR, *Interpretation of the correlation coefficient: A basic review*, Journal of Diagnostic Medical Sonography, 6 (1990), pp. 35–39.
- [61] A. TOULOUMIS, *Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings*, Computational Statistics & Data Analysis, 83 (2015), pp. 251–261.
- [62] C. UDRISTE, *Convex Functions and Optimization Methods on Riemannian Manifolds*, Springer, 2013.
- [63] H. R. VAN DER VAART, *On certain characteristics of the distribution of the latent roots of a symmetric random matrix under general conditions*, The Annals of Mathematical Statistics, 32 (1961), pp. 864–873.
- [64] W. N. VAN WIERINGEN, *Lecture Notes on Ridge Regression*, arXiv preprint arXiv:1509.09169, (2015).
- [65] C. VILLANI, *Optimal Transport: Old and New*, Springer, 2008.
- [66] H. VU, T. TRAN, M.-C. YUE, AND V. A. NGUYEN, *Distributionally robust fair principal components via geodesic descents*, in International Conference on Learning Representations, 2022.
- [67] M. J. WAINWRIGHT, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.
- [68] W. WOLBERG, O. MANGASARIAN, N. STREET, AND W. STREET, *Breast cancer Wisconsin (diagnostic) dataset*. UCI Machine Learning Repository, 1992. <https://doi.org/10.24432/C5DW2B>.
- [69] M.-C. YUE, *A matrix generalization of the Hardy-Littlewood-Pólya rearrangement inequality and its applications*, arXiv preprint arXiv:2006.08144, (2020).
- [70] P. ZHANG, J. ZHANG, AND S. SRA, *Sion’s minimax theorem in geodesic metric spaces and a Riemannian extragradient algorithm*, SIAM Journal on Optimization, 33 (2023), pp. 2885–2908.
- [71] M. ZORZI, *Robust Kalman filtering under model perturbations*, IEEE Transactions on Automatic Control, 62 (2017), pp. 2902–2907.

APPENDIX

The appendix is organized as follows. In Appendix A, we prove Theorem 1 and derive basic properties of γ^* and x_i^* , which will be used in Appendix B to establish the computational, structural and statistical properties of the distributionally robust estimators. Appendices C and D verify Assumptions 1 and 2 for all divergences in Table 1, respectively. The insights of Appendices C and D are used in Appendix E to prove the results of Section 4.

APPENDIX A. PROOF OF THEOREM 1

A.1. Proof of Proposition 2

To simplify the subsequent discussions, for any minimization problem designated by “P,” say, we use “Min(P),” “Argmin(P)” and “Fea(P)” to denote its minimum/infimum, the set of its optimal solutions and its feasible region, respectively.

Proof of Proposition 2. Select any $\Sigma \in \text{Fea}(\text{P}_{\text{Mat}})$, and use $\Sigma = V_{\Sigma} \text{Diag}(\lambda(\Sigma)) V_{\Sigma}^{\top}$ to denote its eigenvalue decomposition. By our notational conventions, we have $0 \leq \lambda_1(\Sigma) \leq \dots \leq \lambda_p(\Sigma)$. We then obtain

$$\begin{aligned} \sum_{i=1}^p d(\lambda_i(\Sigma), \hat{x}_i) &= D(\text{Diag}(\lambda(\Sigma)), \text{Diag}(\hat{x})) \leq D(\widehat{V}^{\top} V_{\Sigma} \text{Diag}(\lambda(\Sigma)) V_{\Sigma}^{\top} \widehat{V}, \text{Diag}(\hat{x})) \\ &= D(V_{\Sigma} \text{Diag}(\lambda(\Sigma)) V_{\Sigma}^{\top}, \widehat{V} \text{Diag}(\hat{x}) \widehat{V}^{\top}) = D(\Sigma, \widehat{\Sigma}) \leq \varepsilon, \end{aligned} \quad (12)$$

where the first equality follows from Assumption 2(b), the first inequality follows from Assumption 2(c), and the second equality follows from Assumption 2(a). This implies that $\lambda(\Sigma) \in \text{Fea}(\text{P}_{\text{Vec}})$.

Next, select any $x \in \text{Fea}(\text{P}_{\text{Vec}})$ such that $\widehat{V} \text{Diag}(x) \widehat{V}^{\top} \in \mathbb{S}_+^p$. We thus have

$$D(\widehat{V} \text{Diag}(x) \widehat{V}^{\top}, \widehat{\Sigma}) = D(\widehat{V} \text{Diag}(x) \widehat{V}^{\top}, \widehat{V} \text{Diag}(\hat{x}) \widehat{V}^{\top}) = D(\text{Diag}(x), \text{Diag}(\hat{x})) = \sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon, \quad (13)$$

where the three equalities follow from the eigenvalue decomposition of $\widehat{\Sigma}$, Assumption 2(a) and Assumption 2(b), respectively. This implies that $\widehat{V} \text{Diag}(x) \widehat{V}^{\top} \in \text{Fea}(\text{P}_{\text{Mat}})$. In summary, we have thus shown that problem (P_{Mat}) is feasible if and only if problem (P_{Vec}) is feasible. This establishes assertion (i).

As for assertion (ii), assume that $\text{Argmin}(\text{P}_{\text{Vec}}) \neq \emptyset$ for otherwise the claim is trivial. Choose then any $x^* \in \text{Argmin}(\text{P}_{\text{Vec}})$, and note that $\widehat{V} \text{Diag}(x^*) \widehat{V}^{\top} \in \text{Fea}(\text{P}_{\text{Mat}})$ by virtue of (13). It remains to be shown that $\widehat{V} \text{Diag}(x^*) \widehat{V}^{\top} \in \text{Argmin}(\text{P}_{\text{Mat}})$. Suppose, for the sake of contradiction, that there is $\Sigma' \in \text{Fea}(\text{P}_{\text{Mat}})$ with

$$\|\Sigma'\|_{\text{F}}^2 < \|\widehat{V} \text{Diag}(x^*) \widehat{V}^{\top}\|_{\text{F}}^2,$$

and let $\Sigma' = V' \text{Diag}(\lambda(\Sigma')) V'^{\top}$ be the eigenvalue decomposition of Σ' for some $V' \in \mathcal{O}_p$. By (12), we then have $\lambda(\Sigma') \in \text{Fea}(\text{P}_{\text{Vec}})$, which contradicts the optimality of x^* in problem (P_{Vec}) because

$$\|\lambda(\Sigma')\|_2^2 = \|\Sigma'\|_{\text{F}}^2 < \|\widehat{V} \text{Diag}(x^*) \widehat{V}^{\top}\|_{\text{F}}^2 = \|x^*\|_2^2.$$

Therefore, $\widehat{V} \text{Diag}(x^*) \widehat{V}^{\top} \in \text{Argmin}(\text{P}_{\text{Mat}})$. This proves assertion (ii).

As for assertion (iii), assume that $\text{Argmin}(\mathbf{P}_{\text{Mat}}) \neq \emptyset$ for otherwise the claim is trivial. Choose then any $\Sigma^* \in \text{Argmin}(\mathbf{P}_{\text{Mat}})$, and note that $\lambda(\Sigma^*) \in \text{Fea}(\mathbf{P}_{\text{Vec}})$ by virtue of (12). It remains to be shown that $\lambda(\Sigma^*) \in \text{Argmin}(\mathbf{P}_{\text{Vec}})$. Suppose, for the sake of contradiction, that there is $x' \in \text{Fea}(\mathbf{P}_{\text{Vec}})$ with

$$\|x'\|_2^2 < \|\lambda(\Sigma^*)\|_2^2.$$

By (13), we then have $\widehat{V} \text{Diag}(x') \widehat{V}^\top \in \text{Fea}(\mathbf{P}_{\text{Mat}})$, which contradicts the optimality of Σ^* in $(\mathbf{P}_{\text{Mat}})$ because

$$\left\| \widehat{V} \text{Diag}(x') \widehat{V}^\top \right\|_{\text{F}}^2 = \|x'\|_2^2 < \|\lambda(\Sigma^*)\|_2^2 = \|\Sigma^*\|_{\text{F}}^2.$$

Therefore, $\lambda(\Sigma^*) \in \text{Argmin}(\mathbf{P}_{\text{Vec}})$. This proves assertion (iii).

Finally, in order to prove assertion (iv), we need to show that any $\Sigma \in \text{Fea}(\mathbf{P}_{\text{Mat}})$ corresponds to some $x \in \text{Fea}(\mathbf{P}_{\text{Vec}})$ with the same objective function value and vice versa. However, this follows in a straightforward manner from the proof of assertion (i). Details are omitted for brevity. \square

A.2. Proof of Proposition 3

The next lemma shows that any solution of problem $(\mathbf{P}_{\text{Vec}})$ shrinks \hat{x} towards the origin. This will imply that our proposed distributionally robust estimators constitute shrinkage estimators. From now on we use $d_b(\cdot)$ as a notational shorthand for the function $d(\cdot, b)$ for any fixed $b \geq 0$.

Lemma 5 (Eigenvalue shrinkage). *If Assumptions 2 and 3(a) hold and x^* solves problem $(\mathbf{P}_{\text{Vec}})$, then we have $x_i^* \in \text{dom}(d_{\hat{x}_i})$ and $x_i^* \leq \hat{x}_i$ for all $i = 1, \dots, p$.*

Proof of Lemma 5. Select any $x^* \in \text{Argmin}(\mathbf{P}_{\text{Vec}})$. As $x^* \in \text{Fea}(\mathbf{P}_{\text{Vec}})$, it is clear that $x_i^* \in \text{dom}(d_{\hat{x}_i})$ for all $i = 1, \dots, p$. Next, suppose that $x_j^* > \hat{x}_j$ for some $j = 1, \dots, p$, and define $\tilde{x} \in \mathbb{R}_+^p$ through

$$\tilde{x}_i = \begin{cases} \hat{x}_j & \text{if } i = j, \\ x_i^* & \text{if } i \neq j. \end{cases}$$

Recall now that if Assumption 2(b) holds, then d constitutes a spectral divergence on \mathbb{R}_+ . Assumption 3(a) further implies that $(\hat{x}_j, \hat{x}_j) \in \text{dom}(d)$. Hence, $d(\hat{x}_j, \hat{x}_j) = 0 < d(x_j^*, \hat{x}_j)$, which ensures that $\tilde{x} \in \text{Fea}(\mathbf{P}_{\text{Vec}})$. However, from the construction of \tilde{x} it is evident that $\|\tilde{x}\|_2^2 < \|x^*\|_2^2$, which contradicts the optimality of x^* in $(\mathbf{P}_{\text{Vec}})$. Thus, we have $x_i^* \leq \hat{x}_i$ for all $i = 1, \dots, p$. This observation completes the proof. \square

Lemma 5 allows us to prove the existence and uniqueness of the proposed robust covariance estimators.

Proposition 9 (Existence and uniqueness of optimal solutions). *If Assumptions 2, 3 and 4 hold, then problems $(\mathbf{P}_{\text{Vec}})$ and $(\mathbf{P}_{\text{Mat}})$ admit a unique optimal solution. In addition, if Assumptions 1, 2, 3 and 4 hold, then there exists a unique distributionally robust estimator that solves problem (3).*

Proof of Proposition 9. Suppose first that only Assumptions 2, 3 and 4 hold. Lemma 5 then implies that problem $(\mathbf{P}_{\text{Vec}})$ has the same set of optimal solutions as the following variant

of (P_{Vec}) with box constraints

$$\begin{aligned} & \inf_{x \in \mathbb{R}^p} \|x\|_2^2 \\ & \text{s. t.} \quad \sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon \\ & \quad 0 \leq x_i \leq \hat{x}_i \quad \forall i = 1, \dots, p. \end{aligned} \tag{P'_{\text{Vec}}}$$

Note that problem (P'_{Vec}) is feasible due to Assumption 3(a), which posits that $d(\hat{x}_i, \hat{x}_i) = 0$ for all $i = 1, \dots, p$. Next, we show that the feasible region of (P'_{Vec}) is compact. To this end, note that $d(x_i, \hat{x}_i)$ is continuous in x_i on the interval $[0, \hat{x}_i]$ for every $i = 1, \dots, p$. Indeed, continuity trivially holds if $\hat{x}_i = 0$, in which case $[0, \hat{x}_i]$ collapses to a point. Otherwise, if $\hat{x}_i > 0$, then continuity follows from Assumption 2(b). This readily implies that the feasible region of (P'_{Vec}) is closed and—thanks to the box constraints—also compact. The solvability of problem (P'_{Vec}) thus follows from Weierstrass' maximum theorem, which applies because the objective function is continuous. Assumption 4 further implies that $d(x_i, \hat{x}_i)$ is convex in x_i on $[0, \hat{x}_i]$ for all $i = 1, \dots, p$, which implies that the feasible region of (P'_{Vec}) is convex. The uniqueness of the optimal solution x^* thus follows from the strong convexity of the objective function. This shows that problem (P_{Vec}) has a unique optimal solution. The other claims immediately follow from Propositions 1 and 2. \square

Proposition 10 (Solution of problem (P_{Vec})). *If Assumptions 2, 3 and 4 hold, then the unique minimizer x^* of problem (P_{Vec}) has the following properties. If $\hat{x}_i = 0$, then $x_i^* = 0$, and if $\hat{x}_i > 0$, then $x_i^* \in (0, \hat{x}_i)$ and*

$$0 = 2x_i^* + \gamma^* d'_{\hat{x}_i}(x_i^*), \tag{14}$$

where γ^* is a solution of the nonlinear equation $\sum_{i=1}^p d(s(\gamma^*, \hat{x}_i), \hat{x}_i) - \varepsilon = 0$.

The following lemma shows that d_b is strictly decreasing on $[0, b]$, which will be used to prove Proposition 10.

Lemma 6 (Derivative of d_b). *If Assumptions 2 and 4 hold, then we have*

$$d'_b(a) \leq -\frac{d(a, b)}{b-a} < -\frac{d(a, b)}{b} < 0 \quad \forall a \in (0, b), \quad \forall b > 0.$$

Proof of Lemma 6. Select any $b > 0$. As $d(\cdot, b)$ is finite and convex on $[0, b]$ thanks to Assumption 4, we have

$$0 = d(b, b) \geq d(a, b) + (b-a) d'_b(a) \quad \forall a \in (0, b).$$

The desired inequality then follows from an elementary rearrangement. \square

Proof of Proposition 10. Lemma 5 allows us to rewrite problem (P_{Vec}) equivalently as

$$\begin{aligned} & \min_{x \in \mathcal{C}} \|x\|_2^2 \\ & \text{s. t.} \quad \sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon, \end{aligned} \tag{P''_{\text{Vec}}}$$

where $\mathcal{C} = \mathcal{C}_1 \times \dots \times \mathcal{C}_p$ with $\mathcal{C}_i = [0, \hat{x}_i] \cap \text{dom}(d_{\hat{x}_i})$ for each $i = 1, \dots, p$. Note that the objective and constraint functions adopt finite values on \mathcal{C} . By Proposition 9 and Lemma 5, problem (P''_{Vec}) has a unique minimizer x^* satisfying $x_i^* = 0$ for all i with $\hat{x}_i = 0$. For such indices i , $d(0, 0) = d(\hat{x}_i, \hat{x}_i) = 0$ by Assumption 3(a). By removing the corresponding decision variables from (P''_{Vec}) and focusing on the optimization problem in the remaining variables,

we can therefore assume without loss of generality that $\hat{x}_i > 0$ for all $i = 1, \dots, p$. Hence, problem (P''_{Vec}) can be viewed as an ordinary convex program in the sense of [52, Section 28].

Following [52, Section 28], we define the Lagrangian $L : \mathbb{R} \times \mathbb{R}^p \rightarrow \overline{\mathbb{R}}$ of problem (P''_{Vec}) through

$$L(\gamma, x) = \begin{cases} \|x\|_2^2 + \gamma(\sum_{i=1}^p d(x_i, \hat{x}_i) - \varepsilon) & \text{if } x \in \mathcal{C}, \gamma \geq 0, \\ -\infty & \text{if } x \in \mathcal{C}, \gamma < 0, \\ +\infty & \text{if } x \notin \mathcal{C}. \end{cases}$$

By [52, Corollary 28.2.1 and Theorem 28.3], problem (P''_{Vec}) is thus equivalent to the minimax problem

$$\min_{x \in \mathbb{R}^p} \sup_{\gamma \in \mathbb{R}} L(\gamma, x) = \max_{\gamma \in \mathbb{R}} \min_{x \in \mathbb{R}^p} L(\gamma, x).$$

Specifically, the dual maximization problem on the right-hand side is solvable, and every maximizer $\gamma^* \geq 0$ gives rise to a saddle point (γ^*, x^*) of the minimax problem. Next, we prove that $\gamma^* > 0$. Suppose for the sake of contradiction that $\gamma^* = 0$. Since $x^* \in \mathcal{C}$, we find $L(\gamma^*, x^*) = L(0, x^*) = \|x^*\|_2^2$. If $x_i^* > 0$ for some i , then

$$0 < \|x^*\|_2^2 = L(0, x^*) \leq L(0, x) = \|x\|_2^2 \quad \forall x \in \mathcal{C},$$

where the second inequality holds because $(0, x^*)$ is a saddle point. However, the discussion after Assumption 4 implies that $\text{dom}(d_{\hat{x}_i})$ either equals \mathbb{R}_+ or \mathbb{R}_{++} for every $i = 1, \dots, p$. Hence, we have $\prod_{i=1}^p (0, \hat{x}_i] \subseteq \mathcal{C}$, that is, \mathcal{C} contains points that are arbitrarily close to 0. This leads to the contradiction

$$0 = \inf_{x \in \mathcal{C}} \|x\|_2^2 \geq \|x^*\|_2^2 > 0.$$

We may thus conclude that if $\gamma^* = 0$, then $x_i^* = 0$ for all i , that is, $x^* = 0$. However, this contradicts Assumption 3(b), which implies that $0 \notin \text{Fea}(P_{\text{Vec}})$. In summary, this shows that $\gamma^* > 0$.

Next, we note that for any dual optimal solution $\gamma^* > 0$, the minimization problem

$$\min_{x \in \mathbb{R}^p} L(\gamma^*, x) = \min_{x \in \mathcal{C}} \|x\|_2^2 + \gamma^* \left(\sum_{i=1}^p d(x_i, \hat{x}_i) - \varepsilon \right) \quad (15)$$

admits a unique optimal solution, and by [52, Corollary 28.1.1] this minimizer must coincide with the unique optimal solution x^* of problem (P''_{Vec}) . Given γ^* , we can thus solve (15) instead of (P''_{Vec}) . This is attractive from a computational point of view because \mathcal{C} is rectangular, whereby problem (15) can be simplified to

$$-\varepsilon\gamma^* + \sum_{i=1}^p \min_{x_i \in \mathcal{C}_i} \{x_i^2 + \gamma^* d(x_i, \hat{x}_i)\} = -\varepsilon\gamma^* + \sum_{i=1}^p \min_{x_i \in [0, \hat{x}_i]} \{x_i^2 + \gamma^* d(x_i, \hat{x}_i)\}.$$

Therefore, it suffices to solve the following simple univariate minimization problem for each $i = 1, \dots, p$.

$$\min_{x_i \in [0, \hat{x}_i]} x_i^2 + \gamma^* d(x_i, \hat{x}_i) \quad (16)$$

If $\hat{x}_i = 0$, then $(0, 0) \in \text{dom}(d)$ by Assumption 3(a), and hence $d(0, 0) = d(\hat{x}_i, \hat{x}_i) = 0$. In this case, $x_i^* = 0$ is the only feasible—and thus unique optimal—solution of (16). Assume next that $\hat{x}_i > 0$. In this case we need to prove that x_i^* falls within the open interval $(0, \hat{x}_i)$ and satisfies (14). We will first show that $x_i^* > 0$. From the discussion after Assumption 4 we know that $d_{x_i^*}$

can evaluate to $+\infty$ only at 0. If $d_{\hat{x}_i}(0) = +\infty$, then we trivially have $x_i^* > 0$. Assume next that $d_{\hat{x}_i}(0) < +\infty$. By Assumption 2(b), $d_{\hat{x}_i}$ is continuous and $d_{\hat{x}_i}(0) > 0$. There exists a threshold $\delta > 0$ such that $d_{\hat{x}_i}(a) \geq \delta$ for all sufficiently small $a \in [0, \hat{x}_i]$. In addition, as the function $a^2 + \gamma^* d(a, \hat{x}_i)$ is convex and differentiable in a by virtue of Assumption 4, we have

$$\begin{aligned} 0^2 + \gamma^* d(0, \hat{x}_i) &\geq a^2 + \gamma^* d(a, \hat{x}_i) + (2a + \gamma^* d'_{\hat{x}_i}(a))(0 - a) \\ &> a^2 + \gamma^* d(a, \hat{x}_i) - 2a^2 + \frac{a\gamma^* d(a, \hat{x}_i)}{\hat{x}_i} \\ &\geq a^2 + \gamma^* d(a, \hat{x}_i) - 2a^2 + \frac{a\gamma^* \delta}{\hat{x}_i} \end{aligned}$$

for all sufficiently small $a \geq 0$. Here, the second inequality follows from Lemma 6, and the third inequality holds because $d_{\hat{x}_i}(a) \geq \delta$ for all sufficiently small $a \geq 0$. This reasoning implies that

$$\gamma^* d(0, \hat{x}_i) > a^2 + \gamma^* d(a, \hat{x}_i) - 2a^2 + \frac{a\gamma^* \delta}{\hat{x}_i} > a^2 + \gamma^* d(a, \hat{x}_i) \quad (17)$$

for all sufficiently small $a \geq 0$. Thus, small $a > 0$ are strictly preferable to 0, that is, $x_i^* > 0$.

Next, we prove that $x_i^* < \hat{x}_i$. As the differentiable function $d_b(a)$ is non-negative and attains its minimum 0 at $a = b$, we may conclude that its derivative $d'_b(a)$ converges to 0 as a tends to b . For any $a < b$ sufficiently close to b we thus have $(b - a)(2a + \gamma^* d'_b(a)) > 0$. As $a^2 + \gamma^* d(a, b)$ is convex in a on $[0, b]$, this ensures that

$$b^2 + \gamma^* d(b, b) \geq a^2 + \gamma^* d(a, b) + (b - a)(2a + \gamma^* d'_b(a)) > a^2 + \gamma^* d(a, b).$$

Hence, any $a < b$ sufficiently close to b is strictly preferable to b . Setting $b = \hat{x}_i$, we thus find $x_i^* < \hat{x}_i$.

Finally, note that since $x_i^* \in (0, \hat{x}_i)$, the constraints of problem (16) are not binding at optimality. Thus, the minimizer of (16) is uniquely determined by the problem's first-order optimality condition (14).

It remains to be shown that γ^* is unique. As $0 \notin \text{Fea}(\text{P}_{\text{Vec}})$ thanks to Assumption 3(b), there exists at least one $i = 1, \dots, p$ with $x_i^* > 0$, and hence $\hat{x}_i > 0$. Since $d_{\hat{x}_i}$ is differentiable on \mathbb{R}_{++} , equation (14) implies

$$\gamma^* = -\frac{2x_i^*}{d'_{\hat{x}_i}(x_i^*)}.$$

Hence, γ^* is unique because x_i^* is unique. Note also that γ^* is the Lagrange multiplier associated with the constraint $\sum_{i=1}^p d(x_i, \hat{x}_i) \leq \varepsilon$ in problem $(\text{P}_{\text{Vec}}'')$. As strong duality holds and $\gamma^* > 0$, we have

$$\sum_{i=1}^p d(x_i^*, \hat{x}_i) - \varepsilon = 0$$

by complementary slackness. Using the definition (8) of the eigenvalue map s , we then obtain

$$\sum_{i=1}^p d(s(\gamma^*, \hat{x}_i), \hat{x}_i) - \varepsilon = 0.$$

This observation completes the proof. \square

A.2.1. Properties of s and γ^*

We first provide a detailed analysis of the nonlinear equation that defines the eigenvalue map s .

Lemma 7 (Properties of s). *If Assumptions 2 and 4 hold, then the the following hold.*

- (i) *If $\gamma > 0$ and $b > 0$, then the equation $0 = 2a + \gamma d'_b(a)$ admits a unique solution in $(0, b)$. Hence, the eigenvalue map $s(\gamma, b)$ is well-defined on \mathbb{R}_+^2 .*
- (ii) *If $b > 0$, then $s_b(\gamma) = s(\gamma, b)$ is continuous and strictly increasing on \mathbb{R}_+ and differentiable on \mathbb{R}_{++} .*
- (iii) *If $b > 0$, then $\lim_{\gamma \downarrow 0} s_b(\gamma) = 0$ and $\lim_{\gamma \rightarrow \infty} s_b(\gamma) = b$.*

Recall that, for and fixed $\gamma > 0$, the function $s_\gamma(b)$ shrinks the input b in the sense that $s_\gamma(b) \leq b$. Lemma 7 further shows that, for any fixed $b > 0$, $s_b(\gamma)$ strictly increases from 0 to b as γ grows. Therefore, we can interpret γ as an inverse shrinkage intensity.

Proof of Lemma 7. Assertion (i) follows directly from the proof of Proposition 10 and is thus not repeated.

Next, we prove assertion (ii). Recall from Assumption 4 that d_b is twice continuously differentiable on \mathbb{R}_{++} . Thus, the function $H(\gamma, a) = 2a + \gamma d'_b(a)$ is continuously differentiable on \mathbb{R}_{++}^2 . Assumption 4 further stipulates that d_b is convex on $[0, b]$. Hence, $H(\gamma, a)$ is strictly increasing in a in the sense that

$$\frac{\partial H(\gamma, a)}{\partial a} = 2 + \gamma d''_b(a) \geq 2 > 0 \quad \forall a \in (0, b].$$

As $s_b(\gamma) \in (0, b)$ by assertion (i), the implicit function theorem ensures that $s_b(\gamma)$ is differentiable (and in particular continuous) at any $\gamma > 0$. It remains to be shown that $s_b(\gamma)$ is continuous at 0. Given any $\epsilon > 0$ and as $s_b(0) = 0$ by definition, we thus need to show that there is $\delta > 0$ such that $s_b(\gamma) \leq \epsilon$ for all $\gamma \in [0, \delta]$. As $s_b(\gamma) \in (0, b)$ for all $\gamma, b > 0$, we may assume without loss of generality that $\epsilon \in (0, b)$. By Lemma 6, we have $d'_b(\epsilon) < 0$, which guarantees that $\delta = -2\epsilon/d'_b(\epsilon)$ is positive. For any $\gamma \in [0, \delta]$, we thus obtain

$$s_b(\gamma) = -\frac{\gamma d'_b(s_b(\gamma))}{2} \leq \frac{\epsilon d'_b(s_b(\gamma))}{d'_b(\epsilon)},$$

where the equality follows from the definition of s_b in (8), and the inequality follows from the definition of δ . This confirms that $s_b(\gamma) \leq \epsilon$. Suppose to the contrary that $s_b(\gamma) > \epsilon$. Then the above inequality implies $d'_b(s_b(\gamma)) < d'_b(\epsilon)$. As d'_b is non-decreasing by virtue of the convexity of d_b , this in turn leads to the contradiction $s_b(\gamma) > \epsilon$. Thus, $s_b(\gamma) \leq \epsilon$ for all $\gamma \in [0, \delta]$. We conclude that $s_b(\gamma)$ is indeed continuous at 0.

To show that $s_b(\gamma)$ is strictly increasing on \mathbb{R}_{++} , recall that $s_b(\gamma)$ is differentiable on \mathbb{R}_{++} . We may thus differentiate both sides of the equation $0 = 2s_b(\gamma) + \gamma d'_b(s_b(\gamma))$ with respect to γ to obtain

$$0 = 2s'_b(\gamma) + d'_b(s_b(\gamma)) + \gamma d''_b(s_b(\gamma))s'_b(\gamma).$$

Rearranging terms then yields

$$s'_b(\gamma) = -\frac{d'_b(s_b(\gamma))}{2 + \gamma d''_b(s_b(\gamma))}, \tag{18}$$

which is strictly positive because $d'_b(s_b(\gamma)) < 0$ thanks to Lemma 6 and $d''_b(s_b(\gamma)) \geq 0$ thanks to the convexity of d_b on $[0, b]$. Hence, $s_b(\gamma)$ is strictly increasing on \mathbb{R}_+ . This completes the proof of assertion (ii).

It remains to prove assertion (iii). The continuity of $s_b(\gamma)$ at $\gamma = 0$ has already been established in assertion (ii). As $s_b(\gamma) \in (0, b)$ is strictly increasing in γ , it is clear that, as γ tends

to infinity, $s_b(\gamma)$ has a well-defined limit not larger than b . By the definition of s_b in (8), we further have

$$\frac{2s_b(\gamma)}{\gamma} + d'_b(s_b(\gamma)) = 0 \quad \forall \gamma > 0.$$

Driving γ to infinity and recalling that $s_b(\gamma) \in (0, b)$ for all $\gamma > 0$ thus shows that

$$0 = \lim_{\gamma \rightarrow \infty} d'_b(s_b(\gamma)) = d'_b \left(\lim_{\gamma \rightarrow \infty} s_b(\gamma) \right),$$

where the second equality follows from the continuity of d'_b on \mathbb{R}_{++} . Note that $\lim_{\gamma \rightarrow \infty} s_b(\gamma)$ exists and falls within the interval $(0, b]$ because s_b is a strictly increasing function mapping \mathbb{R}_+ to $(0, b)$. These arguments imply that the limit must be a root of d'_b within $(0, b]$. Lemma 6 implies that d'_b has *no* root in the open interval $(0, b)$. We may thus conclude that $\lim_{\gamma \rightarrow \infty} s_b(\gamma)$ must coincide with b . As a sanity check, one readily verifies that $0 = d'_b(b)$ because $d_b(a)$ attains its minimum of 0 at $a = b$. Thus, assertion (iii) follows. \square

We now prove that the function $F(\gamma) = \sum_{i=1}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon$ has one and only one root. By the proof of Proposition 10, this root must coincide with the unique optimal solution γ^* of the problem dual to (P_{Vec}) .

Lemma 8. *If Assumptions 2, 3 and 4 hold, then the equation $F(\gamma) = 0$ has a unique root, which is positive.*

Proof of Lemma 8. Recall that $s(\gamma, 0) = 0$ by the definition of s in (8). Recall also that if $\hat{x}_i = 0$, then $d(s(\gamma, \hat{x}_i), \hat{x}_i) = d(0, 0) = 0$ by virtue of Assumptions 2 and 3(a). Therefore, vanishing components of \hat{x} do not contribute to the function $F(\gamma)$. In addition, Assumption 3(b) ensures that there exists at least one $i \in \{1, \dots, p\}$ with $x_i^* > 0$ and hence also with $\hat{x}_i > 0$. For these reasons, we henceforth assume without loss of generality that $\hat{x}_i > 0$ for all $i = 1, \dots, p$. By Lemma 7(ii), $s(\gamma, \hat{x}_i)$ constitutes a continuous real-valued function of $\gamma \in \mathbb{R}_+$. Similarly, by Assumption 2(b), $d(x_i, \hat{x}_i)$ constitutes a continuous extended real-valued function of $x_i \in \mathbb{R}_+$. Therefore, the extended real-valued function $F(\gamma)$ is continuous on \mathbb{R}_+ . Assumption 3(b) implies that $F(0) = \sum_{i=1}^p d(0, \hat{x}_i) - \varepsilon > 0$. Recall now from Lemma 7(iii) that $s(\gamma, \hat{x}_i)$ converges to \hat{x}_i as γ tends to infinity. By the continuity of $d(x_i, \hat{x}_i)$ in x_i we thus have

$$\lim_{\gamma \rightarrow \infty} F(\gamma) = \sum_{i=1}^p d(\hat{x}_i, \hat{x}_i) - \varepsilon = -\varepsilon < 0.$$

All of this implies that the equation $F(\gamma) = 0$ has at least one positive root. In the remainder we prove that this root is unique. As $\hat{x}_i > 0$, Lemma 7 implies that $s(\gamma, \hat{x}_i)$ strictly increases from 0 (at $\gamma = 0$) to \hat{x}_i (as γ tends to infinity). Lemma 6 further implies that $d_{\hat{x}_i}$ is strictly decreasing on $[0, \hat{x}_i]$. Thus, the composite function $d(s(\gamma, \hat{x}_i), \hat{x}_i)$ is strictly decreasing in γ for every i . This readily shows that $F(\gamma)$ is strictly decreasing in γ throughout \mathbb{R}_+ , thus implying that the equation $F(\gamma) = 0$ has only one root. \square

We are now ready to prove Proposition 3.

Proof of Proposition 3. The proof is a direct consequence of Propositions 9 and 10 and Lemmas 7 and 8. \square

APPENDIX B. PROOFS OF SECTION 3.3

Proof of Proposition 4. In view of the proof of Lemma 8, it only remains to be shown that $F(\gamma)$ is differentiable at any $\gamma > 0$. Towards that end, recall that vanishing components of \hat{x} do not contribute to $F(\gamma)$ such that

$$F(\gamma) = \sum_{i=1}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon = \sum_{\substack{i=1: \\ \hat{x}_i > 0}}^p d(s(\gamma, \hat{x}_i), \hat{x}_i) - \varepsilon.$$

For any fixed $\hat{x}_i > 0$, $s(\gamma, \hat{x}_i)$ is differentiable with respect to $\gamma \in \mathbb{R}_{++}$ by Lemma 7(ii), and $d(x, \hat{x}_i)$ is differentiable with respect to $x \in \mathbb{R}_{++}$ by Assumption 4. Therefore, $F(\gamma)$ is differentiable at any $\gamma > 0$. \square

From the proof of Proposition 10 we know that the problem dual to (P_{Vec}) has a unique optimal solution γ^* . Thus, γ^* can be viewed as a function $\gamma^*(\varepsilon)$ of the radius $\varepsilon > 0$ of the divergence ball (2).

Lemma 9 (Monotonicity of γ^*). *If Assumptions 2, 3 and 4 hold, then $\gamma^*(\varepsilon)$ is non-increasing on $(0, \bar{\varepsilon})$.*

Proof of Lemma 9. The proof of Proposition 10 implies that $\gamma^*(\varepsilon)$ is the unique maximizer of the problem dual to (P_{Vec}) . By inverting its objective function, this problem can be recast as the minimization problem

$$\min_{\gamma > 0} \varepsilon \gamma + G(\gamma), \quad (19)$$

where the function $G : \mathbb{R}_{++} \rightarrow \overline{\mathbb{R}}$ is defined through

$$G(\gamma) = - \sum_{\substack{i=1: \\ \hat{x}_i > 0}}^p \min_{x_i \in [0, \hat{x}_i]} \{x_i^2 + \gamma d(x_i, \hat{x}_i)\} = - \sum_{\substack{i=1: \\ \hat{x}_i > 0}}^p \left((s_{\hat{x}_i}(\gamma))^2 + \gamma d_{\hat{x}_i}(s_{\hat{x}_i}(\gamma)) \right).$$

Note also that the non-negativity constraint on γ in (19) is strict because $\gamma = 0$ cannot be optimal, or, dually, because the constraint in (P_{Vec}) must be binding at optimality for $\varepsilon < \bar{\varepsilon}$. By construction, $G(\gamma)$ constitutes a pointwise maximum of multiple linear functions and is, therefore, convex. Next, select $\varepsilon_1, \varepsilon_2 \in (0, \bar{\varepsilon}]$ with $0 < \varepsilon_1 < \varepsilon_2$, and introduce the notational shorthands $\gamma_1 = \gamma^*(\varepsilon_1)$ and $\gamma_2 = \gamma^*(\varepsilon_2)$. By the optimality of γ_1 and γ_2 in problem (19) at ε_1 and ε_2 , there exist subgradients $g_1 \in \partial G(\gamma_1)$ and $g_2 \in \partial G(\gamma_2)$ satisfying the first-order optimality conditions $\varepsilon_1 + g_1 = 0$ and $\varepsilon_2 + g_2 = 0$, respectively. Since $G(\gamma)$ is convex, its subdifferential is monotone, whereby $(\gamma_2 - \gamma_1)(g_2 - g_1) \geq 0$. Together with the first-order optimality conditions, this implies that $(\gamma_2 - \gamma_1)(\varepsilon_1 - \varepsilon_2) \geq 0$. As $\varepsilon_1 < \varepsilon_2$, we may thus conclude that $\gamma_2 \leq \gamma_1$. Hence, the claim follows. \square

Proof of Proposition 5. Note that $x_i^*(\varepsilon) = s(\gamma^*(\varepsilon), \hat{x}_i)$ for every $\varepsilon \in (0, \bar{\varepsilon})$ thanks to Proposition 3, and recall that $x_i^*(\bar{\varepsilon}) = 0$ by definition. We aim to show that $x_i^*(\varepsilon)$ is non-increasing on $[0, \bar{\varepsilon}]$ and that $\lim_{\varepsilon \uparrow \bar{\varepsilon}} x_i^*(\varepsilon) = 0$. To this end, note first that both claims are trivially satisfied if $\hat{x}_i = 0$, in which case $x_i^*(\varepsilon) = 0$ for all $\varepsilon \in (0, \bar{\varepsilon})$ thanks to Proposition 3 and our conventions that $x_i^*(0) = \hat{x}_i$ and $x_i^*(\bar{\varepsilon}) = 0$. Assume next that $\hat{x}_i > 0$. Recall that $\gamma^*(\varepsilon)$ is non-increasing on $(0, \bar{\varepsilon})$ thanks to Lemma 9, while $s_{\hat{x}_i}(x_i) = s(x_i, \hat{x}_i)$ is strictly increasing on \mathbb{R}_+ thanks to Lemma 7(ii), which applies because $\hat{x}_i > 0$. Therefore, $x_i^*(\varepsilon) = s(\gamma^*(\varepsilon), \hat{x}_i)$ is non-increasing on $(0, \bar{\varepsilon})$. We also have $x_i^*(\varepsilon) \in (0, \hat{x}_i)$ for all $\varepsilon \in (0, \bar{\varepsilon})$ thanks to Proposition 3, and we have

$x_i^*(0) = \hat{x}_i$ and $x_i^*(\bar{\varepsilon}) = 0$ by definition. All of this readily implies that $x_i^*(\varepsilon)$ is non-increasing on $[0, \bar{\varepsilon}]$. In order to prove that $\lim_{\varepsilon \uparrow \bar{\varepsilon}} x_i^*(\varepsilon) = 0$, note first that $\lim_{\varepsilon \uparrow \bar{\varepsilon}} x_i^*(\varepsilon)$ must exist because $x_i^*(\varepsilon)$ is non-negative as well as non-increasing in ε . Next, recall from Lemma 6 that the function $d_{\hat{x}_i}(x_i) = d(x_i, \hat{x}_i)$ is strictly decreasing on $(0, \hat{x}_i)$. In fact, this monotonicity property extends to $[0, \hat{x}_i]$ because $d_{\hat{x}_i}$ is continuous thanks to Assumption 2(b). We then choose an arbitrary tolerance $\delta > 0$ and assume without loss of generality that δ is smaller than the smallest non-vanishing component of \hat{x} . Next, consider a vector $x \in \mathbb{R}_+^p$ defined through $x_i = 0$ if $\hat{x}_i = 0$ and $x_i = \delta$ if $\hat{x}_i > 0$, $i = 1, \dots, p$, and set $\varepsilon = \sum_{i=1}^p d(x_i, \hat{x}_i)$. By construction, we have

$$\varepsilon = \sum_{i=1}^p d(x_i, \hat{x}_i) < \sum_{i=1}^p d(0, \hat{x}_i) = \bar{\varepsilon},$$

where the strict inequality holds because \hat{x} has at least one strictly positive component and because $d(x_i, \hat{x}_i) < d(0, \hat{x}_i)$ whenever $\hat{x}_i > 0$ thanks to the monotonicity properties of d established above. Hence, x is feasible in P_{Vec} , and ε is consistent with Assumption 3(b). In addition, one readily verifies that the objective function value of x satisfies $\|x\|_2^2 \leq p\delta^2$. By the optimality of $x^*(\varepsilon)$ in P_{Vec} , we thus find

$$x_i^*(\varepsilon)^2 \leq \|x^*(\varepsilon)\|_2^2 \leq p\delta^2 \quad \forall i = 1, \dots, p.$$

Thus, for any sufficiently small $\delta > 0$ there exists $\varepsilon > 0$ with $x_i^*(\varepsilon) \leq \sqrt{p}\delta$. As $x_i^*(\varepsilon)$ is non-increasing on $[0, \bar{\varepsilon}]$, this implies indeed that $\lim_{\varepsilon \uparrow \bar{\varepsilon}} x_i^*(\varepsilon) = 0$. It remains to be shown that X^* constitutes a shrinkage estimator. This is now evident, however, because $\hat{\Sigma} = \hat{V} \text{Diag}(\hat{x}) \hat{V}^\top = \hat{V} \text{Diag}(x^*(0)) \hat{V}^\top$. \square

Proof of Lemma 1. Throughout this proof we fix any $\gamma > 0$. We first aim to show that the function

$$K(b) = \frac{1}{b} \frac{\partial d(s(\gamma, b), b)}{\partial a}$$

is non-decreasing on \mathbb{R}_{++} . To this end, note that $d(a, b)$ is twice continuously differentiable on \mathbb{R}_{++}^2 by Assumption 5. Using the implicit function theorem as in Lemma 7, one can thus show that $s(\gamma, b)$ is differentiable with respect to b and that $s(\gamma, b) \in (0, b)$ for every $b > 0$. Recall also that $-\frac{2}{\gamma} s(\gamma, b) = \frac{\partial d}{\partial a}(s(\gamma, b), b)$ by the definition of s in (8).

$$-\frac{2}{\gamma} \frac{\partial s(\gamma, b)}{\partial b} = \frac{d}{db} \left(\frac{\partial d(s(\gamma, b), b)}{\partial a} \right) = \frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b} + \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2} \frac{\partial s(\gamma, b)}{\partial b}. \quad (20)$$

This in turn implies that

$$\frac{\partial s(\gamma, b)}{\partial b} = - \left(\frac{2}{\gamma} + \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2} \right)^{-1} \frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b}, \quad (21)$$

which is well-defined because $\gamma > 0$ and $d(\cdot, b)$ is convex by Assumption 4. We then find

$$\frac{dK(b)}{db} = -\frac{1}{b^2} \frac{\partial d(s(\gamma, b), b)}{\partial a} + \frac{1}{b} \frac{d}{db} \left(\frac{\partial d(s(\gamma, b), b)}{\partial a} \right).$$

The second term on the right hand side of the above expression satisfies

$$\begin{aligned} \frac{1}{b} \frac{d}{db} \left(\frac{\partial d(s(\gamma, b), b)}{\partial a} \right) &= -\frac{2}{b\gamma} \frac{\partial s(\gamma, b)}{\partial b} = \frac{2}{b\gamma} \left(\frac{2}{\gamma} + \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2} \right)^{-1} \frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b} \\ &= \frac{2}{b} \left(\frac{\frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b}}{2 - \frac{2s(\gamma, b)}{\frac{\partial d(s(\gamma, b), b)}{\partial a}} \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2}} \right) = \frac{1}{b} \left(\frac{\frac{\partial d(s(\gamma, b), b)}{\partial a} \frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b}}{\frac{\partial d(s(\gamma, b), b)}{\partial a} - s(\gamma, b) \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2}} \right), \end{aligned}$$

where the first and the second equalities follow from (20) and (21), respectively, and the third equality follows from the defining equation of s in (8). Combining the last two equations finally yields

$$\frac{dK(b)}{db} = -\frac{1}{b^2} \frac{\partial d(s(\gamma, b), b)}{\partial a} \left(1 - \frac{b \frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b}}{\frac{\partial d(s(\gamma, b), b)}{\partial a} - s(\gamma, b) \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2}} \right).$$

Recall now that $\frac{\partial d(a, b)}{\partial a} < 0$ for every $a \in (0, b)$ thanks to Lemma 6 and that $s(\gamma, b) \in (0, b)$ thanks to Lemma 7. This implies that the derivative of $K(b)$ is non-negative if and only if

$$\frac{\partial^2 d(s(\gamma, b), b)}{\partial a \partial b} \geq \frac{1}{b} \left(\frac{\partial d(s(\gamma, b), b)}{\partial a} - s(\gamma, b) \frac{\partial^2 d(s(\gamma, b), b)}{\partial a^2} \right). \quad (22)$$

Assumption 5 guarantees that (22) holds indeed for all $b > 0$. Hence, $K(b)$ is a non-decreasing function.

We now prove the desired inequality. By the defining equation of s in (8) we have

$$-2\gamma b_1 s(\gamma, b_2) = b_1 \frac{\partial d(s(\gamma, b_2), b_2)}{\partial a} \geq b_2 \frac{\partial d(s(\gamma, b_1), b_1)}{\partial a} = -2\gamma b_2 s(\gamma, b_1)$$

for any $b_2 \geq b_1 > 0$, where the inequality follows from the monotonicity of K established above. This implies that $s(\gamma, b_2)/s(\gamma, b_1) \leq b_2/b_1$ for all $b_1, b_2 \in \mathbb{R}_{++}$ with $b_2 \geq b_1$. Hence, the claim follows. \square

Proof of Proposition 7. Throughout the proof we use the shorthands $x_{i,n}^* = \lambda_i(X_n^*)$ and $\hat{x}_{i,n} = \lambda_i(\hat{\Sigma}_n)$ for all $i = 1, \dots, p$ and $n \in \mathbb{N}$. By the strong consistency assumption, $\hat{\Sigma}_n$ converges almost surely to Σ_0 . Fix now temporarily a particular realization of the uncertainties, for which $\hat{\Sigma}_n$ converges deterministically to Σ_0 . In this case, $\hat{x}_{i,n}$ converges to $\lambda_i(\Sigma_0)$ because the eigenvalue map λ_i is continuous [4, Corollary VI.1.6], and the sequence $\{x_{i,n}^*\}_{n \in \mathbb{N}}$ is bounded by Lemma 5. Thus, any convergent subsequence $\{x_{i,n_k}^*\}_{k \in \mathbb{N}}$ satisfies

$$\lim_{k \rightarrow \infty} x_{i,n_k}^* \in [0, \lim_{k \rightarrow \infty} \hat{x}_{i,n_k}] = [0, \lim_{n \rightarrow \infty} \hat{x}_{i,n}] = [0, \lambda_i(\Sigma_0)].$$

In addition, we have

$$d(x_{i,n_k}^*, \hat{x}_{i,n_k}) \leq \sum_{j=1}^p d(x_{j,n_k}^*, \hat{x}_{j,n_k}) = D(X_{n_k}^*, \hat{\Sigma}_{n_k}) \leq \varepsilon_{n_k} \quad \forall k \in \mathbb{N},$$

where the first equality holds because of Assumptions 2(a) and 2(b) and because $X_{n_k}^*$ and $\hat{\Sigma}_{n_k}$ share the same eigenvectors. The second inequality follows from Proposition 1(ii), which ensures that $X_{n_k}^*$ is feasible in problem (P_{Mat}). As ε_{n_k} converges to 0 and as d is continuous on $\mathbb{R}_+ \times \mathbb{R}_{++}$, the above implies that

$$d(\lim_{k \rightarrow \infty} x_{i,n_k}^*, \lambda_i(\Sigma_0)) = d(\lim_{k \rightarrow \infty} x_{i,n_k}^*, \lim_{k \rightarrow \infty} \hat{x}_{i,n_k}) = \lim_{k \rightarrow \infty} d(x_{i,n_k}^*, \hat{x}_{i,n_k}) = 0.$$

Recall now from Assumption 2 that d satisfies the identity of indiscernibles. Thus we find $\lim_{k \rightarrow \infty} x_{i,n_k}^* = \lambda_i(\Sigma_0)$. This shows that every convergent subsequence of the bounded sequence $\{x_{i,n}^*\}_{n \in \mathbb{N}}$ must have the same limit $\lambda_i(\Sigma_0)$. By [1, Exercise 2.5.5], the eigenvalue $x_{i,n}^*$ therefore converges to $\lambda_i(\Sigma_0)$. This reasoning applies to every uncertainty realization under which $\hat{\Sigma}_n$ converges to Σ_0 . As $\hat{\Sigma}_n$ converges almost surely to Σ_0 , we have thus shown that $x_{i,n}^*$ converges

almost surely to $\lambda_i(\Sigma_0)$. This in turn implies that

$$\begin{aligned} \mathbb{P}[\lim_{n \rightarrow \infty} \|X_n^* - \Sigma_0\|_F = 0] &\geq \mathbb{P}[\lim_{n \rightarrow \infty} (\|X_n^* - \widehat{\Sigma}_n\|_F + \|\widehat{\Sigma}_n - \Sigma_0\|_F) = 0] \\ &= \mathbb{P}[\lim_{n \rightarrow \infty} (\|x_n^* - \hat{x}_n\|_2 + \|\widehat{\Sigma}_n - \Sigma_0\|_F) = 0] \\ &\geq \mathbb{P}[\lim_{n \rightarrow \infty} (\|x_n^* - \lambda(\Sigma_0)\|_2 + \|\lambda(\Sigma_0) - \hat{x}_n\|_2 + \|\widehat{\Sigma}_n - \Sigma_0\|_F) = 0] = 1, \end{aligned}$$

where both inequalities hold thanks to the triangle inequality, the first equality follows from Theorem 1, which ensures that X_n^* and $\widehat{\Sigma}_n$ share the same eigenvectors, and the second equality exploits the almost sure convergence of x_n^* and \hat{x}_n to $\lambda(\Sigma_0)$ established above and the almost sure convergence of $\widehat{\Sigma}_n$ to Σ_0 . This shows that X_n^* converges almost surely to Σ_0 and therefore completes the proof. \square

From now on we use $\|X\|_*$ to denote the nuclear norm of $X \in \mathbb{S}^p$ (*i.e.*, the sum of all singular values of X), which is the norm dual to the spectral norm $\|X\|$ (*i.e.*, the largest singular value of X). The proof of Proposition 8 relies on the following well-known result from high-dimensional statistics.

Lemma 10 ([67, Theorem 6.5]). *Under the assumptions of Proposition 8, there exists a universal constant $c_0 > 0$ independent of \mathbb{P} such that*

$$\mathbb{P}^n \left[\|\widehat{\Sigma}_n - \Sigma_0\| \leq \rho(p, n, \eta) \right] \geq 1 - \frac{\eta}{2}$$

for every $n \in \mathbb{N}$ and $\eta \in (0, 1)$, where

$$\rho(p, n, \eta) = c_0 \sigma^2 \left(\frac{p + \log \eta^{-1}}{n} + \sqrt{\frac{p + \log \eta^{-1}}{n}} \right).$$

Proof of Proposition 8. For any divergence function D from Table 1 we will prove that there exist a constant $c > 0$ and a function $n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ that may depend on \mathbb{P} via σ^2 and $\lambda_1(\Sigma_0)$ such that

$$\mathbb{P}^n \left[D(\Sigma_0, \widehat{\Sigma}_n) \leq c \|\Sigma_0 - \widehat{\Sigma}_n\| \right] \geq 1 - \frac{\eta}{2} \quad (23)$$

for all $n \geq n_{\min}(p, \eta)$ and $\eta \in (0, 1)$. Indeed, if such an inequality holds, then Lemma 10 and the union bound imply that $\mathbb{P}^n[D(\Sigma_0, \widehat{\Sigma}_n) \leq c\rho(p, n, \eta)] \geq 1 - \eta$. The claim then follows by setting $\varepsilon_{\min}(p, n, \eta) = c\rho(p, n, \eta)$.

Stein, Inverse Stein and Symmetrized Stein Divergences: Note that the sum of the Stein and inverse Stein divergences equals twice the symmetrized Stein divergence. Recall also that all divergences are non-negative. Thus, if the ball of radius ε with respect to the symmetrized Stein divergence contains Σ_0 with probability at least $1 - \eta$, then the ball of radius 2ε with respect to the Stein or inverse Stein divergence contains Σ_0 with probability at least $1 - \eta$. It thus suffices to focus on the symmetrized Stein divergence. Suppose now that the smallest eigenvalue of $\widehat{\Sigma}_n$ is no smaller than half of the smallest eigenvalue of Σ_0 . As $\Sigma_0 \succ 0$, this implies in particular that $\widehat{\Sigma}_n$ is positive definite and that $\widehat{\Sigma}_n^{-1}$ exists. Rewriting the symmetrized Stein divergence as $\frac{1}{2} \text{Tr}[(\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1})(\widehat{\Sigma}_n - \Sigma_0)]$, we may then use the matrix Hölder's inequality to obtain

$$\text{Tr}[(\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1})(\widehat{\Sigma}_n - \Sigma_0)] \leq \|\Sigma_0 - \widehat{\Sigma}_n\| \|\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1}\|_*.$$

In the following we use $x_i = \lambda_i(\Sigma_0)$ and $\hat{x}_{i,n} = \lambda_i(\widehat{\Sigma}_n)$ to denote i -th smallest population and sample eigenvalues for $i = 1, \dots, p$, respectively. By the definitions of the nuclear and spectral

norms, we then have

$$\begin{aligned}
\|\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1}\|_* &\leq p\|\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1}\| \\
&= p \max \left\{ \lambda_p(\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1}), -\lambda_1(\Sigma_0^{-1} - \widehat{\Sigma}_n^{-1}) \right\} \\
&\leq p \max \left\{ \lambda_p(\Sigma_0^{-1}) - \lambda_1(\widehat{\Sigma}_n^{-1}), \lambda_p(\widehat{\Sigma}_n^{-1}) - \lambda_1(\Sigma_0^{-1}) \right\} \\
&= p \max \left\{ \frac{1}{x_1} - \frac{1}{\widehat{x}_{p,n}}, \frac{1}{\widehat{x}_{1,n}} - \frac{1}{x_p} \right\} \\
&\leq p \max \left\{ \frac{1}{x_1}, \frac{1}{\widehat{x}_{1,n}} \right\} \leq \frac{2p}{x_1},
\end{aligned}$$

where the first equality holds because the singular values of a symmetric matrix coincide with the absolute values of the eigenvalues of that matrix. The second inequality follows from a classical result by Weyl, which asserts that $\lambda_1(A+B) \leq \lambda_1(A) + \lambda_p(B) \leq \lambda_p(A+B)$ for any $A, B \in \mathbb{S}^p$, and the second equality holds because $\lambda_i(A^{-1}) = 1/\lambda_{p-i+1}(A)$ for any $i = 1, \dots, p$ and $A \in \mathbb{S}_{++}^p$. The third inequality exploits our assumption that all population and sample eigenvalues are strictly positive, and the last inequality follows from the assumption that $\widehat{x}_{1,n} \geq x_1/2$. We have thus shown that if $\widehat{x}_{1,n} \geq x_1/2$, then $D(\Sigma_0, \widehat{\Sigma}_n) \leq \frac{p}{x_1}\|\Sigma_0 - \widehat{\Sigma}_n\|$. Hence, we find

$$\mathbb{P}^n \left[D(\Sigma_0, \widehat{\Sigma}_n) \leq \frac{p}{x_1}\|\Sigma_0 - \widehat{\Sigma}_n\| \right] \geq \mathbb{P}^n \left[\widehat{x}_{1,n} \geq \frac{x_1}{2} \right].$$

As $\widehat{x}_{1,n} \geq x_1 - \|\Sigma_0 - \widehat{\Sigma}_n\|$ by virtue of Weyl's inequality and by Lemma 10, the last probability satisfies

$$\mathbb{P}^n \left[\widehat{x}_{1,n} \geq \frac{x_1}{2} \right] \geq \mathbb{P}^n \left[\|\Sigma_0 - \widehat{\Sigma}_n\| \leq \frac{x_1}{2} \right] \geq \mathbb{P}^n \left[\|\Sigma_0 - \widehat{\Sigma}_n\| \leq \rho(p, n, \eta) \right] \geq 1 - \frac{\eta}{2} \quad (24)$$

whenever $x_1/2 \geq \rho(p, n, \eta)$. By the definition of $\rho(p, n, \eta)$, a sufficient condition for this inequality to hold is

$$n \geq n_{\min}(p, \eta) = \max \left\{ 1, \frac{16c_0^2\sigma^4}{x_1^2} \right\} (p + \log \eta^{-1}).$$

The above estimates imply that (23) holds for all $n \geq n_{\min}(p, \eta)$ and $\eta \in (0, 1)$ if we set $c = p/x_1$. In addition, the minimal sample size and the minimal radius of the uncertainty set satisfy $n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ and

$$\varepsilon_{\min}(p, n, \eta) = c\rho(p, n, \eta) = \frac{pc_0\sigma^2}{x_1} \left(\frac{p + \log \eta^{-1}}{n} + \sqrt{\frac{p + \log \eta^{-1}}{n}} \right) = \mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}}),$$

where the last equality holds because $n \geq p + \log \eta^{-1}$. This establishes the claim for the Stein, the inverse Stein and the symmetrized Stein divergences.

Wasserstein Divergence: From the proof of [44, Theorem 4] we know that if $\widehat{x}_{1,n} \geq \frac{x_1}{2}$, then

$$D(\Sigma_0, \widehat{\Sigma}_n) \leq \frac{1}{(\widehat{x}_{1,n} + x_1)^2} \|\Sigma_0 - \widehat{\Sigma}_n\|_{\text{F}}^2 \leq \frac{p}{(\widehat{x}_{1,n} + x_1)^2} \|\Sigma_0 - \widehat{\Sigma}_n\|^2 \leq \frac{4p}{9x_1^2} \|\Sigma_0 - \widehat{\Sigma}_n\|^2.$$

We also know from (24) that $\mathbb{P}^n[\widehat{x}_{1,n} \geq \frac{x_1}{2}] \geq 1 - \frac{\eta}{2}$ for all $n \geq \mathcal{O}(p + \log \eta^{-1})$. Thus, we have

$$\mathbb{P}^n \left[D(\Sigma_0, \widehat{\Sigma}_n) \leq \frac{4p}{9x_1^2} \|\Sigma_0 - \widehat{\Sigma}_n\|^2 \right] \geq \mathbb{P}^n \left[\widehat{x}_{1,n} \geq \frac{x_1}{2} \right] \geq 1 - \frac{\eta}{2} \quad (25)$$

for all $n \geq \mathcal{O}(p + \log \eta^{-1})$. Lemma 10 further implies that

$$\mathbb{P}^n \left[\|\Sigma_0 - \widehat{\Sigma}_n\| \leq 1 \right] \geq \mathbb{P}^n \left[\|\Sigma_0 - \widehat{\Sigma}_n\| \leq \rho(p, n, \eta) \right] \geq 1 - \frac{\eta}{2}, \quad (26)$$

whenever

$$1 \geq \rho(p, n, \eta) = c_0 \sigma^2 \left(\frac{p + \log \eta^{-1}}{n} + \sqrt{\frac{p + \log \eta^{-1}}{n}} \right).$$

A sufficient condition for this inequality to hold is that $n \geq \mathcal{O}(p + \log \eta^{-1})$. Combining (25) and (26) and using the union bound implies that there is a function $n_{\min}(p, \eta)$ that grows at most as $\mathcal{O}(p + \log \eta^{-1})$ with

$$\mathbb{P}^n \left[D(\Sigma_0, \hat{\Sigma}_n) \leq \frac{4p}{9x_1^2} \|\Sigma_0 - \hat{\Sigma}_n\| \right] \geq 1 - \eta$$

for all $n \geq n_{\min}(p, \eta)$. Thus, (23) holds for all $n \geq n_{\min}(p, \eta)$ and $\eta \in (0, 1)$ if we set $c = 4p/(9x_1^2)$. Similar calculations as in the last part of the proof reveal that $\varepsilon_{\min}(p, n, \eta) = c\rho(p, n, \eta)$ grows at most as $\mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}})$. This establishes the claim for the Wasserstein divergence.

Quadratic Divergence: Since $\|A\|_F \leq \sqrt{p}\|A\|$ for all $A \in \mathbb{S}^p$, we have

$$D(\Sigma_0, \hat{\Sigma}_n) = \|\Sigma_0 - \hat{\Sigma}_n\|_F^2 \leq p \|\Sigma_0 - \hat{\Sigma}_n\|^2.$$

From (26) we already know that $\mathbb{P}^n[\|\Sigma_0 - \hat{\Sigma}_n\| \leq 1] \geq 1 - \eta$ for all $n \geq \mathcal{O}(p + \log \eta^{-1})$. Thus, there is a function $n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ such that (23) holds for all $n \geq n_{\min}(p, \eta)$ and $\eta \in (0, 1)$ if we set $c = p$. As usual, one verifies that $\varepsilon_{\min}(p, n, \eta) = c\rho(p, n, \eta) = \mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}})$. This proves the claim for the quadratic divergence.

Weighted Quadratic Divergence: As $\text{Tr}[AB] \leq \|A\| \|B\|_* \leq p\|A\| \|B\|$ for all $A, B \in \mathbb{S}^p$, we have

$$D(\Sigma_0, \hat{\Sigma}_n) = \text{Tr}[(\Sigma_0 - \hat{\Sigma}_n)^2 \hat{\Sigma}_n^{-1}] \leq p \|(\Sigma_0 - \hat{\Sigma}_n)^2\| \|\hat{\Sigma}_n^{-1}\| \leq \frac{p}{\hat{x}_{1,n}} \|\Sigma_0 - \hat{\Sigma}_n\|^2 \leq \frac{2p}{x_1} \|\Sigma_0 - \hat{\Sigma}_n\|^2$$

whenever $\hat{x}_{1,n} \geq \frac{x_1}{2}$. Recall also that $\hat{\Sigma}_n$ is indeed invertible under this assumption. Together with (24) and (26), the above inequality implies that there exists a function $n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ such that

$$\mathbb{P}^n \left[D(\Sigma_0, \hat{\Sigma}_n) \leq \frac{2p}{x_1} \|\Sigma_0 - \hat{\Sigma}_n\| \right] \geq 1 - \eta,$$

for all $n \geq n_{\min}(p, \eta)$. Thus, (23) holds for all $n \geq n_{\min}(p, \eta)$ and $\eta \in (0, 1)$ if we set $c = 2p/x_1$. As usual, we have $\varepsilon_{\min}(p, n, \eta) = \mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}})$. This proves the claim for the weighted quadratic divergence.

Fisher-Rao Divergence: As $\log^2 x \leq x - 2 + x^{-1}$ for all $x > 0$, the Fisher-Rao divergence satisfies

$$D(X, Y) = \sum_{i=1}^p \log^2 \lambda_i(XY^{-1}) \leq \sum_{i=1}^p \left(\lambda_i(XY^{-1}) - 2 + \frac{1}{\lambda_i(XY^{-1})} \right) = \text{Tr}[XY^{-1}] - 2p + \text{Tr}[YX^{-1}]$$

for all $X, Y \in \mathbb{S}_{++}^p$, where the last expression equals twice the symmetrized Stein divergence of X and Y . We have already shown that (23) holds for symmetrized Stein divergence for all $n \geq n_{\min}(p, \eta) = \mathcal{O}(p + \log \eta^{-1})$ and $\eta \in (0, 1)$ provided that $c = \frac{p}{x_1}$. Thus, (23) must also hold for the Fisher-Rao divergence if $c = \frac{2p}{x_1}$. As usual, we have $\varepsilon_{\min}(p, n, \eta) = \mathcal{O}(pn^{-\frac{1}{2}}(p + \log \eta^{-1})^{\frac{1}{2}})$. This proves the claim for the Fisher-Rao divergence. \square

APPENDIX C. VERIFICATION OF THE MINIMAX PROPERTY

Proposition 11. *All the divergences listed in Table 1 satisfy Assumption 1.*

Proof of Proposition 11. Our goal is to prove the minimax equality

$$\min_{X \in \mathbb{S}_+^p} \max_{\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})} \text{Tr}[X^2] - 2\langle \Sigma, X \rangle = \max_{\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})} \min_{X \in \mathbb{S}_+^p} \text{Tr}[X^2] - 2\langle \Sigma, X \rangle. \quad (27)$$

If D is the Kullback-Leibler, Fisher-Rao, inverse Stein, symmetrized Stein or weighted quadratic divergence and if $\widehat{\Sigma}$ is singular, then $(\Sigma, \widehat{\Sigma}) \notin \text{dom}(D)$ for every $\Sigma \in \mathbb{S}_+^p$. In this case, the uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma}) = \{\Sigma \in \mathbb{S}_+^p : D(\Sigma, \widehat{\Sigma}) \leq \varepsilon\}$ is empty, and the minimax equality (27) holds trivially because both sides of (27) evaluate to ∞ . Thus, we may always assume that $\widehat{\Sigma} \in \mathbb{S}_{++}^p$ for these divergences.

The objective function $\text{Tr}[X^2] - 2\langle \Sigma, X \rangle$ of the minimax problem (27) is convex and continuous in X for any fixed $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$, and it is concave and continuous in Σ for any fixed $X \in \mathbb{S}_+^p$. If $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is convex and compact, then (27) follows readily from Sion's classical minimax theorem. We will argue below that this is true for the Kullback-Leibler, Wasserstein, symmetrized Stein, quadratic, and weighted quadratic divergences. The uncertainty sets associated with the quadratic and weighted quadratic divergences constitute ellipsoids and are, therefore, trivially convex and compact. In addition, the convexity and compactness of the uncertainty set induced by the Wasserstein divergence follow from [45, Lemma A.6]. We next show that the Kullback-Leibler and symmetrized Stein divergences also induce convex and compact uncertainty sets.

Kullback-Leibler Divergence: For any fixed $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, the Kullback-Leibler divergence $D(\Sigma, \widehat{\Sigma})$ constitutes a continuous extended real-valued function of Σ . Indeed, one can show that $D(\Sigma, \widehat{\Sigma})$ tends to infinity as Σ approaches the boundary of \mathbb{S}_+^p and $\widehat{\Sigma} \in \mathbb{S}_{++}^p$ is kept fixed. Therefore, the uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is closed as a sublevel set of a continuous function. As $t - 1 - \log t \geq 0$ for every $t > 0$, any $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$ satisfies

$$\varepsilon \geq D(\Sigma, \widehat{\Sigma}) = \frac{1}{2} \sum_{i=1}^p \left(\lambda_i(\widehat{\Sigma}^{-1}\Sigma) - 1 - \log \lambda_i(\widehat{\Sigma}^{-1}\Sigma) \right) \geq \frac{1}{2} \left(\lambda_p(\widehat{\Sigma}^{-1}\Sigma) - 1 - \log \lambda_p(\widehat{\Sigma}^{-1}\Sigma) \right).$$

Note that the function $t - 1 - \log t$ grows indefinitely as t tends to infinity. Consequently, the above inequality implies that there exists $\bar{\lambda} > 0$ with $\lambda_p(\widehat{\Sigma}^{-1}\Sigma) \leq \bar{\lambda}$ for all $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$. Recall now that the spectral norm of any positive definite matrix coincides with its maximum eigenvalue. For any $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$ we thus have

$$\|\Sigma\| = \|\widehat{\Sigma}^{\frac{1}{2}} \widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}} \widehat{\Sigma}^{\frac{1}{2}}\| \leq \|\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}\| \|\widehat{\Sigma}\| = \lambda_p(\Sigma \widehat{\Sigma}^{-1}) \lambda_p(\widehat{\Sigma}) \leq \bar{\lambda} \lambda_p(\widehat{\Sigma}),$$

where the second equality holds because $\|\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}\| = \lambda_p(\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}})$ and because $\Sigma \widehat{\Sigma}^{-1}$ has the same eigenvalues as $\widehat{\Sigma}^{-\frac{1}{2}} \Sigma \widehat{\Sigma}^{-\frac{1}{2}}$. This shows that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is bounded and thus compact. Finally, note that $D(\Sigma, \widehat{\Sigma})$ is convex in Σ because $\text{Tr}[\widehat{\Sigma}^{-1}\Sigma]$ is linear and $\log \det(\widehat{\Sigma} \Sigma^{-1})$ is convex in Σ . Hence, $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is convex.

Symmetrized Stein Divergence: For any fixed $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, the symmetrized Stein divergence $D(\Sigma, \widehat{\Sigma})$ is continuous in Σ . Thus, the corresponding uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is closed. Also, any $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$ satisfies

$$\varepsilon \geq D(\Sigma, \widehat{\Sigma}) = \frac{1}{2} \sum_{i=1}^p \left(\lambda_i(\widehat{\Sigma}^{-1}\Sigma) + \lambda_i^{-1}(\widehat{\Sigma}^{-1}\Sigma) - 2 \right) \geq \frac{1}{2} \left(\lambda_p(\widehat{\Sigma}^{-1}\Sigma) + \lambda_p^{-1}(\widehat{\Sigma}^{-1}\Sigma) - 2 \right),$$

where the second inequality holds because all eigenvalues of $\widehat{\Sigma}^{-1}\Sigma$ are positive. Note that $t + t^{-1} - 2$ grows indefinitely as t tends to infinity. Hence, there exists $\bar{\lambda} > 0$ with $\lambda_p(\widehat{\Sigma}^{-1}\Sigma) \leq \bar{\lambda}$ for all $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$. By using a similar reasoning as for the Kullback-Leibler divergence, we can

thus show that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is compact. To prove convexity, we need to show that $D(\Sigma, \widehat{\Sigma})$ is a convex function of Σ . But this follows from [9, Exercise 3.18(a)].

The uncertainty sets induced by the Fisher-Rao and inverse Stein divergences fail to be convex in the standard Euclidean sense; see Section C.1. We will show, however, that these uncertainty sets are geodesically convex with respect to a certain Riemannian geometry on the cone \mathbb{S}_{++}^p . This will allow us to prove the minimax equality (27) by appealing to Theorem 3, which establishes a generalized version of Sion's minimax theorem for geodesic quasi-convex-quasi-concave minimax problems on Hadamard manifolds.

In order to apply Theorem 3, we embed the feasible set \mathbb{S}_+^p of the minimization problem in (27) into \mathbb{S}^p equipped with the usual Euclidean geometry. Recall from Example 3 that \mathbb{S}^p can be viewed as a Hadamard manifold and that the associated geodesic convexity coincides with the usual Euclidean convexity. Thus, the feasible set \mathbb{S}_+^p constitutes a convex subset of the Hadamard manifold \mathbb{S}^p . In addition, we embed the feasible set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ of the maximization problem in (27) into \mathbb{S}_{++}^p . Recall from Example 4 that \mathbb{S}_{++}^p also constitutes a Hadamard manifold. The objective function $\text{Tr}[X^2] - 2\langle \Sigma, X \rangle$ of (27) is ostensibly convex and continuous in X . Similarly, by Lemma 12, the objective function is geodesically concave and continuous in Σ . Hence, Theorem 3 applies, and the desired minimax equality (27) follows if we can prove that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is geodesically convex as well as compact with respect to the metric topology induced by the Riemannian geometry on \mathbb{S}_{++}^p . By Remark 1, however, this notion of compactness is equivalent to the usual compactness notion with respect to the Euclidean space \mathbb{S}^p . Therefore, it suffices to show that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is compact in the usual sense.

As for the Fisher-Rao divergence, the compactness and geodesic convexity of $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ follow from Lemma 11. It thus remains to prove the desired properties of $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ for the inverse Stein divergence.

Inverse Stein Divergence: For any fixed $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, the inverse Stein divergence $D(\Sigma, \widehat{\Sigma})$ is continuous in Σ . Therefore, the corresponding uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is closed. In addition, any $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$ satisfies

$$\varepsilon \geq D(\Sigma, \widehat{\Sigma}) = \frac{1}{2} \sum_{i=1}^p \left(\lambda_i(\Sigma^{-1}\widehat{\Sigma}) - 1 - \log \lambda_i(\Sigma^{-1}\widehat{\Sigma}) \right) \geq \frac{1}{2} \left(\lambda_1(\Sigma^{-1}\widehat{\Sigma}) - 1 - \log \lambda_1(\Sigma^{-1}\widehat{\Sigma}) \right),$$

where the second inequality holds because $t - 1 - \log t \geq 0$ for all $t > 0$. As $t - 1 - \log t$ grows indefinitely when t tends to 0, the above inequality implies that there exists $\underline{\lambda} > 0$ with $\lambda_1(\Sigma^{-1}\widehat{\Sigma}) \geq \underline{\lambda}$ for all $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$. This in turn implies that $\lambda_p(\widehat{\Sigma}^{-1}\Sigma) = \lambda_1^{-1}(\Sigma^{-1}\widehat{\Sigma}) \leq \underline{\lambda}^{-1}$ for all $\Sigma \in \mathcal{B}_\varepsilon(\widehat{\Sigma})$. We may thus conclude that $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is compact. Finally, since $D(\Sigma, \widehat{\Sigma}) = \frac{1}{2} \left(\text{Tr}[\Sigma^{-1}\widehat{\Sigma}] - p + \log \det \Sigma - \log \det \widehat{\Sigma} \right)$, $D(\Sigma, \widehat{\Sigma})$ is a geodesically convex function of Σ thanks to Lemmas 12(ii) and 12(iii). Therefore, $\mathcal{B}_\varepsilon(\widehat{\Sigma})$ is a geodesically convex set by virtue of Proposition 12. \square

C.1. Inapplicability of Sion's Minimax Theorem

We now show through counterexamples that if $D(\Sigma, \widehat{\Sigma})$ is the Fisher-Rao or inverse Stein divergence, then the corresponding uncertainty set $\mathcal{B}_\varepsilon(\widehat{\Sigma}) = \{\Sigma \in \mathbb{S}_+^p : D(\Sigma, \widehat{\Sigma}) \leq \varepsilon\}$ fails to be a convex subset of \mathbb{S}^p . Hence, for these divergences, we cannot appeal to Sion's classical minimax theorem to prove (27). More precisely, we will show that $D(\Sigma, \widehat{\Sigma})$ fails to be quasi-convex and thus has non-convex sublevel sets.

Definition 4 (Quasi-convex function). *A function $\psi : \mathbb{S}_+^p \rightarrow \bar{\mathbb{R}}$ is quasi-convex if for any $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^p$ and $\lambda \in [0, 1]$, we have $\psi(\lambda\Sigma_1 + (1-\lambda)\Sigma_2) \leq \max\{\psi(\Sigma_1), \psi(\Sigma_2)\}$.*

Example 1 (Non-convexity of the Fisher-Rao uncertainty set). *The divergence $D(\Sigma, \hat{\Sigma}) = \|\log(\hat{\Sigma}^{-\frac{1}{2}}\Sigma\hat{\Sigma}^{-\frac{1}{2}})\|_{\mathbb{F}}^2$ is not quasi-convex in Σ for any fixed $\hat{\Sigma} \in \mathbb{S}_{++}^3$. To see this, assume first that $\hat{\Sigma} = I_3$. Setting*

$$\Sigma_1 = \begin{pmatrix} 33 & -5 & -10 \\ -5 & 6 & 3 \\ -10 & 3 & 4 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 6 & -4 & 5 \\ -4 & 11 & -2 \\ 5 & -2 & 18 \end{pmatrix},$$

one readily verifies that $\Sigma_1, \Sigma_2 \succ 0$, while $D(\Sigma_1, I_3) = 16.4501$ and $D(\Sigma_2, I_3) = 16.2111$. In addition, we find

$$D(\tfrac{1}{2}\Sigma_1 + \tfrac{1}{2}\Sigma_2, I_3) = 18.6796 > \max\{16.4501, 16.2111\} = \max\{D(\Sigma_1, I_3), D(\Sigma_2, I_3)\}.$$

This shows that $D(\Sigma, I_3)$ fails to be quasi-convex in Σ . For a generic $\hat{\Sigma} \in \mathbb{S}_{++}^3$, we define $\Sigma'_1 = \hat{\Sigma}^{\frac{1}{2}}\Sigma_1\hat{\Sigma}^{\frac{1}{2}}$ and $\Sigma'_2 = \hat{\Sigma}^{\frac{1}{2}}\Sigma_2\hat{\Sigma}^{\frac{1}{2}}$. The above inequality then immediately implies that

$$D(\tfrac{1}{2}\Sigma'_1 + \tfrac{1}{2}\Sigma'_2, \hat{\Sigma}) > \max\{D(\Sigma'_1, \hat{\Sigma}), D(\Sigma'_2, \hat{\Sigma})\}.$$

Consequently, the function $D(\Sigma, \hat{\Sigma})$ fails to be quasi-convex in Σ irrespective of $\hat{\Sigma} \in \mathbb{S}_{++}^3$.

Example 2 (Non-convexity of the inverse Stein uncertainty set). *The function $D(\Sigma, \hat{\Sigma}) = \frac{1}{2}(\text{Tr}[\Sigma^{-1}\hat{\Sigma}] - 3 + \log \det(\Sigma\hat{\Sigma}^{-1}))$ is not quasi-convex in Σ for any fixed $\hat{\Sigma} \in \mathbb{S}_{++}^3$. Indeed, if $\hat{\Sigma} = I_3$, we may set*

$$\Sigma_1 = \begin{pmatrix} 30 & 13 & 23 \\ 13 & 12 & 9 \\ 23 & 9 & 20 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 27 & 13 & 23 \\ 13 & 10 & 14 \\ 23 & 14 & 30 \end{pmatrix}.$$

It can be verified that $\Sigma_1, \Sigma_2 \succ 0$, while $D(\Sigma_1, I_3) = 4.0427$ and $D(\Sigma_2, I_3) = 4.3020$. In addition, we find

$$D(\tfrac{1}{2}\Sigma_1 + \tfrac{1}{2}\Sigma_2, I_3) = 4.3262 > \max\{4.0427, 4.3020\} = \max\{D(\Sigma_1, I_3), D(\Sigma_2, I_3)\}.$$

This shows that $D(\Sigma, I_3)$ fails to be quasi-convex in Σ . For a generic $\hat{\Sigma} \in \mathbb{S}_{++}^3$, we define $\Sigma'_1 = \hat{\Sigma}^{\frac{1}{2}}\Sigma_1\hat{\Sigma}^{\frac{1}{2}}$ and $\Sigma'_2 = \hat{\Sigma}^{\frac{1}{2}}\Sigma_2\hat{\Sigma}^{\frac{1}{2}}$. The above inequality then immediately implies that

$$D(\tfrac{1}{2}\Sigma'_1 + \tfrac{1}{2}\Sigma'_2, \hat{\Sigma}) > \max\{D(\Sigma'_1, \hat{\Sigma}), D(\Sigma'_2, \hat{\Sigma})\}$$

that is, the function $D(\Sigma, \hat{\Sigma})$ fails to be quasi-convex in Σ irrespective of $\hat{\Sigma} \in \mathbb{S}_{++}^3$.

C.2. Riemannian Geometry and Geodesic Convexity

In order to keep this paper self-contained, we now briefly review some basic concepts from Riemannian geometry. For a more comprehensive survey of this topic, we refer to the excellent textbooks [29, 37].

Definition 5 (Riemannian manifold). *A Riemannian manifold is a pair $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ consisting of a differentiable manifold \mathcal{M} and a smooth family of inner products $\{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}}$ defined on the tangent spaces $T_u\mathcal{M}$ of \mathcal{M} . That is, for any $u \in \mathcal{M}$, $\langle \cdot, \cdot \rangle_u$ represents a symmetric, positive definite bilinear map on $T_u\mathcal{M}$. The family $\{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}}$ of inner products is called a Riemannian metric on \mathcal{M} .*

Throughout this paper we will restrict attention to Hadamard manifolds.

Definition 6 (Hadamard manifolds). *A Hadamard manifold is a complete, simply connected Riemannian manifold that has everywhere non-positive sectional curvature.*

Intuitively, the sectional curvature of a Riemannian manifold is non-positive at a point u if and only if the area of any small two-dimensional disc centered at u is larger or equal to the area of a disc with the same radius in flat space. For a formal definition see [29, p. 236] or [37, p. 154]. All piecewise continuously differentiable curves on a Riemannian manifold—and, in particular, on a Hadamard manifold—can be assigned a length.

Definition 7 (Length of a curve). *The length of a continuously differentiable curve $c : [0, 1] \rightarrow \mathcal{M}$ on a Riemannian manifold $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ is defined as*

$$L(c) = \int_0^1 \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)}} dt.$$

If c is piecewise continuously differentiable, then its length is defined as the sum of the lengths of its pieces.

The Riemannian distance between two points $u_1, u_2 \in \mathcal{M}$ is defined as $d_{\mathcal{M}}(u_1, u_2) = \min_c L(c)$, where the minimum is over all continuously differentiable curves c with constant speed $(\langle \dot{c}(t), \dot{c}(t) \rangle_{c(t)})^{\frac{1}{2}}$ that connect u_1 and u_2 . For complete and connected Riemannian manifolds, the minimum is guaranteed to exist, and any minimizer is a geodesic. Moreover, by the Hopf-Rinow theorem [29, 37], any two points on a Hadamard manifold are connected by a unique geodesic. This greatly simplifies the study of convexity on such manifolds.

Definition 8 (Geodesically convex sets). *If $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ is a Hadamard manifold, then $\mathcal{U} \subseteq \mathcal{M}$ is geodesically convex if, for any $u_1, u_2 \in \mathcal{U}$, the image of the geodesic connecting u_1 and u_2 lies within \mathcal{U} .*

Definition 9 (Geodesically (quasi-)convex function). *If $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ is a Hadamard manifold and $\mathcal{U} \subseteq \mathcal{M}$ is geodesically convex, then the function $\psi : \mathcal{U} \rightarrow \mathbb{R}$ is geodesically (quasi-)convex if the composition $\psi \circ c : [0, 1] \rightarrow \mathbb{R}$ is (quasi-)convex function in the usual Euclidean sense for every geodesic c connecting two arbitrary points in \mathcal{U} . In addition, ϕ is geodesically (quasi-)concave if $-\phi$ is geodesically (quasi-)convex.*

Definition 9 makes sense because a geodesic is always parametrized proportionally to arc length. It readily implies that all sublevel sets of a geodesically quasi-convex function are geodesically convex.

Proposition 12 ([62, Theorem 3.4]). *If $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ is a Hadamard manifold and $\psi : \mathcal{M} \rightarrow \mathbb{R}$ is geodesically quasi-convex, then the sublevel set $\{u \in \mathcal{M} : \psi(u) \leq \alpha\}$ is geodesically convex for any $\alpha \in \mathbb{R}$.*

The examples below are useful for our theoretical development and used in the proof of Proposition 11.

Example 3. *The Euclidean spaces \mathbb{R}^p and \mathbb{S}^p equipped with their usual inner products constitute Hadamard manifolds. In both cases, geodesic convexity (of sets as well as functions) reduces to Euclidean convexity.*

Example 4. The cone of positive definite matrices \mathbb{S}_{++}^p represents a differentiable manifold [5, 29]. The tangent space $T_\Sigma \mathbb{S}_{++}^p$ at $\Sigma \in \mathbb{S}_{++}^p$ is naturally identified with \mathbb{S}^p , that is, all tangent vectors constitute symmetric matrices. We can assign every $\Sigma \in \mathbb{S}_{++}^p$ an inner product $\langle \cdot, \cdot \rangle_\Sigma : \mathbb{S}^p \times \mathbb{S}^p \rightarrow \mathbb{R}$ defined through

$$\langle \Sigma_1, \Sigma_2 \rangle_\Sigma = \text{Tr}[\Sigma^{-1} \Sigma_1 \Sigma^{-1} \Sigma_2] \quad \forall \Sigma_1, \Sigma_2 \in \mathbb{S}^p.$$

By [29, Theorem XII 1.2], \mathbb{S}_{++}^p equipped with the inner products $\langle \cdot, \cdot \rangle_\Sigma$, $\Sigma \in \mathbb{S}_{++}^p$, is a Hadamard manifold.

Remark 1. By definition, any Hadamard manifold $(\mathcal{M}, \{\langle \cdot, \cdot \rangle_u\}_{u \in \mathcal{M}})$ is simply connected and therefore, in particular, connected. Hence, [38, Theorem 13.29] implies that the metric topology on \mathcal{M} induced by the Riemannian distance $d_{\mathcal{M}}$ coincides with the manifold topology. For instance, the metric topology on the Hadamard manifold \mathbb{S}_{++}^p from Example 4 coincides with the subspace topology on \mathbb{S}_{++}^p inherited from the ambient vector space \mathbb{S}^p , which is the standard (Euclidean norm) topology used for matrices.

In the following lemmas, we treat \mathbb{S}_{++}^p as a Hadamard manifold in the sense of Example 4.

Lemma 11 (Compactness and convexity [46, Theorem 2.5]). *For any fixed $\Sigma' \in \mathbb{S}_{++}^p$, the set*

$$\left\{ \Sigma \in \mathbb{S}_{++}^p : \|\log(\Sigma'^{-\frac{1}{2}} \Sigma \Sigma'^{-\frac{1}{2}})\|_{\text{F}}^2 \leq \varepsilon^2 \right\}$$

constitutes a compact and geodesically convex subset of \mathbb{S}_{++}^p .

We now show that several popular matrix functions are geodesically convex. Here, we adopt the standard terminology whereby a function that is both geodesically convex and concave is called geodesically linear.

Lemma 12 (Geodesic convexity of popular matrix functions). *The following hold.*

- (i) $g(\Sigma) = \text{Tr}[X\Sigma]$ is geodesically convex on \mathbb{S}_{++}^p for every $X \in \mathbb{S}_+^p$.
- (ii) $g(\Sigma) = \text{Tr}[X\Sigma^{-1}]$ is geodesically convex on \mathbb{S}_{++}^p for every $X \in \mathbb{S}_+^p$.
- (iii) $g(\Sigma) = \log \det \Sigma$ is geodesically linear on \mathbb{S}_{++}^p .

Proof of Lemma 12. We can prove assertion (i) by showing that, for every fixed $\Sigma \in \mathbb{S}_{++}^p$, the Riemannian Hessian of the function $g(\Sigma) = \text{Tr}[X\Sigma]$ is positive semidefinite on the tangent space $T_\Sigma \mathbb{S}_{++}^p \cong \mathbb{S}^p$ [2, 62]. To this end, note first that the Euclidean gradient of g is given by $\nabla g(\Sigma) = X$ and that the Euclidean Hessian $\nabla^2 g(\Sigma)$ coincides with the zero map from \mathbb{S}^p to \mathbb{S}^p . By [14, § 4.2], the Riemannian Hessian of g thus satisfies

$$\text{Hess } g(\Sigma)[S] = \frac{1}{2}(SX\Sigma + \Sigma XS) \quad \forall S \in \mathbb{S}^p.$$

This implies that

$$\langle \text{Hess } g(\Sigma)[S], S \rangle_\Sigma = \text{Tr}[SXS\Sigma^{-1}] \geq 0 \quad \forall S \in \mathbb{S}^p,$$

where the inequality holds because $SXS \in \mathbb{S}_+^p$ and $\Sigma^{-1} \in \mathbb{S}_{++}^p$. Thus, the Riemannian Hessian of g is positive semidefinite on the tangent space $T_\Sigma \mathbb{S}_{++}^p \cong \mathbb{S}^p$. As $\Sigma \in \mathbb{S}_{++}^p$ was chosen freely, this shows via [62, Theorem 6.2] that g is geodesically convex throughout \mathbb{S}_{++}^p .

Assertions (ii) and (iii) are proved similarly. As for assertion (ii), note that the gradient of $g(\Sigma) = \text{Tr}[X\Sigma^{-1}]$ is given by $\nabla g(\Sigma) = -\Sigma^{-1}X\Sigma^{-1}$ [50, § 2.2]. Also, the Hessian of g is a

linear operator on \mathbb{S}^p satisfying

$$\begin{aligned}\nabla^2 g(\Sigma)[S] &= \left. \frac{d\nabla g(\Sigma + tS)}{dt} \right|_{t=0} = - \left. \frac{d(\Sigma + tS)^{-1}}{dt} \right|_{t=0} X\Sigma^{-1} - \Sigma^{-1}X \left. \frac{d(\Sigma + tS)^{-1}}{dt} \right|_{t=0} \\ &= \Sigma^{-1}S\Sigma^{-1}X\Sigma^{-1} + \Sigma^{-1}X\Sigma^{-1}S\Sigma^{-1},\end{aligned}$$

where the third equality exploits [50, § 2.2]. By [14, § 4.2], the Riemannian Hessian of g thus satisfies

$$\text{Hess } g(\Sigma)[S] = \frac{1}{2}(\Sigma^{-1}S\Sigma^{-1}X + X\Sigma^{-1}S\Sigma^{-1}) \quad \forall S \in \mathbb{S}^p.$$

This implies that

$$\langle \text{Hess } g(\Sigma)[S], S \rangle_\Sigma = \frac{1}{2} \text{Tr}[\Sigma^{-1}(S\Sigma^{-1}X + X\Sigma^{-1}S\Sigma^{-1})\Sigma^{-1}S] = \text{Tr}[S\Sigma^{-1}S\Sigma^{-1}X\Sigma^{-1}] \geq 0 \quad \forall S \in \mathbb{S}^p,$$

where the inequality holds because $S\Sigma^{-1}S$ and $\Sigma^{-1}X\Sigma^{-1}$ are positive semidefinite. Thus, the Riemannian Hessian of g is positive semidefinite on the tangent space $T_\Sigma \mathbb{S}_{++}^p \cong \mathbb{S}^p$, and the claim follows.

As for assertion (iii), the gradient of $g(\Sigma) = \log \det \Sigma$ is given by $\nabla g(\Sigma) = -\Sigma^{-1}$, and the Hessian of g is a linear operator on \mathbb{S}^p satisfying $\nabla^2 g(\Sigma)[S] = \Sigma^{-1}S\Sigma^{-1}$ [50, § 2.2]. By [14, § 4.2], the Riemannian Hessian of g thus satisfies $\text{Hess } g(\Sigma)[S] = 0$ for all $S \in \mathbb{S}^p$. Hence, g is both geodesically convex and concave on \mathbb{S}_{++}^p . \square

C.3. A Riemannian Generalization of Sion's Minimax Theorem

We now present a generalization of Sion's minimax theorem for geodesically convex-concave saddle functions on Hadamard manifolds.⁴

Theorem 3 (Sion's minimax theorem for geodesically convex-concave saddle problems). *Let \mathcal{U} and \mathcal{V} be geodesically convex subsets of two Hadamard manifolds, and assume that \mathcal{U} is compact. Also, let $\psi : \mathcal{U} \times \mathcal{V} \rightarrow \mathbb{R}$ be a function with $\psi(u, \cdot)$ being upper semi-continuous and geodesically quasi-concave on \mathcal{V} for any fixed $u \in \mathcal{U}$ and with $\psi(\cdot, v)$ being lower semi-continuous and geodesically quasi-convex on \mathcal{U} for every fixed $v \in \mathcal{V}$. Then,*

$$\min_{u \in \mathcal{U}} \sup_{v \in \mathcal{V}} \psi(u, v) = \sup_{v \in \mathcal{V}} \min_{u \in \mathcal{U}} \psi(u, v).$$

The proof of this Riemannian minimax theorem is omitted because it closely follows the approach in [26], with the natural modification that line segments between two points in linear spaces are replaced by (unique) geodesics connecting two points on Hadamard manifolds.

APPENDIX D. VERIFICATION OF THE REARRANGEMENT PROPERTY

Proposition 13. *All the divergences listed in Table 1 satisfy Assumption 2(c).*

Proof of Proposition 13. Let D be the Kullback-Leibler, Fisher-Rao, inverse Stein or symmetric Stein divergence. In either case, if x or y contains any vanishing entry, then both sides of the rearrangement inequality in Assumption 2(c) evaluate to $+\infty$; see the definitions in Table 1. Thus, Assumption 2(c) is trivially satisfied. It therefore suffices to prove the inequality for $x, y \in \mathbb{R}_{++}^p$. Next, let D be the weighted quadratic divergence. Hence, if y contains any vanishing entry, then both sides of the rearrangement inequality evaluate again to $+\infty$, and

⁴While finalizing this paper, we discovered a concurrent work describing a result akin to Theorem 3 [70].

Divergence	$h_1(t)$	$h_2(t)$	$g_1(t)$	$g_2(t)$	$f(t)$	$tf'(t)$
Kullback-Leibler	$-\frac{1}{4}$	$-\frac{1}{4}$	t	$\frac{1}{t}$	$\frac{1}{2}(t - \log t)$	$\frac{1}{2}(t - 1)$
Wasserstein	t	t	t	t	$-2\sqrt{t}$	$-\sqrt{t}$
Fisher-Rao	0	0	t	$\frac{1}{t}$	$(\log t)^2$	$2 \log t$
Inverse Stein	$-\frac{1}{4}$	$-\frac{1}{4}$	t	$\frac{1}{t}$	$\frac{1}{2}(\frac{1}{t} + \log t)$	$\frac{1}{2}(1 - \frac{1}{t})$
Symmetrized Stein	$-\frac{1}{2}$	$-\frac{1}{2}$	t	$\frac{1}{t}$	$\frac{1}{2}(t + \frac{1}{t})$	$\frac{1}{2}(t - \frac{1}{t})$
Quadratic	t^2	t^2	t	t	$-2t$	$-2t$
Weighted quadratic	$-2t$	t	t^2	$\frac{1}{t}$	t	t

TABLE 4. Functions h_1 , h_2 , g_1 , g_2 and f in the representation (28) of the divergences of Table 1.

Assumption 2(c) is trivially satisfied. It therefore suffices to assume that $y \in \mathbb{R}_{++}^p$. With these assumptions in place, both sides of the rearrangement inequality are guaranteed to be finite.

The subsequent proof requires additional notation. We use $\sigma_i(S)$ to denote the i -th smallest singular value of the matrix $S \in \mathbb{S}^p$. The vector $\sigma(S) \in \mathbb{R}_+^p$ is then defined through $(\sigma(S))_i = \sigma_i(S)$ for all $i = 1, \dots, p$. Any univariate function $g : \mathbb{R} \rightarrow \mathbb{R}$ naturally induces multivariate functions $g : \mathbb{R}^p \rightarrow \mathbb{R}^p$ and $g : \mathbb{S}^p \rightarrow \mathbb{S}^p$, which, by slight abuse of notation, are represented by the same symbol g . Specifically, for any $x \in \mathbb{R}^p$, we define $g(x) \in \mathbb{R}^p$ through $(g(x))_i = g(x_i)$ for all $i = 1, \dots, p$. Similarly, for any $S \in \mathbb{S}^p$ with eigenvalue decomposition $S = V_S \text{Diag}(\lambda(S)) V_S^\top$ with $V_S \in \mathcal{O}_p$, we define $g(S) \in \mathbb{S}^p$ through $g(S) = V_S \text{Diag}(g(\lambda(S))) V_S^\top$.

Observe now that all divergences listed in Table 1 are representable as

$$D(X, Y) = \sum_{i=1}^p (h_1(\lambda_i(X)) + h_2(\lambda_i(Y))) + \sum_{i=1}^p f(\lambda_i(g_2(Y^{\frac{1}{2}})g_1(X)g_2(Y^{\frac{1}{2}}))) \quad (28)$$

for some functions f , h_1 , h_2 , g_1 and g_2 from \mathbb{R} to $\overline{\mathbb{R}}$ as specified in Table 4. As the spectrum of any matrix is invariant under conjugation with an orthogonal matrix $V \in \mathcal{O}_p$, we have

$$\sum_{i=1}^p (h_1(\lambda_i(V \text{Diag}(x^\uparrow) V^\top)) + h_2(\lambda_i(\text{Diag}(y^\uparrow)))) = \sum_{i=1}^p (h_1(\lambda_i(\text{Diag}(x^\uparrow))) + h_2(\lambda_i(\text{Diag}(y^\uparrow))))$$

for all $x, y \in \mathbb{R}^p$. In view of the representation (28) and the above identity, it remains to be shown that

$$\sum_{i=1}^p f(\lambda_i(\text{Diag}(\sqrt{y^\uparrow}) V g_1(\text{Diag}(x^\uparrow)) V^\top g_2(\text{Diag}(\sqrt{y^\uparrow})))) \geq \sum_{i=1}^p f(\lambda_i(g_1(\text{Diag}(x^\uparrow)) g_2(\text{Diag}(y^\uparrow)))) \quad (29)$$

for all $x, y \in \mathbb{R}_+^p$ and $V \in \mathcal{O}_p$. Table 4 shows that always either of the following two conditions holds:

- $t \mapsto tf'(t)$ is strictly increasing, g_1 is strictly increasing and g_2 is strictly decreasing;
- $t \mapsto tf'(t)$ is strictly decreasing, and g_1 and g_2 are both strictly increasing.

The desired inequality (29) then follows from [69, Theorem 3]. Inspecting the proofs of [69, Theorem 3 and Lemma 1] further reveals that (29) holds if and only if $Vg_1(\text{Diag}(x^\uparrow))V^\top = g_1(\text{Diag}(x^\uparrow))$, which is equivalent to $V\text{Diag}(x^\uparrow)V^\top = \text{Diag}(x^\uparrow)$ because g_1 is strictly increasing. This observation completes the proof. \square

APPENDIX E. PROOFS OF SECTION 4

Proof of Theorem 2. We prove the assumptions one by one. Note first that, by Proposition 11, every divergence D in Table 1 satisfies the minimax property specified in Assumption 1.

Assumption 2 requires D to be a spectral divergence. To show that D is orthogonally equivariant, recall that the spectrum of a matrix is preserved under similarity transformations. As the trace and the determinant are spectral functions, the orthogonal equivariance of all divergences in Table 1 is easily verified using elementary rules of matrix algebra. It is also straightforward to verify that every divergence D in Table 1 is spectral with generator d as specified in Table 2. In addition, the domain of d contains a point (a, b) with $b > 0$, and d is ostensibly continuous throughout its domain. The rearrangement property holds thanks to Proposition 13.

Assumption 4 follows immediately from definitions of the generators in Table 2. For example, it is clear that the generator $d_b(\cdot) = d(\cdot, b) = (\log(\cdot/b))^2$ of the Fisher-Rao divergence is twice continuously differentiable on \mathbb{R}_{++} for any fixed $b > 0$. In addition, we have $d_b''(a) = 2(1 - \log(a/b))/a^2 > 0$ for any $a \in (0, b]$ and $b > 0$, which shows that d_b is convex on $[0, b]$. Similarly, one can prove Assumption 4 for all other divergences.

It remains to be shown that all generators in Table 2 satisfy the differential inequality of Assumption 5. For example, the generator $d(a, b) = (\log(a/b))^2$ of the Fisher-Rao divergence satisfies

$$\frac{\partial}{\partial a}d(a, b) = \frac{2}{a} \log \frac{a}{b}, \quad \frac{\partial^2}{\partial a^2}d(a, b) = \frac{2}{a^2} \left(1 - \log \frac{a}{b}\right) \quad \text{and} \quad \frac{\partial^2}{\partial b \partial a}d(a, b) = -\frac{2}{ab} \quad \forall a, b \in \mathbb{R}_{++}.$$

Therefore, we obtain

$$a \frac{\partial^2}{\partial a^2}d(a, b) + b \frac{\partial^2}{\partial a \partial b}d(a, b) - \frac{\partial}{\partial a}d(a, b) = \frac{2}{a} \left(1 - \log \frac{a}{b}\right) - \frac{2}{a} - \frac{2}{a} \log \frac{a}{b} = -\frac{4}{a} \log \frac{a}{b} > 0$$

for all for any $b > a > 0$. Hence, Assumption 5 holds for the Fisher-Rao divergence. Similarly, Assumption 5 can be proved for all other divergences using the basic rules of calculus. \square

We now prove Corollaries 1, 2 and 3, which characterize the eigenvalue map as well as the inverse shrinkage intensity of the KL, Wasserstein and Fisher-Rao covariance shrinkage estimators, respectively.

Proof of Corollary 1. The generator of the KL divergence is given by $d(a, b) = \frac{1}{2}(\frac{a}{b} - 1 - \log \frac{a}{b})$; see Table 2. Note that Assumptions 1, 2, 4 and 5 hold by Theorem 2, Assumption 3(a) holds because $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, and Assumption 3(b) holds because $d(0, b) = +\infty$ for any $b > 0$. Therefore, Theorem 1 applies, which implies that problem (4) is uniquely solved by $X^\star = \widehat{V} \text{Diag}(x^\star) \widehat{V}^\top$, where $x_i^\star = s(\gamma^\star, \hat{x}_i)$ for every $i = 1, \dots, p$. Next, we construct the eigenvalue map s defined in (8). If $b > 0$, then $s(\gamma, b)$ is the unique solution $a^\star \geq 0$ of

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

We thus obtain

$$s(\gamma, b) = \frac{-\gamma + \sqrt{\gamma^2 + 16b^2\gamma}}{8b}.$$

It remains to find a formula for γ^* . By Theorem 1, γ^* is the unique positive root of the equation

$$\sum_{i=1}^p d(s(\gamma^*, \hat{x}_i), \hat{x}_i) - \varepsilon = 0 \iff 2\varepsilon + p + \sum_{i=1}^p \left[-\frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} + \log \frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} \right] = 0.$$

To show that γ_{KL} provides an upper bound on γ^* , note that the above equation implies that

$$0 = 2\varepsilon + p + \sum_{i=1}^p \left[-\frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} + \log \frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} \right] \geq 2\varepsilon + \sum_{i=1}^p \log \frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i} \geq 2\varepsilon + p \log \frac{s(\gamma^*, \hat{x}_p)}{\hat{x}_p}.$$

Here, the two inequalities follow from Lemmas 7 and 1, which imply that $s(\gamma, b) < b$ for all $\gamma, b > 0$ and that $s(\gamma, b)/b$ is non-increasing in b , respectively. Rearranging the above inequality yields $\hat{x}_p e^{-\frac{2\varepsilon}{p}} \geq s(\gamma^*, \hat{x}_p)$. As $s(\gamma, \hat{x}_p)$ is strictly increasing in γ by virtue of Lemma 7(ii), the unique solution γ_{KL} of the equation

$$\hat{x}_p e^{-\frac{2\varepsilon}{p}} = s(\gamma_{\text{KL}}, \hat{x}_p) = \frac{-\gamma_{\text{KL}} + \sqrt{\gamma_{\text{KL}}^2 + 16\hat{x}_p^2\gamma_{\text{KL}}}}{8\hat{x}_p}$$

provides an upper bound on γ^* . The desired formula for γ_{KL} is obtained by solving this equation. \square

Proof of Corollary 2. The generator of the Wasserstein divergence is given by $d(a, b) = a + b - 2\sqrt{ab}$; see Table 2. Assumptions 1, 2, 4 and 5 hold by Theorem 2, Assumption 3(a) holds because $\hat{\Sigma} \in \mathbb{S}_+^p$, and Assumption 3(b) holds because $\varepsilon \in (0, \text{Tr}[\hat{\Sigma}])$, which implies that $\sum_{i=1}^p d(0, \hat{x}_i) = \sum_{i=1}^p \hat{x}_i = \text{Tr}[\hat{\Sigma}] > \varepsilon$. Thus, Theorem 1 applies. Recall now from (8) that if $\gamma > 0$, then $s(\gamma, b)$ is defined as the unique solution $a^* \geq 0$ of

$$0 = 2a^* + \gamma \frac{\partial}{\partial a} d(a^*, b) = 2a^* + \gamma \left(1 - \sqrt{\frac{b}{a^*}} \right).$$

Solving a cubic equation in $\sqrt{a^*}$ thus reveals that $s(\gamma, b)$ is given by (10a). Theorem 1 further implies that the inverse shrinkage intensity γ^* is the unique positive root of the equation (10b). To show that γ_{W} provides an upper bound on γ^* , let $i' \in \{1, \dots, p\}$ be the smallest index i with $\hat{x}_i > 0$. As $s(\gamma^*, 0) = 0$, (10b) implies

$$\begin{aligned} 0 &= \varepsilon - \sum_{i=i'}^p \left(\sqrt{\hat{x}_i} - \sqrt{s(\gamma^*, \hat{x}_i)} \right)^2 \geq \varepsilon - \hat{x}_p \sum_{i=i'}^p \left(1 - \sqrt{\frac{s(\gamma^*, \hat{x}_i)}{\hat{x}_i}} \right)^2 \\ &\geq \varepsilon - p\hat{x}_p \left(1 - \sqrt{\frac{s(\gamma^*, \hat{x}_p)}{\hat{x}_p}} \right)^2 = \varepsilon - p \left(\sqrt{\hat{x}_p} - \sqrt{s(\gamma^*, \hat{x}_p)} \right)^2, \end{aligned} \quad (30)$$

where the first inequality holds because $\hat{x}_i \leq \hat{x}_p$, and the second inequality follows from Lemmas 1 and 7, which imply that $s(\gamma, b)/b$ is non-increasing in b and that $0 < s(\gamma, b) < b$ for all $\gamma, b > 0$, respectively. The defining equation for $s(\gamma^*, \hat{x}_p)$ further implies that

$$\left(\sqrt{\hat{x}_p} - \sqrt{s(\gamma^*, \hat{x}_p)} \right)^2 = \frac{4s(\gamma^*, \hat{x}_p)^3}{\gamma^{*2}}. \quad (31)$$

Substituting (31) into (30) yields

$$0 \geq \varepsilon - \frac{4ps(\gamma^*, \hat{x}_p)^3}{\gamma^{*2}} \geq \varepsilon - \frac{4p\hat{x}_p^3}{\gamma^{*2}} \iff \gamma^* \leq 2\sqrt{\frac{p\hat{x}_p^3}{\varepsilon}} = \gamma_{\text{W}}.$$

This observation completes the proof. \square

Proof of Corollary 3. The generator of the Fisher-Rao divergence is $d(a, b) = (\log \frac{a}{b})^2$; see Table 2. Assumptions 1, 2, 4 and 5 hold by Theorem 2, Assumption 3(a) holds because $\widehat{\Sigma} \in \mathbb{S}_{++}^p$, and Assumption 3(b) holds because $d(0, b) = +\infty$ for any $b > 0$. Thus, Theorem 1 applies. If $b > 0$, $s(\gamma, b)$ is the unique solution $a^* \geq 0$ of

$$0 = 2a^* + \gamma \frac{\partial}{\partial a} d(a^*, b) = 2a^* + \frac{2\gamma}{a^*} \log \frac{a^*}{b} \quad \Longleftrightarrow \quad \frac{2(a^*)^2}{\gamma} e^{\frac{2(a^*)^2}{\gamma}} = \frac{2b^2}{\gamma}.$$

Recall now that, for any $t > -e^{-1}$, the principal branch of the Lambert W -function is defined as the unique solution $W_0(t)$ of the equation $We^W = t$. Identifying W with $2(a^*)^2/\gamma$ and t with $2b^2/\gamma > 0$, we thus find

$$s(\gamma, b) = \sqrt{\frac{\gamma}{2} W_0\left(\frac{2b^2}{\gamma}\right)} = b \exp\left(-\frac{1}{2} W_0\left(\frac{2b^2}{\gamma}\right)\right), \quad (32)$$

where the second equality holds because $W_0(t) = te^{-W_0(t)}$. This proves (11a). Theorem 1 further implies that the inverse shrinkage intensity γ^* is the unique positive root of the equation (11b). It remains to prove that γ_{FR} upper bounds γ^* . Recalling that $0 \leq W_0(t) = t \exp(-W_0(t)) \leq t$ for any $t \geq 0$, (11b) implies that

$$4\varepsilon = \sum_{i=1}^p W_0^2\left(\frac{2\hat{x}_i^2}{\gamma^*}\right) \leq \sum_{i=1}^p \frac{4\hat{x}_i^4}{\gamma^{*2}} \quad \Longrightarrow \quad \gamma^* \leq \sqrt{\sum_{i=1}^p \frac{\hat{x}_i^4}{\varepsilon}} \leq \|\widehat{\Sigma}\|_{\text{F}}^2 \sqrt{\varepsilon} = \gamma_{\text{FR}}.$$

This observation completes the proof. □