Distributionally Robust Inverse Covariance Estimation
The Wasserstein Shrinkage Estimator

Nguyen, Viet Anh; Kuhn, Daniel; Esfahani, Peyman Mohajerin

DOI
10.1287/opre.2020.2076

Publication date
2022

Document Version
Final published version

Published in
Operations Research

Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator

Viet Anh Nguyen, Daniel Kuhn, Peyman Mohajerin Esfahani

To cite this article:
Viet Anh Nguyen, Daniel Kuhn, Peyman Mohajerin Esfahani (2022) Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator. Operations Research 70(1):490-515. https://doi.org/10.1287/opre.2020.2076

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 70, No. 1, January–February 2022, pp. 490–515
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator

Viet Anh Nguyen,[a] Daniel Kuhn,[b] Peyman Mohajerin Esfahani[c]

[a] Department of Management Science and Engineering, Stanford University, Stanford, California 94305; [b] Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [c] Delft Center for Systems and Control, Delft University of Technology, 2628 CD Delft, Netherlands

Contact: viet-anh.nguyen@stanford.edu, https://orcid.org/0000-0002-9607-7891 (VAN); daniel.kuhn@epfl.ch (DK); p.mohajerinesfahani@tudelft.nl, https://orcid.org/0000-0003-1286-8782 (PME)

Received: March 29, 2018
Revised: March 27, 2019
Accepted: July 16, 2020
Published Online in Articles in Advance: July 23, 2021

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2020.2076

Abstract. We introduce a distributionally robust maximum likelihood estimation model with a Wasserstein ambiguity set to infer the inverse covariance matrix of a p-dimensional Gaussian random vector from n independent samples. The proposed model minimizes the worst case (maximum) of Stein's loss across all normal reference distributions within a prescribed Wasserstein distance from the normal distribution characterized by the sample mean and the sample covariance matrix. We prove that this estimation problem is equivalent to a semidefinite program that is tractable in theory but beyond the reach of general-purpose solvers for practically relevant problem dimensions p. In the absence of any prior structural information, the estimation problem has an analytical solution that is naturally interpreted as a nonlinear shrinkage estimator. Besides being invertible and well conditioned even for $p > n$, the new shrinkage estimator is rotation equivariant and preserves the order of the eigenvalues of the sample covariance matrix. These desirable properties are not imposed ad hoc but emerge naturally from the underlying distributionally robust optimization model. Finally, we develop a sequential quadratic approximation algorithm for efficiently solving the general estimation problem subject to conditional independence constraints typically encountered in Gaussian graphical models.

Funding: The authors gratefully acknowledge financial support from the Swiss National Science Foundation [Grants BSCGI0_157733 and P2EZP2_165264] and the ERC [Grant TRUST-949796].

Keywords: distributionally robust optimization • data-driven optimization • Wasserstein distance • shrinkage estimator • maximum likelihood estimation

1. Introduction

The covariance matrix $\Sigma := \mathbb{E}^{\mathbb{P}}[(\xi - \mathbb{E}^{\mathbb{P}}[\xi])(\xi - \mathbb{E}^{\mathbb{P}}[\xi])^\top]$ of a random vector $\xi \in \mathbb{R}^p$ governed by a distribution $\mathbb{P}$ collects basic information about the spreads of all individual components and the linear dependencies among all pairs of components of ξ. The inverse Σ^{-1} of the covariance matrix is called the *precision matrix*. This terminology captures the intuition that a large spread reflects a low precision, and vice versa. Although the covariance matrix appears in the *formulations* of many problems in engineering, science, and economics, it is often the precision matrix that emerges in their *solutions*. For example, the optimal classification rule in linear discriminant analysis (Fisher 1936), the optimal investment portfolio in Markowitz's celebrated mean-variance model (Markowitz 1952), and the optimal array vector of the beamforming problem in signal processing (Du et al. 2010) all depend on the precision matrix. Moreover, the optimal fingerprint method used to detect a multivariate climate change signal blurred by weather noise requires knowledge of the climate vector's precision matrix (Ribes et al. 2009).

1.1. Background on Precision Matrix Estimation

If the distribution $\mathbb{P}$ of ξ is known, then the covariance matrix Σ and the precision matrix Σ^{-1} can at least principally be calculated in closed form. In practice, however, $\mathbb{P}$ is never known and only indirectly observable through n independent training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_n$ from $\mathbb{P}$. In this setting, Σ and Σ^{-1} need to be estimated from the training data. Arguably the simplest estimator for Σ is the sample covariance matrix $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n (\widehat{\xi}_i - \widehat{\mu})(\widehat{\xi}_i - \widehat{\mu})^\top$, where $\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n \widehat{\xi}_i$ stands for the sample mean. Note that $\widehat{\mu}$ and $\widehat{\Sigma}$ simply represent the actual mean and covariance matrix of the uniform distribution on the training samples. For later convenience, $\widehat{\Sigma}$ is defined here without Bessel's correction and thus constitutes a biased estimator.[1] Moreover, as a sum of n rank 1 matrices, $\widehat{\Sigma}$ is rank deficient in the big data regime ($p > n$). In this case, $\widehat{\Sigma}$ cannot be inverted to obtain a precision matrix estimator, which is often the actual quantity of interest.

If ξ follows a normal distribution with unknown mean μ and precision matrix $X > 0$, which we will

assume throughout the rest of the paper, then the log-likelihood function of the training data can be expressed as

$$\begin{aligned}\widehat{\mathcal{L}}(\mu, X) := & -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det X \\ & -\frac{1}{2}\sum_{i=1}^{n}\left(\widehat{\xi}_i - \mu\right)^\top X\left(\widehat{\xi}_i - \mu\right) \\ = & -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det X - \frac{n}{2}\mathrm{Tr}\left[\widehat{\Sigma}X\right] \\ & -\frac{n}{2}(\widehat{\mu} - \mu)^\top X(\widehat{\mu} - \mu). \end{aligned} \tag{1}$$

Note that $\widehat{\mathcal{L}}(\mu, X)$ is strictly concave in μ and X (Boyd and Vandenberghe 2004, chap. 7) and depends on the training samples only through the sample mean and the sample covariance matrix. It is clear from the last expression that $\widehat{\mathcal{L}}(\mu, X)$ is maximized by $\mu^* = \widehat{\mu}$ for any fixed X. The maximum likelihood estimator X^* for the precision matrix is thus obtained by maximizing $\widehat{\mathcal{L}}(\widehat{\mu}, X)$ over all $X \succ 0$, which is tantamount to solving the convex program

$$\inf_{X \succ 0} -\log\det X + \mathrm{Tr}\left[\widehat{\Sigma}X\right]. \tag{2}$$

If $\widehat{\Sigma}$ is rank deficient, which necessarily happens for $p > n$, then Problem (2) is unbounded. Indeed, expressing the sample covariance matrix as $\widehat{\Sigma} = R\Lambda R^\top$ with R orthogonal and $\Lambda \geq 0$ diagonal, we may set $X_k = R\Lambda_k R^\top$ for any $k \in \mathbb{N}$, where $\Lambda_k \succ 0$ is the diagonal matrix with $(\Lambda_k)_{ii} = 1$ if $\lambda_i > 0$ and $(\Lambda_k)_{ii} = k$ if $\lambda_i = 0$. By construction, the objective value of X_k in (2) tends to $-\infty$ as k grows. If $\widehat{\Sigma}$ is invertible, on the other hand, then the first-order optimality conditions can be solved analytically, showing that the minimum of Problem (2) is attained at $X^* = \widehat{\Sigma}^{-1}$. This implies that maximum likelihood estimation under normality simply recovers the sample covariance matrix but fails to yield a precision matrix estimator for $p > n$.

Adding an ℓ_1-regularization term to its objective function guarantees that Problem (2) has a unique minimizer $X^* \succ 0$, which constitutes a proper (invertible) precision matrix estimator (Hsieh et al. 2014). Moreover, as the ℓ_1-norm represents the convex envelope of the cardinality function on the unit hypercube, the ℓ_1-norm regularized maximum likelihood estimation problem promotes sparse precision matrices that encode interpretable Gaussian graphical models (Banerjee et al. 2008, Friedman et al. 2008). Indeed, under the given normality assumption, one can show that $X_{ij} = 0$ if and only if the random variables ξ_i and ξ_j are conditionally independent given $\{\xi_k\}_{k \notin \{i,j\}}$ (Lauritzen 1996). The sparsity pattern of the precision matrix X thus captures the conditional independence structure of ξ.

In theory, the ℓ_1-norm regularized maximum likelihood estimation problem can be solved in polynomial time via modern interior point algorithms. In practice, however, scalability to high dimensions remains challenging because of the problem's semidefinite nature, and larger problem instances must be addressed with special-purpose methods such as the Newton-type QUIC algorithm (Hsieh et al. 2014).

Instead of penalizing the ℓ_1-norm of the precision matrix, one may alternatively penalize its inverse X^{-1} with the goal of promoting sparsity in the covariance matrix and thus controlling the *marginal* independence structure of ξ (Bien and Tibshirani 2011). Despite its attractive statistical properties, this alternative model leads to a hard nonconvex and nonsmooth optimization problem, which can only be solved approximately.

By the Fisher–Neyman factorization theorem, $\widehat{\Sigma}$ is a sufficient statistic for the true covariance matrix Σ of a normally distributed random vector; that is, $\widehat{\Sigma}$ contains the same information about Σ as the entire training data set. Without any loss of generality, we may thus focus on estimators that depend on the data only through $\widehat{\Sigma}$. If neither the covariance matrix Σ nor the precision matrix Σ^{-1} is known to be sparse, and if there is no prior information about the orientation of their eigenvectors, it is reasonable to restrict attention to *rotation-equivariant* estimators. A precision matrix estimator $\widehat{X}(\widehat{\Sigma})$ is called rotation equivariant if $\widehat{X}(R\widehat{\Sigma}R^\top) = R\widehat{X}(\widehat{\Sigma})R^\top$ for any rotation matrix R. This definition requires that the estimator for the rotated data coincides with the rotated estimator for the original data. One can show that rotation-equivariant estimators have the same eigenvectors as the sample covariance matrix (see, e.g., Perlman (2007), lemma 5.3, for a simple proof) and are thus uniquely determined by their eigenvalues. Hence, imposing rotation equivariance reduces the degrees of freedom from $p(p+1)/2$ to p. Using an entropy loss function introduced in James and Stein (1961), Stein was the first to demonstrate that superior covariance estimators in the sense of statistical decision theory can be constructed by shrinking the eigenvalues of the sample covariance matrix (Stein 1975, 1986). Unfortunately, his proposed shrinkage transformation may alter the order of the eigenvalues and even undermine the positive semidefiniteness of the resulting estimator when $p > n$, which necessitates an ad hoc correction step involving an isotonic regression. Various refinements of this approach are reported in Dey and Srinivasan (1985), Haff (1991), Yang and Berger (1994), and the references therein, but most of these works focus on the low-dimensional case when $n \geq p$.

Jensen's inequality suggests that the largest (respectively, smallest) eigenvalue of the sample covariance matrix $\widehat{\Sigma}$ is biased upward (respectively, downward), which implies that $\widehat{\Sigma}$ tends to be ill-conditioned

(van der Vaart 1961). This effect is most pronounced for $\Sigma \approx I$. A promising shrinkage estimator for the covariance matrix is thus obtained by forming a convex combination of the sample covariance matrix and the identity matrix scaled by the average of the sample eigenvalues (Ledoit 2004a). If its convex weights are chosen optimally in view of the Frobenius risk, the resulting shrinkage estimator can be shown to be both well conditioned and more accurate than $\widehat{\Sigma}$. Alternative shrinkage targets include the constant correlation model, which preserves the sample variances but equalizes all pairwise correlations (Ledoit 2004b); the single-index model, which assumes that each random variable is explained by one systematic and one idiosyncratic risk factor (Ledoit and Wolf 2003); and the diagonal matrix of the sample eigenvalues (Touloumis 2015), for example.

The *linear* shrinkage estimators just described are computationally attractive because evaluating convex combinations is cheap. Computing the corresponding precision matrix estimators requires a matrix inversion and is therefore more expensive. We emphasize that linear shrinkage estimators for the precision matrix itself, obtained by forming a cheap convex combination of the inverse sample covariance matrix and a shrinkage target, are not available in the big data regime when $p > n$ and $\widehat{\Sigma}$ fails to be invertible.

More recently, insights from random matrix theory have motivated a new rotation-equivariant shrinkage estimator that applies an individualized shrinkage intensity to every sample eigenvalue (Ledoit 2012). Although this *nonlinear* shrinkage estimator offers significant improvements over linear shrinkage, its evaluation necessitates the solution of a hard nonconvex optimization problem, which becomes cumbersome for large values of p. Alternative nonlinear shrinkage estimators can be obtained by imposing an upper bound on the condition number of the covariance matrix in the underlying maximum likelihood estimation problem (Won et al. 2013).

Alternatively, multifactor models familiar from the arbitrage pricing theory can be used to approximate the covariance matrix by a sum of a low-rank and a diagonal component, both of which have only a few free parameters and are thus easier to estimate. Such a dimensionality reduction leads to stable estimators (Fan et al. 2008, Chun et al. 2018).

1.2. Problem Statement and Contributions

This paper endeavors to develop a principled approach to precision matrix estimation, which is inspired by recent advances in distributionally robust optimization (Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014). For the sake of argument, assume that the true distribution of ξ is given by $\mathbb{P} = \mathcal{N}(\mu_0, \Sigma_0)$, where $\Sigma_0 \succ 0$. If μ_0 and Σ_0 were known, the quality of some estimators μ and X for μ_0 and Σ_0^{-1}, respectively, could conveniently be measured by Stein's loss (James and Stein 1961):

$$\begin{aligned} L(X,\mu) &:= -\log\det(\Sigma_0 X) + \mathrm{Tr}[\Sigma_0 X] \\ &\quad + (\mu_0-\mu)^\top X(\mu_0-\mu) - p \\ &= -\log\det X + \mathbb{E}^{\mathbb{P}}\left[(\xi-\mu)^\top X(\xi-\mu)\right] \\ &\quad - \log\det\Sigma_0 - p, \end{aligned} \tag{3}$$

which is reminiscent of the log-likelihood function (1). It is easy to verify that Stein's loss is nonnegative for all $\mu \in \mathbb{R}^p$ and $X \in \mathbb{S}_+^p$ and vanishes only at the true mean $\mu = \mu_0$ and the true precision matrix $X = \Sigma_0^{-1}$. Of course, we cannot minimize Stein's loss directly because $\mathbb{P}$ is unknown. As a naïve remedy, one could instead minimize an approximation of Stein's loss obtained by removing the (unknown but irrelevant) normalization constant $-\log\det\Sigma_0 - p$ and replacing $\mathbb{P}$ in (3) with the empirical distribution $\widehat{\mathbb{P}}_n = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$. However, in doing so, we simply recover the standard maximum likelihood estimation problem, which is unbounded for $p > n$ and outputs the sample mean and the inverse sample covariance matrix for $p \leq n$. This motivates us to robustify the empirical loss minimization problem by exploiting that $\widehat{\mathbb{P}}_n$ is close to $\mathbb{P}$ in Wasserstein distance.

Definition 1.1 (Wasserstein Distance)**.** The type-2 Wasserstein distance between two arbitrary distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^p$ with finite second moments is defined as

$$\mathbb{W}(\mathbb{P}_1,\mathbb{P}_2) := \inf_{\Pi}\left\{ \begin{array}{l} \left(\int_{\mathbb{R}^p\times\mathbb{R}^p} \|\xi_1-\xi_2\|^2\,\Pi(\mathrm{d}\xi_1,\mathrm{d}\xi_2)\right)^{\frac{1}{2}}: \\ \quad \Pi \textit{ is a joint distribution of } \xi_1 \textit{ and } \xi_2 \\ \quad \textit{with marginals } \mathbb{P}_1 \textit{ and } \mathbb{P}_2, \textit{respectively} \end{array} \right\}.$$

The squared Wasserstein distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ can be interpreted as the cost of moving the distribution $\mathbb{P}_1$ to the distribution $\mathbb{P}_2$, where $\|\xi_1-\xi_2\|^2$ quantifies the cost of moving unit mass from ξ_1 to ξ_2.

A central limit type theorem for the Wasserstein distance between empirical normal distributions implies that $n\cdot\mathbb{W}(\widehat{\mathbb{P}}_n,\mathbb{P})^2$ converges weakly to a quadratic functional of independent normal random variables as the number n of training samples tends to infinity (Rippl et al. 2016, theorem 2.3). We may thus conclude that for every $\eta \in (0,1)$ there exists $q(\eta) > 0$ such that $\mathbb{P}^n[\mathbb{W}(\widehat{\mathbb{P}}_n,\mathbb{P}) \leq q(\eta)n^{-\frac{1}{2}}] \geq 1-\eta$ for all n large enough. In the following we denote by $\mathcal{N}^p$ the family of all normal distributions on $\mathbb{R}^p$ and by

$$\mathcal{P}_\rho = \left\{\mathbb{Q} \in \mathcal{N}^p : \mathbb{W}\left(\mathbb{Q}, \widehat{\mathbb{P}}_n\right) \leq \rho\right\}$$

the ambiguity set of all normal distributions whose Wasserstein distance to $\widehat{\mathbb{P}}_n$ is at most $\rho \geq 0$. Note that $\mathcal{P}_\rho$ depends on the unknown true distribution $\mathbb{P}$ only through the training data and, for $\rho \geq q(\eta)n^{-\frac{1}{2}}$, contains $\mathbb{P}$ with confidence $1-\eta$ asymptotically as n tends to infinity. It is thus natural to formulate a *distributionally robust* estimation problem for the precision matrix that minimizes Stein's loss—modulo an irrelevant normalization constant—in the worst case across all reference distributions $\mathbb{Q} \in \mathcal{P}_\rho$:

$$\mathcal{J}\left(\widehat{\mu},\widehat{\Sigma}\right) := \inf_{\mu\in\mathbb{R}^p,\, X\in\mathcal{X}} \left\{ -\log\det X + \sup_{\mathbb{Q}\in\mathcal{P}_\rho} \mathbb{E}^{\mathbb{Q}}\left[(\xi-\mu)^\top X(\xi-\mu)\right] \right\}. \tag{4}$$

Here, $\mathcal{X} \subseteq \mathbb{S}_{++}^p$ denotes the set of admissible precision matrices. In the absence of any prior structural information, the only requirement is that X be positive semidefinite and invertible, in which case $\mathcal{X} = \mathbb{S}_{++}^p$. Known conditional independence relationships impose a sparsity pattern on X, which is easily enforced through linear equality constraints in $\mathcal{X}$. By adopting a worst-case perspective, we hope that the minimizers of (4) will have low Stein's loss with respect to all distributions in $\mathcal{P}_\rho$ including the unknown true distribution $\mathbb{P}$. As Stein's loss with respect to the empirical distribution is proportional to the log-likelihood function (1), Problem (4) can also be interpreted as a robust maximum likelihood estimation problem that hedges against perturbations in the training samples. As we will show in what follows, this robustification is tractable and has a regularizing effect.

Recently, it has been discovered that distributionally robust optimization models with Wasserstein ambiguity sets centered at *discrete* distributions on $\mathbb{R}^p$ (and *without* any normality restrictions) are often equivalent to tractable convex programs (Mohajerin Esfahani and Kuhn 2017, Zhao and Guan 2018). Extensions of these results to general Polish spaces are reported in Blanchet and Murthy (2016) and Gao and Kleywegt (2016). The explicit convex reformulations of Wasserstein distributionally robust models have not only facilitated efficient solution procedures but also revealed insightful connections between distributional robustness and regularization in machine learning. Indeed, many classical regularization schemes of supervised learning such as the Lasso method can be explained by a Wasserstein distributionally robust model. This link was first discovered in the context of logistic regression (Shafieezadeh-Abadeh et al. 2015) and later extended to other popular regression and classification models (Blanchet and Murthy 2016, Shafieezadeh-Abadeh et al. 2017), and even to generative adversarial networks in deep learning (Gao et al. 2016).

Model (4) differs fundamentally from all existing distributionally robust optimization models in that the ambiguity set contains only normal distributions. As the family of normal distributions fails to be closed under mixtures, the ambiguity set is thus nonconvex. In the remainder of the paper, we devise efficient solution methods for Problem (4), and we investigate the properties of the resulting precision matrix estimator.

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

- Leveraging an analytical formula for the Wasserstein distance between two normal distributions derived in Givens and Shortt (1984), we prove that the distributionally robust estimation problem (4) is equivalent to a tractable semidefinite program—despite the nonconvex nature of the underlying ambiguity set.
- We prove that Problem (4) and its unique minimizer depend on the training data only through $\widehat{\Sigma}$ (but not through $\widehat{\mu}$), which is reassuring because $\widehat{\Sigma}$ is a sufficient statistic for the precision matrix.
- In the absence of any structural information, we demonstrate that Problem (4) has an analytical solution that is naturally interpreted as a nonlinear shrinkage estimator. Indeed, the optimal precision matrix estimator shares the eigenvectors of the sample covariance matrix, and as the radius ρ of the Wasserstein ambiguity set grows, its eigenvalues are shrunk toward 0 while preserving their order. At the same time, the condition number of the optimal estimator steadily improves and eventually converges to 1 even for $p > n$. These desirable properties are not enforced ex ante but emerge naturally from the underlying distributionally robust optimization model.
- In the presence of conditional independence constraints, the semidefinite program equivalent to (4) is beyond the reach of general-purpose solvers for practically relevant problem dimensions p. We thus devise an efficient sequential quadratic approximation method reminiscent of the QUIC algorithm (Hsieh et al. 2014), which can solve instances of Problem (4) with $p \lesssim 10^4$ on a standard PC.
- We derive an analytical formula for the extremal distribution that attains the supremum in (4).

An important aspect of the distributionally robust estimation problem (4) is the choice of the radius $\rho \geq 0$ of the ambiguity set $\mathcal{P}_\rho$. Ideally, this hyperparameter should be tuned so as to minimize the distance between the precision matrix estimator $X^*(\rho)$ that solves (4) and the unknown true precision matrix Σ^{-1}. While this paper was under review, Blanchet and Si (2019) managed to prove that if distances in $\mathbb{S}_+^p$ are measured via Stein's loss function, and there are no conditional independence constraints, then the Wasserstein radius that minimizes the *expected* distance between the estimator $X^*(\rho)$ and the true precision matrix Σ^{-1} scales

linearly with the sample size as n^{-1}, where the proportionality constant is a function of the true covariance matrix that is known in closed form (Blanchet and Si 2019, theorem 1). This result is surprising vis-à-vis the central limit type theorem (Rippl et al. 2016, theorem 2.3), which suggests a canonical square root scaling of the form $n^{-\frac{1}{2}}$. In practice, ρ should be calibrated in view of the training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_n$—for example, via cross-validation using application-specific performance measures. Concrete examples of different cross-validation schemes are described in Section 6.

The rest of this paper is structured as follows. Section 2 demonstrates that the distributionally robust estimation problem (4) admits an exact reformulation as a tractable semidefinite program. Section 3 derives an analytical solution of this semidefinite program in the absence of any structural information, and Section 4 develops an efficient sequential quadratic approximation algorithm for the problem with conditional independence constraints. The extremal distribution that attains the worst-case expectation in (4) is characterized in Section 5, and numerical experiments based on synthetic and real data are reported in Section 6.

1.2.1. Notation. For any $A \in \mathbb{R}^{p\times p}$, we use $\mathrm{Tr}[A]$ to denote the trace and $\|A\| = \sqrt{\mathrm{Tr}[A^\top A]}$ to denote the Frobenius norm of A. By slight abuse of notation, the Euclidean norm of $v \in \mathbb{R}^p$ is also denoted by $\|v\|$. Moreover, I stands for the identity matrix. Its dimension is usually evident from the context. For any $A, B \in \mathbb{R}^{p\times p}$, we use $\langle A, B\rangle = \mathrm{Tr}[A^\top B]$ to denote the inner product and $A \otimes B \in \mathbb{R}^{p^2\times p^2}$ to denote the Kronecker product of A and B. The space of all symmetric matrices in $\mathbb{R}^{p\times p}$ is denoted by $\mathbb{S}^p$. We use $\mathbb{S}^p_+$ ($\mathbb{S}^p_{++}$) to represent the cone of symmetric positive semidefinite (positive definite) matrices in $\mathbb{S}^p$. For any $A, B \in \mathbb{S}^p$, the relation $A \succeq B$ ($A \succ B$) means that $A - B \in \mathbb{S}^p_+$ ($A - B \in \mathbb{S}^p_{++}$).

2. Tractable Reformulation

Throughout this paper we assume that the random vector $\xi \in \mathbb{R}^p$ is normally distributed. This is in line with the common practice in statistics and in the natural and social sciences, whereby normal distributions are routinely used to model random vectors whose distributions are unknown. The normality assumption is often justified by the central limit theorem, which suggests that random vectors influenced by many small and unrelated disturbances are approximately normally distributed. Moreover, the normal distribution maximizes entropy across all distributions with given first- and second-order moments, and as such, it constitutes the least prejudiced distribution compatible with a given mean vector and covariance matrix.

2.1. Preliminaries

In order to facilitate rigorous statements, we first provide a formal definition of normal distributions.

Definition 2.1 (Normal Distributions)**.** We say that $\mathbb{P}$ is a normal distribution on $\mathbb{R}^p$ with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_+$; that is, $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ if $\mathbb{P}$ is supported on $\mathrm{supp}(\mathbb{P}) = \{\mu + Ev : v \in \mathbb{R}^k\}$ and if the density function of $\mathbb{P}$ with respect to the Lebesgue measure on $\mathrm{supp}(\mathbb{P})$ is given by

$$\varrho_{\mathbb{P}}(\xi) := \frac{1}{\sqrt{(2\pi)^k \det(D)}} e^{-(\xi-\mu)^\top E D^{-1} E^\top (\xi-\mu)},$$

where $k = \mathrm{rank}(\Sigma)$, $D \in \mathbb{S}^p_{++}$ is the diagonal matrix of the positive eigenvalues of Σ, and $E \in \mathbb{R}^{p\times k}$ is the matrix whose columns correspond to the orthonormal eigenvectors of the positive eigenvalues of Σ. The family of all normal distributions on $\mathbb{R}^p$ is denoted by $\mathcal{N}^p$, and the subfamily of all distributions in $\mathcal{N}^p$ with zero means and arbitrary covariance matrices is denoted by $\mathcal{N}^p_0$.

Definition 2.1 explicitly allows for degenerate normal distributions with rank-deficient covariance matrices.

The normality assumption also has distinct computational advantages. In fact, whereas the Wasserstein distance between two generic distributions is only given implicitly as the solution of a mass transportation problem, the Wasserstein distance between two normal distributions is known in closed form. It can be expressed explicitly as a function of the mean vectors and covariance matrices of the two distributions.

Proposition 2.2 (Givens and Shortt (1984, proposition 7))**.** *The type* 2 *Wasserstein distance between two normal distributions* $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ *and* $\mathbb{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$*, with* $\mu_1, \mu_2 \in \mathbb{R}^p$ *and* $\Sigma_1, \Sigma_2 \in \mathbb{S}^p_+$*, amounts to*

$$\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) = \sqrt{\begin{array}{r}\|\mu_1 - \mu_2\|^2 + \mathrm{Tr}[\Sigma_1] + \mathrm{Tr}[\Sigma_2] \\ - 2\mathrm{Tr}\left[\sqrt{\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2}}\right].\end{array}}$$

If $\mathbb{P}_1$ and $\mathbb{P}_2$ share the same mean vector (e.g., if $\mu_1 = \mu_2 = 0$), then the Wasserstein distance $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2)$ reduces to a function of the covariance matrices Σ_1 and Σ_2 only, thereby inducing a metric on the cone $\mathbb{S}^p_+$.

Definition 2.3 (Induced Metric on $\mathbb{S}^p_+$)**.** Let $\mathbb{W}_S : \mathbb{S}^p_+ \times \mathbb{S}^p_+ \to \mathbb{R}_+$ be the metric on $\mathbb{S}^p_+$ induced by the type 2 Wasserstein metric on the family of normal distributions with equal means. Thus, for all $\Sigma_1, \Sigma_2 \in \mathbb{S}^p_+$, we set

$$\mathbb{W}_S(\Sigma_1, \Sigma_2) := \sqrt{\mathrm{Tr}[\Sigma_1] + \mathrm{Tr}[\Sigma_2] - 2\mathrm{Tr}\left[\sqrt{\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2}}\right]}.$$

The definition of $\mathbb{W}_S$ implies via Proposition 2.2 that $\mathbb{W}(\mathbb{P}_1,\mathbb{P}_2)=\mathbb{W}_S(\Sigma_1,\Sigma_2)$ for all $\mathbb{P}_1=\mathcal{N}(\mu_1,\Sigma_1)$ and $\mathbb{P}_2=\mathcal{N}(\mu_2,\Sigma_2)$ with $\mu_1=\mu_2$. Thanks to its interpretation as the restriction of $\mathbb{W}$ to the space of normal distributions with a fixed mean, it is easy to verify that $\mathbb{W}_S$ is symmetric and positive definite and satisfies the triangle inequality. In other words, $\mathbb{W}_S$ inherits the property of being a metric from $\mathbb{W}$.

Corollary 2.4 (Commuting Covariance Matrices)**.** *If $\Sigma_1,\Sigma_2\in\mathbb{S}_+^p$ commute ($\Sigma_1\Sigma_2=\Sigma_2\Sigma_1$), then the induced Wasserstein distance $\mathbb{W}_S$ simplifies to the trace norm between the square roots of Σ_1 and Σ_2; that is,*

$$\mathbb{W}_S(\Sigma_1,\Sigma_2)=\left\|\sqrt{\Sigma_1}-\sqrt{\Sigma_2}\right\|.$$

Proof. The commutativity of Σ_1 and Σ_2 implies that $\sqrt{\Sigma_2}\Sigma_1\sqrt{\Sigma_2}=\Sigma_1\Sigma_2$, whereby

$$\begin{aligned}\mathbb{W}_S(\Sigma_1,\Sigma_2)&=\sqrt{\mathrm{Tr}[\Sigma_1]+\mathrm{Tr}[\Sigma_2]-2\mathrm{Tr}\left[\sqrt{\Sigma_1\Sigma_2}\right]}\\&=\sqrt{\mathrm{Tr}\left[\left(\sqrt{\Sigma_1}-\sqrt{\Sigma_2}\right)^2\right]}\\&=\left\|\sqrt{\Sigma_1}-\sqrt{\Sigma_2}\right\|.\end{aligned}$$

Thus, the claim follows.Proposition 2.2 reveals that the Wasserstein distance between any two (possibly degenerate) normal distributions is finite. By contrast, the Kullback–Leibler divergence between degenerate and nondegenerate normal distributions is infinite.

Remark 2.5 (Kullback–Leibler Divergence Between Normal Distributions)**.** A simple calculation shows that the Kullback–Leibler divergence from $\mathbb{P}_2=\mathcal{N}(\mu_2,\Sigma_2)$ to $\mathbb{P}_1=\mathcal{N}(\mu_1,\Sigma_1)$ amounts to

$$\begin{aligned}D_{\mathrm{KL}}(\mathbb{P}_1\|\mathbb{P}_2)=\frac{1}{2}\big[&(\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1)\\&+\mathrm{Tr}\big[\Sigma_1\Sigma_2^{-1}\big]-p-\log\det\Sigma_1+\log\det\Sigma_2\big]\end{aligned}$$

whenever $\mu_1,\mu_2\in\mathbb{R}^p$ and $\Sigma_1,\Sigma_2\in\mathbb{S}_{++}^p$. If either $\mathbb{P}_1$ or $\mathbb{P}_2$ is degenerate (i.e., if Σ_1 is singular and Σ_2 invertible, or vice versa), then $\mathbb{P}_1$ fails to be absolutely continuous with respect to $\mathbb{P}_2$, which implies that $D_{\mathrm{KL}}(\mathbb{P}_1\|\mathbb{P}_2)=\infty$. Moreover, from the above-mentioned formula, it is easy to verify that $D_{\mathrm{KL}}(\mathbb{P}_1\|\mathbb{P}_2)$ diverges if either Σ_1 or Σ_2 tends to a singular matrix.

In the big data regime ($p>n$), the sample covariance matrix $\widehat{\Sigma}$ is singular even if the samples are drawn from a nondegenerate normal distribution $\mathbb{P}=\mathcal{N}(\mu,\Sigma)$ with $\Sigma\in\mathbb{S}_{++}^p$. In this case, the Kullback–Leibler distance between the empirical distribution $\widehat{\mathbb{P}}=\mathcal{N}(\widehat{\mu},\widehat{\Sigma})$ and $\mathbb{P}$ is infinite, and thus $\widehat{\mathbb{P}}$ and $\mathbb{P}$ are perceived as maximally dissimilar despite their intimate relation. By contrast, their Wasserstein distance is finite.

2.2. Precision Matrix Estimation When the Mean Vector Is Known

Before investigating the general problem (4), we first address a simpler problem variant where the true mean μ_0 of ξ is known to vanish. Thus, we temporarily assume that ξ follows $\mathcal{N}(0,\Sigma_0)$. In this setting, it makes sense to focus on the modified ambiguity set $\mathcal{P}_\rho^0:=\{\mathbb{Q}\in\mathcal{N}_0^p:\mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}})\le\rho\}$, which contains all normal distributions with zero mean that have a Wasserstein distance of at most $\rho\ge0$ from the empirical distribution $\widehat{\mathbb{P}}=\mathcal{N}(0,\widehat{\Sigma})$. Under these assumptions, the estimation problem (4) thus simplifies to

$$\mathcal{J}\left(\widehat{\Sigma}\right):=\inf_{X\in\mathcal{X}}\left\{-\log\det X+\sup_{\mathbb{Q}\in\mathcal{P}_\rho^0}\mathbb{E}^{\mathbb{Q}}\left[\left\langle\xi\xi^\top,X\right\rangle\right]\right\}.\tag{5}$$

We are now ready to state the first main result of this section.

Theorem 2.6 (Convex Reformulation)**.** *For any fixed $\rho>0$ and $\widehat{\Sigma}\succeq0$, the simplified distributionally robust estimation problem (5) is equivalent to*

$$\mathcal{J}\left(\widehat{\Sigma}\right)=\begin{cases}\inf\limits_{X,\gamma} & -\log\det X+\gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)\\ & \qquad+\gamma^2\left\langle(\gamma I-X)^{-1},\widehat{\Sigma}\right\rangle\\ \text{s.t.} & \gamma I\succ X\succ0,\quad X\in\mathcal{X}.\end{cases}\tag{6}$$

Moreover, the optimal value function $\mathcal{J}(\widehat{\Sigma})$ is continuous in $\widehat{\Sigma}\in\mathbb{S}_+^p$.

The proof of Theorem 2.6 relies on several auxiliary results. A main ingredient to derive the convex program (6) is a reformulation of the worst-case expectation function $g:\mathbb{S}_+^p\times\mathbb{S}_+^p\to\mathbb{R}$ defined through

$$g\left(\widehat{\Sigma},X\right):=\sup_{\mathbb{Q}\in\mathcal{P}_\rho^0}\mathbb{E}^{\mathbb{Q}}\left[\left\langle\xi\xi^\top,X\right\rangle\right].\tag{7}$$

In Proposition 2.8 we will demonstrate that $g(\widehat{\Sigma},X)$ is continuous and coincides with the optimal value of an explicit semidefinite program, a result that depends on the following preparatory lemma.

Lemma 2.7 (Continuity Properties of Partial Infima)**.** *Consider a function $\varphi:\mathcal{E}\times\Gamma\to\mathbb{R}$ on two normed spaces $\mathcal{E}$ and Γ, and define the partial infimum with respect to γ as $\Phi(\varepsilon):=\inf_{\gamma\in\Gamma}\varphi(\varepsilon,\gamma)$ for every $\varepsilon\in\mathcal{E}$.*

i. *If $\varphi(\varepsilon,\gamma)$ is continuous in ε at $\varepsilon_0\in\mathcal{E}$ for every $\gamma\in\Gamma$, then $\Phi(\varepsilon)$ is upper semicontinuous at ε_0.*

ii. *If $\varphi(\varepsilon,\gamma)$ is calm from below at $\varepsilon_0\in\mathcal{E}$ uniformly in $\gamma\in\Gamma$—that is, if there exists a constant $L\ge0$ such that $\varphi(\varepsilon,\gamma)-\varphi(\varepsilon_0,\gamma)\ge-L\|\varepsilon_0-\varepsilon\|$ for all $\gamma\in\Gamma$—then $\Phi(\varepsilon)$ is lower semicontinuous at ε_0.*

Proof. As for assertion (i), we have

$$\begin{aligned}\limsup_{\varepsilon\to\varepsilon_0}\Phi(\varepsilon) &= \inf_{\delta>0}\sup_{\|\varepsilon-\varepsilon_0\|\le\delta}\Phi(\varepsilon) = \inf_{\delta>0}\sup_{\|\varepsilon-\varepsilon_0\|\le\delta}\inf_{\gamma\in\Gamma}\varphi(\varepsilon,\gamma)\\ &\le \inf_{\gamma\in\Gamma}\inf_{\delta>0}\sup_{\|\varepsilon-\varepsilon_0\|\le\delta}\varphi(\varepsilon,\gamma)\\ &= \inf_{\gamma\in\Gamma}\limsup_{\varepsilon\to\varepsilon_0}\varphi(\varepsilon,\gamma) = \inf_{\gamma\in\Gamma}\varphi(\varepsilon_0,\gamma) = \Phi(\varepsilon_0),\end{aligned}$$

where the inequality follows from interchanging the infimum and supremum operators, whereas the penultimate equality in the last line relies on the continuity assumption. As for assertion (ii), note that

$$\begin{aligned}\liminf_{\varepsilon\to\varepsilon_0}\Phi(\varepsilon) &= \sup_{\delta>0}\inf_{\|\varepsilon-\varepsilon_0\|\le\delta}\Phi(\varepsilon) = \sup_{\delta>0}\inf_{\|\varepsilon-\varepsilon_0\|\le\delta}\inf_{\gamma\in\Gamma}\varphi(\varepsilon,\gamma)\\ &= \sup_{\delta>0}\inf_{\gamma\in\Gamma}\inf_{\|\varepsilon-\varepsilon_0\|\le\delta}\varphi(\varepsilon,\gamma)\\ &\ge \sup_{\delta>0}\inf_{\gamma\in\Gamma}\inf_{\|\varepsilon-\varepsilon_0\|\le\delta}\big(\varphi(\varepsilon_0,\gamma) - L\|\varepsilon_0-\varepsilon\|\big).\\ &= \sup_{\delta>0}\inf_{\gamma\in\Gamma}\big(\varphi(\varepsilon_0,\gamma) - L\delta\big)\\ &= \inf_{\gamma\in\Gamma}\varphi(\varepsilon_0,\gamma) = \Phi(\varepsilon_0),\end{aligned}$$

where the inequality in the second line holds as a result of the calmness assumption.

Proposition 2.8 (Worst-Case Expectation Function). *For any fixed $\rho>0$, $\widehat{\Sigma}\succeq 0$, and $X\succ 0$, the worst-case expectation $g(\widehat{\Sigma},X)$ defined in (7) coincides with the optimal value of the tractable semidefinite program*

$$\begin{aligned}\inf_{\gamma}\quad & \gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)+\gamma^2\left\langle(\gamma I-X)^{-1},\widehat{\Sigma}\right\rangle\\ \text{s.t.}\quad & \gamma I\succ X.\end{aligned}\tag{8}$$

Moreover, the optimal value function $g(\widehat{\Sigma},X)$ is continuous in $(\widehat{\Sigma},X)\in\mathbb{S}_+^p\times\mathbb{S}_{++}^p$.

Proof. Using the definitions of the worst-case expectation $g(\widehat{\Sigma},X)$ and the ambiguity set $\mathcal{P}_\rho^0$, we find

$$\begin{aligned}g\left(\widehat{\Sigma},X\right) &= \sup_{\mathbb{Q}\in\mathcal{P}_\rho^0}\left\langle\mathbb{E}^{\mathbb{Q}}\left[\xi\xi^\top\right],X\right\rangle\\ &= \sup_{S\in\mathbb{S}_+^p}\left\{\langle S,X\rangle:\mathbb{W}_S\left(S,\widehat{\Sigma}\right)\le\rho\right\},\end{aligned}$$

where the second equality holds because the metric $\mathbb{W}_S$ on $\mathbb{S}_+^p$ is induced by the type 2 Wasserstein metric $\mathbb{W}$ on $\mathcal{M}_0^p$, meaning that there is a one-to-one correspondence between distributions $\mathbb{Q}\in\mathcal{M}_0^p$ with $\mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}})\le\rho$ and covariance matrices $S\in\mathbb{S}_+^p$ with $\mathbb{W}_S(S,\widehat{\Sigma})\le\rho$. The continuity of $g(\widehat{\Sigma},X)$ thus follows from Berge's (1963) maximum theorem (pp. 115–116), which applies because $\langle S,X\rangle$ and $\mathbb{W}_S(S,\widehat{\Sigma})$ are continuous[2] in $(S,\widehat{\Sigma},X)\in\mathbb{S}_+^p\times\mathbb{S}_+^p\times\mathbb{S}_{++}^p$, whereas $\{S\in\mathbb{S}_+^p:\mathbb{W}_S(S,\widehat{\Sigma})\le\rho\}$ is nonempty and compact for every $\widehat{\Sigma}\in\mathbb{S}_+^p$ and $\rho>0$.

By the definition of the induced metric $\mathbb{W}_S$, we then obtain

$$g\left(\widehat{\Sigma},X\right) = \sup_{S\in\mathbb{S}_+^p}\left\{\langle S,X\rangle:\mathrm{Tr}\left[\widehat{\Sigma}\right]+\mathrm{Tr}[S]-2\mathrm{Tr}\left[\sqrt{\widehat{\Sigma}^{\frac12}S\widehat{\Sigma}^{\frac12}}\right]\le\rho^2\right\}.\tag{9}$$

To establish the equivalence between (8) and (9), we first assume that $\widehat{\Sigma}\succ 0$. The generalization to rank-deficient sample covariance matrices will be addressed later. By dualizing the explicit constraint in (9) and introducing the constant matrix $M=\widehat{\Sigma}^{\frac12}$, which inherits invertibility from $\widehat{\Sigma}$, we find

$$\begin{aligned}g\left(\widehat{\Sigma},X\right) &= \sup_{S\in\mathbb{S}_+^p}\inf_{\gamma\ge0}\langle S,X-\gamma I\rangle + 2\gamma\left\langle\sqrt{MSM},I\right\rangle+\gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)\\ &= \inf_{\gamma\ge0}\sup_{S\in\mathbb{S}_+^p}\langle S,X-\gamma I\rangle + 2\gamma\left\langle\sqrt{MSM},I\right\rangle+\gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)\\ &= \inf_{\gamma\ge0}\Big\{\gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)+\sup_{B\in\mathbb{S}_+^p}\big\{\left\langle B^2,M^{-1}(X-\gamma I)M^{-1}\right\rangle+2\gamma\langle B,I\rangle\big\}\Big\}.\end{aligned}\tag{10}$$

Here, the first equality exploits the identity $\mathrm{Tr}[A]=\langle A,I\rangle$ for any $A\in\mathbb{R}^{p\times p}$; the second equality follows from strong duality, which holds because $\widehat{\Sigma}$ constitutes a Slater point for Problem (9) when $\rho>0$; and the third equality relies on the substitution $B\leftarrow\sqrt{MSM}$, which implies that $S=M^{-1}B^2M^{-1}$. Introducing the shorthand $\Delta=M^{-1}(X-\gamma I)M^{-1}$ allows us to simplify the inner maximization problem over B in (10) to

$$\sup_{B\in\mathbb{S}_+^p}\left\{\left\langle B^2,\Delta\right\rangle+2\gamma\langle B,I\rangle\right\}.\tag{11}$$

If $\Delta\not\prec 0$, then (11) is unbounded. To see this, denote by $\overline{\lambda}(\Delta)$ the largest eigenvalue of Δ and by $\overline{v}$ a corresponding eigenvector. If $\overline{\lambda}(\Delta)>0$, then the objective value of $B_k=k\cdot\overline{v}\,\overline{v}^\top\succeq 0$ in (11) grows quadratically with k. If $\overline{\lambda}(\Delta)=0$, then $\gamma>0$, for otherwise, $X\preceq 0$ contrary to our assumption, and thus the objective value of B_k in (11) grows linearly with k. In both cases, (11) is indeed unbounded.

If $\Delta\prec 0$, then (11) becomes a convex optimization problem that can be solved analytically. Indeed, the objective function of (11) is minimized by $B^*=-\gamma\Delta^{-1}$, which satisfies the first-order optimality condition

$$B\Delta+\Delta B+2\gamma I=0\tag{12}$$

and is strictly feasible in (11) because $\Delta < 0$. Moreover, as (12) is naturally interpreted as a continuous Lyapunov equation, its solution B^* can be shown to be unique; see, for example, Hespanha (2009, theorem 12.5). We may thus conclude that B^* is the unique maximizer of (11) and that the maximum of (11) amounts to $-\gamma^2 \mathrm{Tr}[\Delta^{-1}]$.

Adding the constraint $\gamma I \succ X$ to the outer minimization problem in (10), thus excluding all values of γ for which $\Delta \not\prec 0$ and the inner supremum is infinite, and replacing the optimal value of the inner maximization problem with $-\gamma^2 \mathrm{Tr}[\Delta^{-1}] = \gamma^2 \langle (\gamma I - X)^{-1}, \widehat{\Sigma} \rangle$ yields (8). This establishes the claim for $\widehat{\Sigma} \succ 0$.

In the second part of the proof, we show that the claim remains valid for rank-deficient sample covariance matrices. To this end, we denote the optimal value of Problem (8) by $g'(\widehat{\Sigma}, X)$. From the first part of the proof we know that $g'(\widehat{\Sigma}, X) = g(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma}, X \in \mathbb{S}_{++}^p$. We also know that $g(\widehat{\Sigma}, X)$ is continuous in $(\widehat{\Sigma}, X) \in \mathbb{S}_+^p \times \mathbb{S}_{++}^p$. It remains to be shown that $g'(\widehat{\Sigma}, X) = g(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma} \in \mathbb{S}_+^p$ and $X \in \mathbb{S}_{++}^p$.

Fix any $\widehat{\Sigma} \in \mathbb{S}_+^p$ and $X \in \mathbb{S}_{++}^p$, and note that $\widehat{\Sigma} + \varepsilon I \succ 0$ for every $\varepsilon > 0$. Defining the intervals $\mathcal{E} = \mathbb{R}_+$ and $\Gamma = \{\gamma \in \mathbb{R} : \gamma I \succ X\}$ as well as the auxiliary functions

$$\Phi(\varepsilon) = g'\left(\widehat{\Sigma} + \varepsilon I, X\right) \quad \text{and}$$
$$\varphi(\varepsilon, \gamma) = \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma} + \varepsilon I\right]\right) + \gamma^2 \left\langle (\gamma I - X)^{-1}, \widehat{\Sigma} + \varepsilon I \right\rangle,$$

it follows from (8) that

$$\Phi(\varepsilon) = \inf_{\gamma \in \Gamma} \varphi(\varepsilon, \gamma) \quad \forall \varepsilon \in \mathcal{E}.$$

One can show via Lemma 2.7 that $\Phi(\varepsilon)$ is continuous at $\varepsilon = 0$. Indeed, $\varphi(\varepsilon, \gamma)$ is linear and thus continuous in ε for every $\gamma \in \Gamma$, which implies via part (a) of Lemma 2.7 that $\Phi(\varepsilon)$ is upper semicontinuous at $\varepsilon = 0$. Moreover, $\varphi(\varepsilon, \gamma)$ is calm from below at $\varepsilon = 0$ with $L = 0$ uniformly in $\gamma \in \Gamma$ because

$$\varphi(\varepsilon, \gamma) - \varphi(0, \gamma) = \gamma \mathrm{Tr}\left[\left(I - \gamma^{-1} X\right)^{-1} - I\right]\varepsilon \geq 0 \quad \forall \gamma \in \Gamma.$$

Here, the inequality holds for all $\gamma \in \Gamma$ because of the conditions $I \succ \gamma^{-1} X \succ 0$, which are equivalent to $0 \prec I - \gamma^{-1} X \prec I$ and imply $(I - \gamma^{-1} X)^{-1} \succ I$. Lemma 2.7, part (b) thus ensures that $\Phi(\varepsilon)$ is lower semicontinuous at $\varepsilon = 0$. In summary, we conclude that $\Phi(\varepsilon)$ is indeed continuous at $\varepsilon = 0$.

Combining these results, we find

$$g\left(\widehat{\Sigma}, X\right) = \lim_{\varepsilon \to 0^+} g\left(\widehat{\Sigma} + \varepsilon I, X\right) = \lim_{\varepsilon \to 0^+} g'\left(\widehat{\Sigma} + \varepsilon I, X\right)$$
$$= \lim_{\varepsilon \to 0^+} \Phi(\varepsilon) = \Phi(0) = g'\left(\widehat{\Sigma}, X\right),$$

where the first inequality holds because of the continuity of $g(\widehat{\Sigma}, X)$ in $\widehat{\Sigma}$, the second from the fact that $g(\widehat{\Sigma}, X) = g'(\widehat{\Sigma}, X)$ for all $\widehat{\Sigma} \succ 0$, the third from the definition of $\Phi(\varepsilon)$, the fourth because of the continuity of $\Phi(\varepsilon)$ at $\varepsilon = 0$, and the fifth, once again, from the definition of $\Phi(\varepsilon)$. The claim now follows because $\widehat{\Sigma} \in \mathbb{S}_+^p$ and $X \in \mathbb{S}_{++}^p$ were chosen arbitrarily.

We have now collected all necessary ingredients for the proof of Theorem 2.6.

Proof of Theorem 2.6. By Proposition 2.8, the worst-case expectation in (5) coincides with the optimal value of the semidefinite program (8). Substituting this semidefinite program into (5) yields (6). Note that the condition $X \succ 0$, which ensures that $\log \det X$ is well defined, is actually redundant because it is implied by the constraint $X \in \mathcal{X}$. Nevertheless, we make it explicit in (6) for the sake of clarity.

It remains to show that $\mathcal{J}(\widehat{\Sigma})$ is continuous. To this end, we first construct bounds on the minimizers of (6) that vary continuously with $\widehat{\Sigma}$. Such bounds can be constructed from any feasible decision (X_0, γ_0). Assume without loss of generality that $\gamma_0 > p/\rho^2$, and denote by $f_0(\widehat{\Sigma})$ the objective value of (X_0, γ_0) in (6), which constitutes a linear function of $\widehat{\Sigma}$. Moreover, define two continuous auxiliary functions

$$\overline{x}\left(\widehat{\Sigma}\right) := \frac{f_0\left(\widehat{\Sigma}\right) - p(1 - \log \gamma_0)}{\rho^2 - p\gamma_0^{-1}} \quad \text{and}$$
$$\underline{x}\left(\widehat{\Sigma}\right) := \frac{e^{-f_0\left(\widehat{\Sigma}\right)}}{\overline{x}\left(\widehat{\Sigma}\right)^{p-1}}, \tag{13}$$

which are strictly positive because $\gamma_0 > p/\rho^2$. Clearly, the infimum of Problem (6) is determined only by feasible decisions (X, γ) with an objective value of at most $f_0(\widehat{\Sigma})$. All such decisions satisfy

$$\begin{aligned} f_0\left(\widehat{\Sigma}\right) &\geq -\log \det X + \gamma\rho^2 + \gamma\left\langle \left(I - \gamma^{-1} X\right)^{-1} - I, \widehat{\Sigma} \right\rangle \\ &\geq -\log \det X + \gamma\rho^2 \\ &\geq -p \log \gamma + \gamma\rho^2 \geq \left(\rho^2 - \gamma_0^{-1} p\right)\gamma + p(1 - \log \gamma_0), \end{aligned} \tag{14}$$

where the second and third inequalites exploit the estimates $(I - \gamma^{-1} X)^{-1} \succ I$ and $\det X \leq \det(\gamma I) = \gamma^p$, respectively, which are both implied by the constraint $\gamma I \succ X \succ 0$, and the last inequality holds because $\log \gamma \leq \log \gamma_0 + \gamma_0^{-1}(\gamma - \gamma_0)$ for all $\gamma > 0$. By rearranging the above-mentioned inequality and recalling the definition of $\overline{x}(\widehat{\Sigma})$, we thus find $\gamma \leq \overline{x}(\widehat{\Sigma})$, which, in turn, implies that $X \prec \gamma I \leq \overline{x}(\widehat{\Sigma}) I$.

Denoting by $\{x_i\}_{i\leq p}$ the eigenvalues of the matrix X and setting $x_{\min} = \min_{i\leq p} x_i$, we further find

$$\begin{aligned} f_0\left(\widehat{\Sigma}\right) &\geq -\log\det X = -\log\left(\prod_{i=1}^{p} x_i\right) \\ &\geq -\log\left(x_{\min}\,\overline{x}\left(\widehat{\Sigma}\right)^{p-1}\right) \\ &= -\log x_{\min} - (p-1)\log\overline{x}\left(\widehat{\Sigma}\right), \end{aligned}$$

where the first inequality follows from (14), whereas the second inequality is based on overestimating all but the smallest eigenvalue of X by $\overline{x}(\widehat{\Sigma})$. By rearranging the above-mentioned inequality and recalling the definition of $\underline{x}(\widehat{\Sigma})$, we thus find $x_{\min} \geq \underline{x}(\widehat{\Sigma})$, which, in turn, implies that $X \succeq \underline{x}(\widehat{\Sigma})I$.

This reasoning shows that the extra constraint $\underline{x}(\widehat{\Sigma})I \preceq X \preceq \overline{x}(\widehat{\Sigma})I$ has no impact on (6); that is,

$$\begin{aligned} \mathcal{J}\left(\widehat{\Sigma}\right) &= \begin{cases} \inf\limits_{X} & -\log\det X + \inf\limits_{\gamma}\left\{\gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right)\right. \\ & \qquad \left. + \gamma^2\left\langle (\gamma I - X)^{-1}, \widehat{\Sigma}\right\rangle : \gamma I \succ X\right\} \\ \text{s.t.} & X \in \mathcal{X}, \quad \underline{x}\left(\widehat{\Sigma}\right)I \preceq X \preceq \overline{x}\left(\widehat{\Sigma}\right)I \end{cases} \\ &= \begin{cases} \inf\limits_{X} & -\log\det X + g\left(\widehat{\Sigma}, X\right) \\ \text{s.t.} & X \in \mathcal{X}, \quad \underline{x}\left(\widehat{\Sigma}\right)I \preceq X \preceq \overline{x}\left(\widehat{\Sigma}\right)I, \end{cases} \end{aligned}$$

where the second equality follows from Proposition 2.8. The continuity of $\mathcal{J}(\widehat{\Sigma})$ now follows directly from Berge's (1963) maximum theorem (pp. 115–116), which applies as a result of the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 2.8, the compactness of the feasible set, and the continuity of $\underline{x}(\widehat{\Sigma})$ and $\overline{x}(\widehat{\Sigma})$.

An immediate consequence of Theorem 2.6 is that the simplified estimation problem (5) is equivalent to an explicit semidefinite program and is therefore, in principle, computationally tractable.

Corollary 2.9 (Tractability). *For any fixed $\rho > 0$ and $\widehat{\Sigma} \succeq 0$, the simplified distributionally robust estimation problem (5) is equivalent to the tractable semidefinite program*

$$\mathcal{J}\left(\widehat{\Sigma}\right) = \begin{cases} \inf\limits_{X,Y,\gamma} & -\log\det X + \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + \mathrm{Tr}[Y] \\ \text{s.t.} & \begin{bmatrix} Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & \gamma I - X \end{bmatrix} \succeq 0, \\ & \gamma I \succ X \succ 0, \quad Y \succeq 0, \quad X \in \mathcal{X}. \end{cases} \tag{15}$$

Proof. We know from Theorem 2.6 that the estimation problem (5) is equivalent to the convex program (6). As X represents a decision variable instead of a parameter, however, Problem (6) fails to be a semidefinite program per se. Indeed, its objective function involves the nonlinear term $h(X,\gamma) := \gamma^2\langle(\gamma I - X)^{-1}, \widehat{\Sigma}\rangle$, which is interpreted as ∞ outside of its domain $\{(X,\gamma) \in \mathbb{S}_+^p \times \mathbb{R} : \gamma I \succ X\}$. However, $h(X,\gamma)$ constitutes a matrix fractional function as described in Boyd and Vandenberghe (2004, example 3.4) and thus admits the semidefinite reformulation:

$$\begin{aligned} h(X,\gamma) &= \inf_{t}\left\{t : \gamma I \succ X, \;\; \gamma^2\left\langle (\gamma I - X)^{-1}, \widehat{\Sigma}\right\rangle \leq t\right\} \\ &= \inf_{Y,t}\left\{t : \gamma I \succ X, \;\; Y \succeq \gamma^2\widehat{\Sigma}^{\frac{1}{2}}(\gamma I - X)^{-1}\widehat{\Sigma}^{\frac{1}{2}}, \; \mathrm{Tr}[Y] \leq t\right\} \\ &= \inf_{Y}\left\{\mathrm{Tr}[Y] : \gamma I \succ X, \;\; \begin{bmatrix} Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & \gamma I - X \end{bmatrix} \succeq 0\right\}, \end{aligned}$$

where the second equality holds because $A \succeq B$ implies $\mathrm{Tr}[A] \geq \mathrm{Tr}[B]$, whereas the third equality follows from a standard Schur complement argument; see, for example, Boyd and Vandenberghe (2004, appendix A.5.5). Thus, $h(X,\gamma)$ is representable as the optimal value of a parametric semidefinite program whose objective and constraint functions are jointly convex in the auxiliary decision variable Y and the parameters X and γ. The postulated reformulation (15) is then obtained by substituting the last expression into (6).

2.3. Joint Estimation of the Mean Vector and the Precision Matrix

Now that we have derived a tractable semidefinite reformulation for the simplified estimation problem (5), we are ready to address the generic estimation problem (4), which does *not* assume knowledge of the mean and is robustified against all distributions in the ambiguity set $\mathcal{P}_\rho$ *without* mean constraints.

Theorem 2.10 (Sufficiency of $\widehat{\Sigma}$). *For any fixed $\rho > 0$, $\widehat{\mu} \in \mathbb{R}^p$, and $\widehat{\Sigma} \in \mathbb{S}_+^p$, the general distributionally robust estimation problem (4) is equivalent to the optimization problem (6) and the tractable semidefinite program (15). Moreover, the optimal value function $\mathcal{J}(\widehat{\mu}, \widehat{\Sigma})$ is constant in $\widehat{\mu}$ and continuous in $\widehat{\Sigma}$.*

Proof. By Proposition 2.2, the optimal value of the estimation problem (4) can be expressed as

$$\begin{aligned} \mathcal{J}\left(\widehat{\mu}, \widehat{\Sigma}\right) &= \inf_{\mu, X\in\mathcal{X}} -\log\det X \\ &\quad + \begin{cases} \sup\limits_{\mu', S\succeq 0} & (\mu' - \mu)^\top X(\mu' - \mu) + \langle S, X\rangle \\ \text{s.t.} & \mathrm{Tr}[S] + \mathrm{Tr}\left[\widehat{\Sigma}\right] - 2\mathrm{Tr}\left[\sqrt{\widehat{\Sigma}^{\frac{1}{2}} S \widehat{\Sigma}^{\frac{1}{2}}}\right] \\ & \qquad \leq \rho^2 - \left\|\mu' - \widehat{\mu}\right\|^2 \end{cases} \\ &= \inf_{\mu, X\in\mathcal{X}} -\log\det X + \sup_{\mu' : \|\mu' - \widehat{\mu}\| \leq \rho} (\mu' - \mu)^\top X(\mu' - \mu) \\ &\quad + \inf_{\gamma : \gamma I \succ X}\left\{\gamma\left(\rho^2 - \left\|\mu' - \widehat{\mu}\right\|^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right)\right. \\ &\quad \left. + \gamma^2\left\langle (\gamma I - X)^{-1}, \widehat{\Sigma}\right\rangle\right\}. \end{aligned}$$

Here, the second equality holds because the Wasserstein constraint is infeasible unless $\|\mu' - \widehat{\mu}\| \leq \rho$ and because the maximization problem over S, which constitutes an instance of (9) with $\rho^2 - \|\mu' - \widehat{\mu}\|^2$ instead of ρ^2, can be reformulated as a minimization problem over γ thanks to Proposition 2.8. By the minimax theorem (Bertsekas 2009, proposition 5.5.4), which applies because μ' ranges over a compact ball and because $X - \gamma I \prec 0$, we may then interchange the maximization over μ' with the minimization over γ to obtain

$$\begin{aligned}\mathcal{J}\left(\widehat{\mu}, \widehat{\Sigma}\right) = \inf_{\substack{\mu, X \in \mathcal{X}, \\ \gamma : \gamma I > X}} & -\log\det X \\ & + \sup_{\mu' : \|\mu' - \widehat{\mu}\| \leq \rho} (\mu' - \mu)^\top X (\mu' - \mu) \\ & + \gamma\left(\rho^2 - \|\mu' - \widehat{\mu}\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) \\ & + \gamma^2 \left\langle (\gamma I - X)^{-1}, \widehat{\Sigma} \right\rangle.\end{aligned}$$

Using the minimax theorem (Bertsekas 2009, proposition 5.5.4) once again to interchange the minimization over μ with the maximization over μ' yields

$$\begin{aligned}\mathcal{J}\left(\widehat{\mu}, \widehat{\Sigma}\right) &= \inf_{\substack{X \in \mathcal{X}, \\ \gamma : \gamma I > X}} -\log\det X \\ &\quad + \sup_{\mu' : \|\mu' - \widehat{\mu}\| \leq \rho} \inf_{\mu} (\mu' - \mu)^\top X (\mu' - \mu) \\ &\quad + \gamma\left(\rho^2 - \|\mu' - \widehat{\mu}\|^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) \\ &\quad + \gamma^2 \left\langle (\gamma I - X)^{-1}, \widehat{\Sigma} \right\rangle \\ &= \inf_{\substack{X \in \mathcal{X}, \\ \gamma : \gamma I > X}} -\log\det X \\ &\quad + \gamma\left(\rho^2 - \operatorname{Tr}\left[\widehat{\Sigma}\right]\right) + \gamma^2 \left\langle (\gamma I - X)^{-1}, \widehat{\Sigma} \right\rangle,\end{aligned}$$

where the second equality holds because μ' is the unique optimal solution of the innermost minimization problem over μ, whereas $\widehat{\mu}$ is the unique optimal solution of the maximization problem over μ'. Thus, the general estimation problem (4) is equivalent to (6), and $\mathcal{J}(\widehat{\mu}, \widehat{\Sigma})$ is manifestly constant in $\widehat{\mu}$. Theorem 2.6 further implies that $\mathcal{J}(\widehat{\mu}, \widehat{\Sigma})$ is continuous in $\widehat{\Sigma}$, whereas Corollary 2.9 implies that (4) is equivalent to the tractable semidefinite program (15). These observations complete the proof.

Theorem 2.10 asserts that the general estimation problem (4) is equivalent to the simplified estimation problem (5), which is based on the hypothesis that the mean of ξ is known to vanish. Theorem 2.10 further reveals that the general estimation problem (4) as well as its (unique) optimal solution depends on the training data only through the sample covariance matrix $\widehat{\Sigma}$. This is reassuring because $\widehat{\Sigma}$ is known to be a sufficient statistic for the precision matrix. As solving (4) is tantamount to solving (5), it suffices to devise solution procedures for the simplified estimation problem (5) or its equivalent reformulations (6) and (15).

We emphasize that the strictly convex log-determinant term in the objective of (15) is supported by state-of-the-art interior point solvers for semidefinite programs such as SDPT3 (Tütüncü et al. 2003). In principle, Problem (15) can therefore be implemented directly in MATLAB using the YALMIP interface (Löfberg 2004), for instance. In spite of its theoretical tractability, however, the semidefinite program (15) quickly becomes excruciatingly large, and direct solution with a general-purpose solver becomes impracticable already for moderate values of p. This motivates us to investigate practically relevant special cases in which the estimation problem (5) can be solved either analytically (Section 3) or numerically using a dedicated fast Newton-type algorithm (Section 4).

3. Analytical Solution Without Sparsity Information

If we have no prior information about the precision matrix, it is natural to set $\mathcal{X} = \mathbb{S}_{++}^p$. In this case, the distributionally robust estimation Problem (5) can be solved in quasi-closed form.

Theorem 3.1. (Analytical Solution Without Sparsity Information). *If $\rho > 0$, $\mathcal{X} = \mathbb{S}_{++}^p$, and $\widehat{\Sigma} \in \mathbb{S}_+^p$ admits the spectral decomposition $\widehat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^\top$ with eigenvalues λ_i and corresponding orthonormal eigenvectors v_i, $i \leq p$, then the unique minimizer of (5) is given by $X^* = \sum_{i=1}^p x_i^* v_i v_i^\top$, where*

$$x_i^* = \gamma^*\left[1 - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*} - \lambda_i\gamma^*\right)\right] \quad \forall i \leq p, \tag{16a}$$

and $\gamma^* > 0$ is the unique positive solution of the algebraic equation

$$\left(\rho^2 - \frac{1}{2}\sum_{i=1}^p \lambda_i\right)\gamma - p + \frac{1}{2}\sum_{i=1}^p \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma} = 0. \tag{16b}$$

Proof. We first demonstrate that the algebraic equation (16b) admits a unique solution in $\mathbb{R}_+$. For ease of exposition, we define $\varphi(\gamma)$ as the left-hand side of (16b). It is easy to see that $\varphi(0) = -p < 0$ and $\lim_{\gamma\to\infty} \varphi(\gamma)/\gamma = \rho^2$, which implies that $\varphi(\gamma)$ grows asymptotically linearly with γ at slope $\rho^2 > 0$. By the intermediate value theorem, we may thus conclude that Equation (16b) has a solution $\gamma^* > 0$.

As $\lambda_i\gamma + 2 > \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma}$, the derivative of $\varphi(\gamma)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\varphi(\gamma) = \rho^2 + \frac{1}{2}\sum_{i=1}^p \lambda_i\left(\frac{\lambda_i\gamma + 2}{\sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma}} - 1\right) > 0,$$

whereby $\varphi(\gamma)$ is strictly increasing in $\gamma \in \mathbb{R}_+$. Thus, the solution γ^* is unique. The positive slope of $\varphi(\gamma)$ further implies via the implicit function theorem that γ^* changes continuously with $\lambda_i \in \mathbb{R}_+$, $i \leq p$.

In analogy to Proposition 2.8, we prove the claim first under the assumption that $\widehat{\Sigma} > 0$ and postpone the generalization to rank-deficient sample covariance matrices. Focusing on $\widehat{\Sigma} > 0$, we will show that (X^*, γ^*) is feasible and optimal in (6). By Theorem 2.6, this will imply that X^* is feasible and optimal in (5).

As $\gamma^* > 0$ and $\widehat{\Sigma} > 0$, which means that $\lambda_i > 0$ for all $i \leq p$, an elementary calculation shows that

$$2 > \sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*} - \lambda_i\gamma^* > 0 \quad \Longleftrightarrow$$
$$1 > 1 - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*} - \lambda_i\gamma^*\right) > 0.$$

Multiplying the last inequality by γ^* proves that $\gamma^* > x_i^* > 0$ for all $i \leq p$, which, in turn, implies that $\gamma^* I > X^* > 0$. Thus, (X^*, γ^*) is feasible in (6), and X^* is feasible in (5).

To prove optimality, we denote by $f(X, \gamma)$ the objective function of Problem (6) and note that its gradient with respect to X vanishes at (X^*, γ^*). Indeed, we have

$$\begin{aligned}\nabla_X f(X^*, \gamma^*) &= -(X^*)^{-1} + (\gamma^*)^2(\gamma^* I - X^*)^{-1}\widehat{\Sigma}(\gamma^* I - X^*)^{-1} \\ &= \sum_{i=1}^{p}\left((\gamma^*)^2(\gamma^* - x_i^*)^{-2}\lambda_i - (x_i^*)^{-1}\right)v_i v_i^\top \\ &= \sum_{i=1}^{p}\frac{(\gamma^*)^2 x_i^*\lambda_i - (\gamma^* - x_i^*)^2}{(\gamma^* - x_i^*)^2 x_i}v_i v_i^\top = 0,\end{aligned}$$

where the first equality exploits the basic rules of matrix calculus (see, e.g., Bernstein 2009, p. 631), the second equality holds because $\widehat{\Sigma}$ and X share the same eigenvectors v_i, $i \leq p$, and the last equation follows from the identity

$$(\gamma^*)^2 x_i^*\lambda_i = (\gamma^* - x_i^*)^2 \quad \forall i \leq p, \tag{17}$$

which is a direct consequence of the definitions of γ^* and x_i^*, $i \leq p$, in (16). Similarly, the partial derivative of $f(X, \gamma)$ with respect to γ vanishes at (X^*, γ^*), too. In fact, we have

$$\begin{aligned}\frac{\partial}{\partial\gamma}f(X^*, \gamma^*) &= \rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right] + 2\gamma^*\mathrm{Tr}\left[(\gamma^* I - X^*)^{-1}\widehat{\Sigma}\right] \\ &\quad - (\gamma^*)^2\mathrm{Tr}\left[(\gamma^* I - X^*)^{-1}\widehat{\Sigma}(\gamma^* I - X^*)^{-1}\right] \\ &= \rho^2 - \sum_{i=1}^{p}\lambda_i\left(1 - \frac{2\gamma^*}{\gamma^* - x_i^*} + \frac{(\gamma^*)^2}{(\gamma^* - x_i^*)^2}\right) \\ &= \rho^2 - \sum_{i=1}^{p}\frac{(x_i^*)^2}{(\gamma^* - x_i^*)^2}\lambda_i \\ &= \frac{1}{(\gamma^*)^2}\left(\rho^2(\gamma^*)^2 - \sum_{i=1}^{p}x_i^*\right) = 0,\end{aligned}$$

where the second equality expresses $\widehat{\Sigma}$ and X in terms of their respective spectral decompositions, the fourth equality holds because of (17), and the last equality follows from the observation that $\rho^2(\gamma^*)^2 = \sum_{i=1}^{p} x_i^*$. In summary, we have shown that (X^*, γ^*) satisfies the first-order optimality conditions of the convex optimization problem (6), which ensures that X^* is optimal in (5).

Consider now any (possibly singular) sample covariance matrix $\widehat{\Sigma} \in \mathbb{S}_+^p$. As $\gamma^* > 0$, similar arguments as in the first part of the proof show that $\gamma^* \geq x_i^* > 0$ for all $i \leq p$, which, in turn, implies that $\gamma^* I \geq X^* > 0$. Moreover, if $\widehat{\Sigma}$ has at least one zero eigenvalue, it is easy to see that $\gamma^* I \not\succ X^*$, in which case (X^*, γ^*) fails to be feasible in (6). However, X^* remains feasible and optimal in (5). To see this, consider the invertible sample covariance matrix $\widehat{\Sigma} + \varepsilon I > 0$ for some $\varepsilon > 0$, and denote by $(X^*(\varepsilon), \gamma^*(\varepsilon))$ the corresponding minimizer of Problem (6) as constructed in (16). As the solution of the algebraic equation (16b) depends continuously on the eigenvalues of the sample covariance matrix, we conclude that the auxiliary variable $\gamma^*(\varepsilon)$ and—by virtue of (16a)—the estimator $X^*(\varepsilon)$ are both continuous in $\varepsilon \in \mathbb{R}_+$. Thus, we find

$$\begin{aligned}\mathcal{J}\left(\widehat{\Sigma}\right) &= \lim_{\varepsilon\to 0^+}\mathcal{J}\left(\widehat{\Sigma} + \varepsilon I\right) \\ &= \lim_{\varepsilon\to 0^+} -\log\det X^*(\varepsilon) + g\left(\widehat{\Sigma} + \varepsilon I, X^*(\varepsilon)\right) \\ &= -\log\det X^* + g\left(\widehat{\Sigma}, X^*\right),\end{aligned}$$

where the first equality follows from the continuity of $\mathcal{J}(\widehat{\Sigma})$ established in Theorem 2.6, the second equality holds because $X^*(\varepsilon)$ is the optimal estimator corresponding to the sample covariance matrix $\widehat{\Sigma} + \varepsilon I > 0$ in Problem (5), and the third equality follows from the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 2.8 and the fact that $\lim_{\varepsilon\to 0^+} X^*(\varepsilon) = X^* > 0$. Thus, X^* is indeed optimal in (5). The strict convexity of $-\log\det X$ further implies that X^* is unique. This observation completes the proof.

Remark 3.2. (Properties of X^*). The optimal distributionally robust estimator X^* identified in Theorem 3.1 commutes with the sample covariance matrix $\widehat{\Sigma}$ because both matrices share the same eigenbasis. Moreover, the eigenvalues of X^* are obtained from those of $\widehat{\Sigma}$ via a nonlinear transformation that depends on the size ρ of the ambiguity set. We emphasize that all eigenvalues of X^* are positive for every $\rho > 0$, which implies that X^* is invertible. These insights suggest that X^* constitutes a nonlinear shrinkage estimator, which enjoys the rotation equivariance property (when all data points are rotated by $R \in \mathbb{R}^{p\times p}$, then X^* changes to $RX^*R^\top$).

Theorem 3.1 characterizes the optimal solution of Problem (5) in quasi-closed form up to the spectral decomposition of $\widehat{\Sigma}$ and the numerical solution of Equation (16b). By Pan and Chen (1999, theorem 1.1), the eigenvalues of $\widehat{\Sigma}$ can be computed to within an absolute error ε in $\mathcal{O}(p^3)$ arithmetic operations. Moreover, as its left-hand side is increasing in γ^*, Equation (16b) can be solved reliably via bisection or by the Newton–Raphson method. The following lemma provides a priori bounds on γ^* that can be used to initialize the bisection interval.

Lemma 3.3 (Bisection Interval)**.** *For $\rho > 0$, the unique solution of (16b) satisfies $\gamma^* \in [\gamma_{\min}, \gamma_{\max}]$, where*

$$\gamma_{\min} = \frac{p^2\lambda_{\max} + 2p\rho^2 - p\sqrt{p^2\lambda_{\max}^2 + 4p\rho^2\lambda_{\max}}}{2\rho^4} > 0,$$
$$\gamma_{\max} = \min\left\{\frac{p}{\rho^2}, \frac{1}{\rho}\sqrt{\sum_{i=1}^{p}\frac{1}{\lambda_i}}\right\}, \tag{18}$$

and $\lambda_{\max}$ denotes the maximum eigenvalue of $\widehat{\Sigma}$.

Proof. By the definitions of γ^* and x_i^* in (16), we have $\lambda_i x_i^* = (\gamma^* - x_i^*)^2/(\gamma^*)^2 < 1$, which implies that $x_i^* \le \frac{1}{\lambda_i}$. Using (16), one can further show that $(\gamma^*)^2 = \frac{1}{\rho^2}\sum_{i=1}^{p} x_i^* \le \frac{1}{\rho^2}\sum_{i=1}^{p}\frac{1}{\lambda_i}$, which is equivalent to $\gamma^* \le \frac{1}{\rho}(\sum_{i=1}^{p}\frac{1}{\lambda_i})^{\frac{1}{2}}$. Note that this upper bound on γ^* is finite only if $\lambda_i > 0$ for all $i \le p$. To derive an upper bound that is universally meaningful, we denote the left-hand side of (16b) by $\varphi(\gamma)$ and note that $\rho^2\gamma - p \le \varphi(\gamma)$ for all $\gamma \ge 0$. This estimate implies that $\gamma^* \le \frac{p}{\rho^2}$. Thus, we find $\gamma^* \le \min\{\frac{p}{\rho^2}, \frac{1}{\rho}(\sum_{i=1}^{p}\frac{1}{\lambda_i})^{\frac{1}{2}}\} = \gamma_{\max}$.

To derive a lower bound on γ^*, we set $\lambda_{\max} = \max_{i\le p}\lambda_i$ and observe that

$$\varphi(\gamma) \le \rho^2\gamma - p + \sum_{i=1}^{p}\sqrt{\lambda_i\gamma} \le \rho^2\gamma - p + p\sqrt{\lambda_{\max}\gamma},$$

where the first inequality holds because $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for all $a, b \ge 0$. As the unique positive zero of the right-hand side, $\gamma_{\min}$ provides a nontrivial lower bound on γ^*. Thus, the claim follows.

Lemma 3.3 implies that γ^* can be computed via the standard bisection algorithm to within an absolute error of ε in $\log_2((\gamma_{\max} - \gamma_{\min})/\varepsilon) = \mathcal{O}(\log_2 p)$ iterations. As evaluating the left-hand side of (16b) requires only $\mathcal{O}(p)$ arithmetic operations, the computational effort for constructing X^* is largely dominated by the cost of the spectral decomposition of the sample covariance matrix.

Remark 3.4 (Numerical Stability)**.** If both γ^* and λ_i are large numbers, then Equation (16a) for x_i^* becomes numerically unstable. A mathematically equivalent but numerically more robust reformulation of (16a) is

$$x_i^* = \gamma^*\left(1 - \frac{2}{1 + \sqrt{1 + \frac{4}{\lambda_i\gamma^*}}}\right).$$

In the following, we investigate the impact of the Wasserstein radius ρ on the optimal Lagrange multiplier γ^* and the corresponding optimal estimator X^*.

Proposition 3.5 (Sensitivity Analysis)**.** *Assume that the eigenvalues of $\widehat{\Sigma}$ are sorted in ascending order; that is, $\lambda_1 \le \cdots \le \lambda_p$. If $\gamma^*(\rho)$ denotes the solution of (16b), and $x_i^*(\rho)$, $i \le p$, represent the eigenvalues of X^* defined in (16a), which makes the dependence on $\rho > 0$ explicit, then the following assertions hold:*

i. *$\gamma^*(\rho)$ decreases with ρ, and $\lim_{\rho\to\infty}\gamma^*(\rho) = 0$;*

ii. *$x_i^*(\rho)$ decreases with ρ, and $\lim_{\rho\to\infty} x_i^*(\rho) = 0$ for all $i \le p$;*

iii. *the eigenvalues of X^* are sorted in descending order—that is, $x_1^*(\rho) \ge \cdots \ge x_p^*(\rho)$ for every $\rho > 0$; and*

iv. *the condition number $x_1^*(\rho)/x_p^*(\rho)$ of X^* decreases with ρ, and $\lim_{\rho\to\infty} x_1^*(\rho)/x_p^*(\rho) = 1$.*

Proof. As the left-hand side of (16b) is strictly increasing in ρ, it is clear that $\gamma^*(\rho)$ decreases with ρ. Moreover, the a priori bounds on $\gamma^*(\rho)$ derived in Lemma 3.3 imply that

$$0 \le \lim_{\rho\to\infty}\gamma^*(\rho) \le \lim_{\rho\to\infty}\frac{p}{\rho^2} = 0.$$

Thus, assertion (i) follows. Next, by the definition of the eigenvalue x_i^* in (16a), we have

$$\frac{\partial x_i^*}{\partial\gamma^*} = 1 + \lambda_i\gamma^* - \frac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*} + \frac{\lambda_i^2(\gamma^*)^2 + 2\lambda_i\gamma^*}{\sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*}}\right)$$
$$= 1 + \lambda_i\gamma^* - \frac{\lambda_i^2(\gamma^*)^2 + 3\lambda_i\gamma^*}{\sqrt{\lambda_i^2(\gamma^*)^2 + 4\lambda_i\gamma^*}}.$$

Elementary algebra indicates that $(1+z)\sqrt{z^2+4z} \ge z^2 + 3z$ for all $z \ge 0$, whereby the right-hand side of the above-mentioned expression is strictly positive for every $\lambda_i \ge 0$ and $\gamma^* \ge 0$. We conclude that x_i^* grows with γ^* and, by the monotonicity of $\gamma^*(\rho)$ established in assertion (i), that $x_i^*(\rho)$ decreases with ρ. As $\gamma^*(\rho)$ drops to 0 for large ρ and as the continuous function (16a) evaluates to 0 at $\gamma^* = 0$, we thus find that $x_i^*(\rho)$ converges to 0 as ρ grows. These observations establish assertion (ii). As for assertion (iii), use (16a) to express the ith eigenvalue of X^* as $x_i^* = 1 - \frac{1}{2}\psi(\lambda_i)$, where the auxiliary function $\psi(\lambda) = \sqrt{\lambda^2(\gamma^*)^2 + 4\lambda\gamma^*} - \lambda\gamma^*$

is defined for all $\lambda \geq 0$. Note that $\psi(\lambda)$ is monotonically increasing because

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi(\lambda) = \frac{\lambda(\gamma^*)^2 + 2\gamma^*}{\sqrt{\lambda^2(\gamma^*)^2 + 4\lambda\gamma^*}} - \gamma^* = \gamma^*\left(\frac{\lambda\gamma^* + 2}{\sqrt{\lambda^2(\gamma^*)^2 + 4\,\lambda\gamma^*}} - 1\right) > 0.$$

As $\lambda_{i+1} \geq \lambda_i$ for all $i < p$, we thus have $\psi(\lambda_{i+1}) \geq \psi(\lambda_i)$, which, in turn, implies that $x^*_{i+1} \leq x^*_i$. Hence, assertion (iii) follows. As for assertion (iv), note that by (16a) the condition number of X^* is given by

$$\frac{x^*_1(\rho)}{x^*_p(\rho)} = \frac{1 - \frac{1}{2}\left(\sqrt{\lambda_1^2\gamma^*(\rho)^2 + 4\lambda_1\gamma^*(\rho)} - \lambda_1\gamma^*(\rho)\right)}{1 - \frac{1}{2}\left(\sqrt{\lambda_p^2\gamma^*(\rho)^2 + 4\lambda_p\gamma^*(\rho)} - \lambda_p\gamma^*(\rho)\right)}.$$

The last expression converges to 1 as ρ tends to infinity because $\gamma^*(\rho)$ vanishes asymptotically because of assertion (i). A tedious but straightforward calculation using (16a) shows that $\frac{\partial}{\partial\gamma^*}\log(x^*_1/x^*_p) > 0$, which implies via the monotonicity of the logarithm that x^*_1/x^*_p increases with γ^*. As $\gamma^*(\rho)$ decreases with ρ by virtue of assertion (i), we may then conclude that the condition number $x^*_1(\rho)/x^*_p(\rho)$ decreases with ρ.

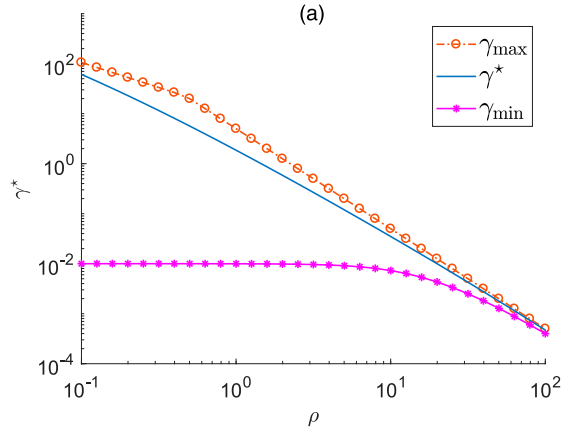
Figure 1 visualizes the dependence of γ^* and X^* on the Wasserstein radius ρ in an example where $p = 5$ and the eigenvalues of $\widehat{\Sigma}$ are given by $\lambda_i = 10^{i-3}$ for $i \leq 5$. Figure 1(a) displays γ^* as well as its a priori bounds $\gamma_{\min}$ and $\gamma_{\max}$ derived in Lemma 3.3. Note first that γ^* drops monotonically to 0 for large ρ, which is in line with Proposition 3.5(i). As γ^* represents the Lagrange multiplier of the Wasserstein constraint, which limits the size of the ambiguity set to ρ, this observation indicates that the worst-case expectation (7) displays a decreasing marginal increase in ρ. Figure 1(b) visualizes the eigenvalues x^*_i, $i \leq 5$, as well as the condition number of X^*. Note that all eigenvalues are monotonically shrunk toward 0 and that their order is preserved as ρ grows, which provides empirical support for Proposition 3.5(ii) and (iii), whereas the condition number decreases monotonically to 1, which corroborates Proposition 3.5(iv).

In summary, we have shown that X^* constitutes a nonlinear shrinkage estimator that is rotation equivariant, positive definite, and well conditioned. Moreover, $(X^*)^{-1}$ preserves the order of the eigenvalues of $\widehat{\Sigma}$. We emphasize that neither the interpretation of X^* as a shrinkage estimator nor any of its desirable properties—most notably, the improvement of its condition number with ρ—were dictated ex ante. Instead, these properties arose naturally from an intuitively appealing distributionally robust estimation scheme. By contrast, existing estimation schemes sometimes impose ad hoc constraints on condition numbers; see, for example, Won et al. (2013). On the downside, as X^* shares the same eigenbasis as the sample covariance matrix $\widehat{\Sigma}$, it does not prompt a new robust principal component analysis. We henceforth refer to X^* as the *Wasserstein shrinkage estimator*.

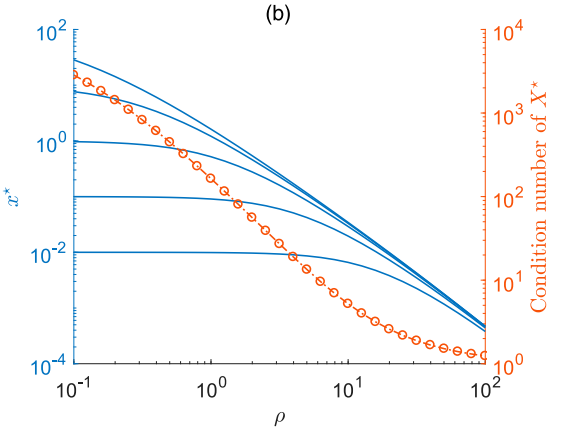
4. Numerical Solution with Sparsity Information

We now investigate a more general setting where $\mathcal{X}$ may be a strict subset of $\mathbb{S}^p_{++}$, which captures a prescribed conditional independence structure of ξ. Specifically, we assume that there exists $\mathcal{E} \subseteq \{1,\ldots,p\}^2$ such that the random variables ξ_i and ξ_j are conditionally independent given $\xi_{-\{i,j\}}$ for any pair $(i,j) \in \mathcal{E}$,

Figure 1. (Color online) Dependence of the Lagrange Multiplier γ^* (Left Panel) as Well as the Eigenvalues x^*_i, $i \leq 5$, and the Condition Number x^*_5/x^*_1 of the Optimal Estimator X^* (Right Panel) on ρ



Lagrange multiplier $\gamma^\star$ and its a priori bounds $\gamma_{\min}$ and $\gamma_{\max}$ from Lemma 3.3.



Eigenvalues (left axis) and condition number (round marker - right axis) of $X^\star$.

where $\xi_{-\{i,j\}}$ represents the truncation of the random vector ξ without the components ξ_i and ξ_j. It is well known that if ξ follows a normal distribution with covariance matrix $S \succ 0$ and precision matrix $X = S^{-1}$, then ξ_i and ξ_j are conditionally independent given $\xi_{-(i,j)}$ if and only if $X_{ij} = 0$. This reasoning forms the basis of the celebrated Gaussian graphical models (see, e.g., Lauritzen (1996)). Any prescribed conditional independence structure of ξ can thus conveniently be captured by the feasible set

$$\mathcal{X} = \{X \in \mathbb{S}_{++}^p : X_{ij} = 0 \quad \forall (i,j) \in \mathcal{E}\}.$$

We may assume without loss of generality that $\mathcal{E}$ inherits symmetry from X; that is, $(i,j) \in \mathcal{E} \Rightarrow (j,i) \in \mathcal{E}$. In Section 3 we have seen that the robust maximum likelihood estimation problem (5) admits an analytical solution when $\mathcal{E} = \emptyset$. In the general case, analytical tractability is lost. Indeed, if $\mathcal{E} \neq \emptyset$, then even the nominal estimation problem obtained by setting $\rho = 0$ requires a numerical solution (Dahl et al. 2005). In this section we develop a Newton-type algorithm to solve (5) in the presence of prior conditional independence information. For the sake of consistency, we will refer to the optimal solution of Problem (5) as the *Wasserstein shrinkage estimator* even in the presence of sparsity constraints.

Remark 4.1 (Conditional Independence Information in $\mathcal{P}_\rho$)**.** We emphasize that our proposed estimation model accounts for the prescribed conditional independence structure only in the feasible set $\mathcal{X}$ but not in the ambiguity set $\mathcal{P}_\rho$. Otherwise, the ambiguity set would have to be redefined as

$$\mathcal{P}_\rho = \left\{\mathbb{Q} \in \mathcal{N}_0^p \;:\; \mathbb{W}\left(\mathbb{Q}, \widehat{\mathbb{P}}\right) \leq \rho, \; \left(\mathbb{E}^{\mathbb{Q}}\left[\xi\xi^\top\right]^{-1}\right)_{ij} = 0 \;\; \forall (i,j) \in \mathcal{E}\right\}.$$

Although conceptually attractive, this new ambiguity set is empty even for some $\rho > 0$ because the inverse sample covariance matrix $\widehat{\Sigma}^{-1}$ violates the prescribed conditional independence relationships with probability 1.

Recall from Theorem 2.6 that the estimation problem (5) is equivalent to the convex program (6) and that the optimal value of (6) depends continuously on $\widehat{\Sigma} \in \mathbb{S}_+^p$. In the remainder of this section we may thus assume without much loss of generality that $\widehat{\Sigma} \succ 0$. Otherwise, we can replace $\widehat{\Sigma}$ with $\widehat{\Sigma} + \varepsilon I$ for some small $\varepsilon > 0$ without significantly changing the estimation problem's solution. Inspired by Oztoprak et al. (2012) and Hsieh et al. (2014), we now develop a sequential quadratic approximation algorithm for solving Problem (6) with sparsity information. Note that the set $\mathcal{X}$ of feasible precision matrices typically fixes many entries to zero, thus reducing the effective problem dimension and making a second-order algorithm attractive even for large instances of (6).

The proposed algorithm starts at $X_0 = I$ and at some $\gamma_0 > 1$, which are trivially feasible in (6). In each iteration the algorithm moves from the current iterate (X_t, γ_t) along a feasible descent direction, which is constructed from a quadratic approximation of the objective function of Problem (6). A judiciously chosen step size guarantees that the next iterate (X_{t+1}, γ_{t+1}) remains feasible and has a better (lower) objective value; see Algorithm 1. The construction of the descent direction relies on the following lemma.

Lemma 4.2 (Fact 7.4.9 in Bernstein 2009)**.** *For any $A, B \in \mathbb{R}^{p\times p}$ and $X \in \mathbb{S}^p$, we have*

$$\mathrm{Tr}[AXBX] = \mathrm{vec}(X)^\top (B \otimes A^\top)\mathrm{vec}(X).$$

Proposition 4.3 (Descent Direction)**.** *Fix $(X,\gamma) \in \mathbb{S}_{++}^p \times \mathbb{R}_{++}$ with $\gamma I \succ X$, and define the orthogonal projection $P : \mathbb{R}^{p^2+1} \to \mathbb{R}^{p^2+1}$ through $(Pz)_k = 0$ if $k = p(j-1)+i$ for some $(i,j) \in \mathcal{E}$, $k = \frac{1}{2} z_{p(j-1)+i} + \frac{1}{2} z_{p(i-1)+j}$ if $k = p(j-1)+i$ for some $i,j \leq p$ with $(i,j) \notin \mathcal{E}$, and $k = z_k$ if $k = p^2+1$. Moreover, define $G := I - \frac{X}{\gamma}$,*

$$H := \begin{bmatrix} X^{-1}\otimes X^{-1} + \frac{2}{\gamma} G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1} & -\frac{1}{\gamma^2}\mathrm{vec}\left(G^{-1}\left[XG^{-1}\widehat{\Sigma} + \widehat{\Sigma}G^{-1}X\right]G^{-1}\right) \\ -\frac{1}{\gamma^2}\mathrm{vec}\left(G^{-1}\left[XG^{-1}\widehat{\Sigma} + \widehat{\Sigma}G^{-1}X\right]G^{-1}\right)^\top & \frac{2}{\gamma^3}\mathrm{Tr}\left[G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}X\right] \end{bmatrix} \in \mathbb{S}^{p^2+1}$$

and

$$g := \begin{bmatrix} \mathrm{vec}\left(G^{-1}\widehat{\Sigma}G^{-1} - X^{-1}\right) \\ \rho^2 + \mathrm{Tr}\left[G^{-1}\widehat{\Sigma}\left(I - \frac{1}{\gamma}G^{-1}X\right) - \widehat{\Sigma}\right] \end{bmatrix} \in \mathbb{R}^{p^2+1}.$$

Then, the unique solution $(\Delta_X^*, \Delta_\gamma^*) \in \mathbb{S}^p \times \mathbb{R}$ of the linear system

$$PH\left(\left(\mathrm{vec}(\Delta_X^*)^\top, \Delta_\gamma^*\right)^\top + g\right) = 0 \;\; \text{and} \;\; (\Delta_X^*)_{ij} = 0 \;\; \forall (i,j) \in \mathcal{E} \tag{19}$$

represents a feasible descent direction for the optimization problem (6) at (X,γ).

Proof. We first expand the objective function of Problem (6) around $(X,\gamma) \in \mathbb{S}_{++}^p \times \mathbb{R}_{++}$ with $\gamma I \succ X$. By the

rules of matrix calculus, the second-order Taylor expansion of the negative log determinant is given by

$$-\log\det(X+\Delta_X) = -\log\det(X) - \mathrm{Tr}\big[X^{-1}\Delta_X\big] + \frac{1}{2}\mathrm{Tr}\big[X^{-1}\Delta_X X^{-1}\Delta_X\big] + \mathcal{O}\big(\|\Delta_X\|^3\big)$$

for $\Delta_X \in \mathbb{S}^p$; see also Boyd and Vandenberghe (2004, p. 644). Moreover, by using a geometric series expansion, we obtain

$$\begin{aligned}\left(I - \frac{X+\Delta_X}{\gamma+\Delta_\gamma}\right)^{-1} &= \left(I - \frac{X+\Delta_X}{\gamma}\left(1 - \frac{\Delta_\gamma}{\gamma} + \frac{\Delta_\gamma^2}{\gamma^2} + \mathcal{O}\big(\|\Delta_\gamma\|^3\big)\right)\right)^{-1} \\ &= \left(I - \frac{X}{\gamma} + \frac{X\Delta_\gamma}{\gamma^2} - \frac{X\Delta_\gamma^2}{\gamma^3} - \frac{\Delta_X}{\gamma} + \frac{\Delta_X\Delta_\gamma}{\gamma^2} + \mathcal{O}\big(\|(\Delta_X,\Delta_\gamma)\|^3\big)\right)^{-1}\end{aligned}$$

for $|\Delta_X| \in \mathbb{R}$. Expanding the matrix inverse as a Neumann series and setting $G = I - \frac{X}{\gamma}$, which is invertible because $\gamma I \succ X$, this expression can be reformulated as

$$\begin{aligned}&G^{-\frac{1}{2}}\left(I + \frac{G^{-\frac{1}{2}}XG^{-\frac{1}{2}}\Delta_\gamma}{\gamma^2} - \frac{G^{-\frac{1}{2}}XG^{-\frac{1}{2}}\Delta_\gamma^2}{\gamma^3} - \frac{G^{-\frac{1}{2}}\Delta_X G^{-\frac{1}{2}}}{\gamma} + \frac{G^{-\frac{1}{2}}\Delta_X G^{-\frac{1}{2}}\Delta_\gamma}{\gamma^2} + \mathcal{O}\big(\|(\Delta_X,\Delta_\gamma)\|^3\big)\right)^{-1} G^{-\frac{1}{2}} \\ &= G^{-1} - \frac{G^{-1}XG^{-1}\Delta_\gamma}{\gamma^2} + \frac{G^{-1}XG^{-1}\Delta_\gamma^2}{\gamma^3} + \frac{G^{-1}\Delta_X G^{-1}}{\gamma} - \frac{G^{-1}\Delta_X G^{-1}\Delta_X G^{-1}\Delta_\gamma}{\gamma^2} \\ &\quad + \frac{G^{-1}XG^{-1}XG^{-1}\Delta_\gamma^2}{\gamma^4} + \frac{G^{-1}\Delta_X G^{-1}\Delta_X G^{-1}}{\gamma^2} - \frac{G^{-1}XG^{-1}\Delta_X G^{-1}\Delta_\gamma}{\gamma^3} - \frac{G^{-1}\Delta_X G^{-1}XG^{-1}\Delta_\gamma}{\gamma^3} \\ &\quad + \mathcal{O}\big(\|(\Delta_X,\Delta_\gamma)\|^3\big).\end{aligned}$$

Thus, the second-order Taylor expansion of the last term in the objective function of (6) is given by

$$\begin{aligned}&(\gamma+\Delta_\gamma)^2\mathrm{Tr}\left[\Big((\gamma+\Delta_\gamma)I - (X+\Delta_X)\Big)^{-1}\widehat{\Sigma}\right] \\ &= (\gamma+\Delta_\gamma)\mathrm{Tr}\left[\left(I - \frac{X+\Delta_X}{\gamma+\Delta_\gamma}\right)^{-1}\widehat{\Sigma}\right] \\ &= \gamma\mathrm{Tr}\Big[G^{-1}\widehat{\Sigma}\Big] + \Delta_\gamma\mathrm{Tr}\left[G^{-1}\widehat{\Sigma}\left(I - \frac{1}{\gamma}G^{-1}X\right)\right] \\ &\quad + \frac{\Delta_\gamma^2}{\gamma^3}\mathrm{Tr}\Big[G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}X\Big] + \mathrm{Tr}\Big[G^{-1}\,\widehat{\Sigma}G^{-1}\Delta_X\Big] \\ &\quad - \frac{\Delta_\gamma}{\gamma^2}\mathrm{Tr}\Big[G^{-1}\widehat{\Sigma}G^{-1}\Delta_X G^{-1}X + G^{-1}\widehat{\Sigma}G^{-1}XG^{-1}\Delta_X\Big] \\ &\quad + \frac{1}{\gamma}\mathrm{Tr}\Big[G^{-1}\Delta_X G^{-1}\widehat{\Sigma}G^{-1}\Delta_X\Big] + \mathcal{O}\Big(\|(\Delta_X,\Delta_\gamma)\|^3\Big),\end{aligned}$$

where the second equality follows from the Taylor expansion of the matrix inverse derived above. Using Lemma 4.2, the objective function of (6) is thus representable as

$$\begin{aligned}&-\log\det(X+\Delta_X) + (\gamma+\Delta_\gamma)\Big(\rho^2 - \mathrm{Tr}\Big[\widehat{\Sigma}\Big]\Big) \\ &\quad + (\gamma+\Delta_\gamma)^2\mathrm{Tr}\Big[\Big((\gamma+\Delta_\gamma)I - (X+\Delta_X)\Big)^{-1}\widehat{\Sigma}\Big] \\ &= c + g^\top(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)^\top \\ &\quad + \frac{1}{2}(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)H(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)^\top + \mathcal{O}\big(\|(\Delta_X,\Delta_\gamma)\|^3\big)\end{aligned}$$

for some $c \in \mathbb{R}$, where the gradient $g \in \mathbb{R}^p$ and the Hessian $H \in \mathbb{S}^p$ are defined as in the proposition statement. A feasible descent direction for Problem (6) is thus obtained by solving the auxiliary quadratic program:

$$\begin{aligned}\min_{\Delta_X,\Delta_\gamma}\ & g^\top(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)^\top + \frac{1}{2}(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma) \\ &\quad \times H(\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)^\top \\ \text{s.t.}\ & \Delta_X \in \mathbb{S}^p,\ (\Delta_X)_{ij} = 0 \quad \forall (i,j) \in \mathcal{E}.\end{aligned} \tag{20}$$

Note that (20) has a unique minimizer because H is positive definite. Indeed, we have

$$\begin{aligned}&\frac{4}{\gamma^4}\mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big)^\top \\ &\quad \times \left(X^{-1}\otimes X^{-1} + \frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}\right)^{-1} \\ &\quad \times \mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big) \\ &< \frac{4}{\gamma^4}\mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big)^\top \left(\frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}\right)^{-1} \\ &\quad \times \mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big) \\ &= \frac{2}{\gamma^3}\mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big)^\top \Big(G\widehat{\Sigma}^{-1}G\otimes G\Big) \\ &\quad \times \mathrm{vec}\Big(G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}\Big) \\ &= \frac{2}{\gamma^3}\mathrm{Tr}\Big[G^{-1}XG^{-1}\widehat{\Sigma}G^{-1}X\Big],\end{aligned}$$

where the inequality holds because $X\otimes X$ is positive definite and $G^{-1}XG^{-1}\widehat{\Sigma}G^{-1} \neq 0$, the first equality follows from Bernstein (2009, proposition 7.1.7), which asserts that $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$ for any $A, B \in \mathbb{S}^p_{++}$, and the second equality follows from Lemma 4.2. The above-mentioned derivation shows that the Schur complement of the positive definite block $X^{-1}\otimes X^{-1} + \frac{2}{\gamma}G^{-1}\widehat{\Sigma}G^{-1}\otimes G^{-1}$ in H is a positive number, which, in turn, implies that the Hessian H is positive definite. In the following, we denote the unique minimizer of (20) by $(\Delta_X^*, \Delta_\gamma^*)$. As $\Delta_X = 0$ and $\Delta_\gamma = 0$ is feasible in (20), it is clear that the objective value of $(\Delta_X^*, \Delta_\gamma^*)$ is

nonpositive. In fact, as $H \succ 0$, the minimum of (20) is negative unless $g = 0$. Thus, $(\Delta_X^*, \Delta_\gamma^*)$ is a feasible descent direction.

Note that P defined in the proposition statement represents the orthogonal projection on the linear space:

$$\mathcal{Z} = \Big\{ z = (\mathrm{vec}(\Delta_X)^\top, \Delta_\gamma)^\top \in \mathbb{R}^{p^2+1} : \Delta_X \in \mathbb{S}^p, \quad (\Delta_X)_{ij} = 0 \quad \forall (i,j) \in \mathcal{E} \Big\}.$$

Indeed, it is easy to verify that $P^2 = P = P^\top$ because the range and the null space of P correspond to $\mathcal{Z}$ and its orthogonal complement, respectively. The quadratic program (20) is thus equivalent to

$$\min_{z \in \mathcal{Z}} \left\{ g^\top z + \frac{1}{2} z^\top H z \right\} = \min_{z \in \mathbb{R}^{p^2+1}} \left\{ g^\top z + \frac{1}{2} z^\top H z : Pz = z \right\}.$$

The minimizer z^* of the last reformulation and the optimal Lagrange multiplier μ^* associated with its equality constraint correspond to the unique solution of the Karush-Kuhn-Tucker optimality conditions:

$$Hz^* + g + (I - P)\mu^* = 0,\ (1 - P)z^* = 0 \iff P(Hz^* + g) = 0,\ (1 - P)z^* = 0,$$

which are mainfestly equivalent to (19). Thus, the claim follows.

Given a descent direction $(\Delta_X^*, \Delta_\gamma^*)$ at a feasible point (X, γ), we use a variant of Armijo's rule (Nocedal and Wright 2006, section 3.1) to choose a step size $\alpha > 0$ that preserves feasibility of the next iterate $(X + \alpha\Delta_X^*, \gamma + \alpha\Delta_\gamma^*)$ and ensures a sufficient decrease of the objective function. Specifically, for a prescribed line search parameter $\sigma \in (0, \frac{1}{2})$, we set the step size α to the largest number in $\{\frac{1}{2^m}\}_{m \in \mathbb{Z}_+}$ satisfying the following two conditions.

Condition 1 (Feasibility)**.** Note that $(\gamma + \alpha\Delta_\gamma^*)I \succ X + \alpha\Delta_X^* \succ 0$.

Condition 2 (Sufficient Decrease)**.** Note that $f(X + \alpha\Delta_X^*, \gamma + \alpha\Delta_\gamma^*) \le f(X, \gamma) + \sigma\alpha\delta$, where $\delta = g^\top (\mathrm{vec}(\Delta_X^*)^\top, \Delta_\gamma^*)^\top < 0$, and g is defined as in Propostion 4.3.

Notice that the sparsity constraints are automatically satisfied at the next iterate thanks to the construction of the descent direction $(\Delta_X^*, \Delta_\gamma^*)$ in (19). Algorithm 1 repeats the procedure previously outlined until $\|g\|$ drops below a given tolerance (10^{-3}) or until the iteration count exceeds a given threshold (10^2). Throughout the numerical experiments in Section 6, we set $\sigma = 10^{-4}$, which is the value recommended in Nocedal and Wright (2006).

Algorithm 1 (Sequential Quadratic Approximation Algorithm)

Data: Sample covariance matrix $\widehat{\Sigma} \succ 0$, Wasserstein radius $\rho > 0$, line search parameter $\sigma \in (0, \frac{1}{2})$.

Initialize $X_0 = I$ and $\gamma_0 > 1$, and set $t \leftarrow 0$;

while *stopping criterion is violated* **do**

- Find the descent direction $(\Delta_X^*, \Delta_\gamma^*)$ at $(X, \gamma) = (X_t, \gamma_t)$ by solving (19);
- Find the largest step size $\alpha_t \in \{\frac{1}{2^m}\}_{m \in \mathbb{Z}_+}$ satisfying Conditions 1 and 2;
- Set $X_{t+1} = X_t + \alpha_t \Delta_X^*$, $\gamma_{t+1} = \gamma_t + \alpha_t \Delta_\gamma^*$;
- Set $t \leftarrow t + 1$;

Remark 4.4 (Steepest Descent Algorithm)**.** The computation of the descent direction in Proposition 4.3 requires second-order information. It is easy to verify that Proposition 4.3 remains valid if the Hessian H is replaced with the identity matrix, in which case the sequential quadratic approximation algorithm reduces to the classical steepest descent algorithm (Nocedal and Wright 2006, chap. 3).

The next proposition establishes that Algorithm 1 converges to the unique minimizer of Problem (6).

Proposition 4.5 (Convergence)**.** *Assume that $\widehat{\Sigma} \succ 0$, $\rho > 0$, and $\sigma \in (0, \frac{1}{2})$. For any initial feasible solution (X_0, γ_0), the sequence $\{(X_t, \gamma_t)\}_{t \in \mathbb{Z}_+}$ generated by Algorithm 1 converges to the unique minimizer (X^*, γ^*) of Problem (6). Moreover, the sequence converges locally quadratically.*

Proof. Denote by $f(X, \gamma)$ the objective function of Problem (6), and define

$$\mathcal{C} := \{(X, \gamma) \in \mathcal{X} \times \mathbb{R}_+ : f(X, \gamma) \le f(X_0, \gamma_0), 0 \prec X \prec \gamma I\}$$

as the set of all feasible solutions that are at least as good as the initial solution (X_0, γ_0). The proof of Theorem 2.6 implies that $\underline{x} I \preceq X \preceq \overline{x} I$ and $\underline{x} \le \gamma \le \overline{x}$ for all $(X, \gamma) \in \mathcal{C}$, where the strictly positive constants $\underline{x}$ and $\overline{x}$ are defined as in (13). Note that, as $\widehat{\Sigma}$ is fixed in this proof, the dependence of $\underline{x}$ and $\overline{x}$ on $\widehat{\Sigma}$ is notationally suppressed to avoid clutter. Thus, $\mathcal{C}$ is bounded. Moreover, as $\widehat{\Sigma} \succ 0$, it is easy to verify $f(X, \gamma)$ tends to infinity if the smallest eigenvalue of X approaches 0 or if the largest eigenvalue of X approaches γ. The continuity of $f(X, \gamma)$ then implies that $\mathcal{C}$ is closed. In summary, we conclude that $\mathcal{C}$ is compact.

By the definition of $f(X, \gamma)$ in (6), any $(X, \gamma) \in \mathcal{C}$ satisfies

$$\begin{aligned} 0 &\le f(X_0, \gamma_0) + \log\det(X) - \gamma\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) \\ &\quad - \gamma\left\langle (I - \gamma^{-1}X)^{-1}, \widehat{\Sigma} \right\rangle \\ &\le f(X_0, \gamma_0) + p\log(\overline{x}) + \overline{x}\,\mathrm{Tr}\left[\widehat{\Sigma}\right] \\ &\quad - \underline{x}\,\lambda_{\min}\mathrm{Tr}\left[(I - \gamma^{-1}X)^{-1}\right], \end{aligned}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $\widehat{\Sigma}$, which is positive by assumption. Thus, we have

$$\mathrm{Tr}\left[\left(I-\gamma^{-1}X\right)^{-1}\right] \leq \frac{1}{\underline{x}\,\lambda_{\min}}\left(f(X_0,\gamma_0)+p\log(\overline{x})+\overline{x}\mathrm{Tr}\left[\widehat{\Sigma}\right]\right),$$

which implies that the eigenvalues of $I-\frac{X}{\gamma}$ are uniformly bounded away from 0 on $\mathcal{C}$. More formally, there exists $c_0>0$ with $I-\frac{X}{\gamma}\succ c_0 I$ for all $(X,\gamma)\in\mathcal{C}$. As the objective function $f(X,\gamma)$ is smooth wherever it is defined, its gradient and Hessian constitute continuous functions on $\mathcal{C}$. Moreover, as $f(X,\gamma)$ is strictly convex on the compact set $\mathcal{C}$, the eigenvalues of its Hessian matrix are uniformly bounded away from 0. This implies that the inverse Hessian matrix and the descent direction $(\Delta_X^*,\Delta_\gamma^*)$ constructed in Proposition 4.3 are also continuous on $\mathcal{C}$. Hence, there exist $c_1,c_2>0$ such that $\Delta_X^*\preceq c_1 I$ and $|\Delta_\gamma^*|\leq c_2$ uniformly on $\mathcal{C}$.

We conclude that any positive step size $\alpha<\underline{x}\min\{c_1^{-1},(c_1+c_2)^{-1}c_0\}$ satisfies the feasibility condition (Condition 1) uniformly on $\mathcal{C}$ because $X+\alpha\Delta_X^*\succ(\underline{x}-\alpha c_1)I\succeq 0$ and

$$\begin{aligned}(\gamma+\alpha\Delta_\gamma^*)I &\succeq X+c_0\underline{x}I+\alpha(\Delta_X^*-\Delta_X^*+\Delta_\gamma^* I).\\ &\succeq X+c_0\underline{x}I+\alpha(\Delta_X^*-(c_1+c_2)I)\succ X+\alpha\Delta_X^*\end{aligned}$$

for all $(X,\gamma)\in\mathcal{C}$. Moreover, by Tseng and Yun (2009, lemma 5(b)), there exists $\overline{\alpha}>0$ such that any positive step size $\alpha\leq\overline{\alpha}$ satisfies the descent condition (Condition 2) for all $(X,\gamma)\in\mathcal{C}$. In summary, there exists $m^*\in\mathbb{Z}_+$ such that

$$\alpha^*=\frac{1}{2^{m^*}}<\min\left\{\overline{\alpha},\underline{x}\min\left\{c_1^{-1},(c_1+c_2)^{-1}c_0\right\}\right\}$$

satisfies both line search conditions (Conditions 1 and 2) uniformly on $\mathcal{C}$. By induction, the iterates $\{(X_t,\gamma_t)\}_{t\in\mathbb{N}}$ generated by Algorithm 1 have nonincreasing objective values and thus all belong to $\mathcal{C}$, whereas the step sizes $\{\alpha_t\}_{t\in\mathbb{N}}$ generated by Algorithm 1 are all larger or equal to α^*. Hence, the algorithm's global convergence is guaranteed by Tseng and Yun (2009, theorem 1), whereas the local quadratic convergence follows from Hsieh et al. (2014, theorem 16).

Remark 4.6 (Refinements of Algorithm 1)**.** For large values of p, computing and storing the exact Hessian matrix H from Proposition 4.3 is prohibitive. In this case, H can be approximated by a low-rank matrix as in the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) method without sacrificing global convergence (Tseng and Yun 2009). Alternatively, one can resort to a coordinate descent method akin to the QUIC algorithm (Hsieh et al. 2014), in which case both the global and local convergence guarantees of Proposition 4.5 remain valid.

Remark 4.7 (Learning the Sparsity Pattern)**.** If the precision matrix is known to be sparse but has an unknown sparsity pattern, then one may set $\mathcal{X}=\mathbb{S}_{++}^p$ and add a weighted ℓ_1-regularization or Lasso term to the objective function of Problem (6) in order to generate sparse Wasserstein shrinkage estimators. Different sparsity patterns can be obtained by tuning the weight of the Lasso term. The regularized nonlinear SDP (6) can then be solved with a variant of the QUIC algorithm (Hsieh et al. 2014). Indeed, the gradient and the Hessian matrix of the smooth part of the objective function (which coincides with the robust version of Stein's loss function) can again be computed efficiently by leveraging Proposition 4.3. Details are omitted for brevity.

5. Extremal Distributions

It is instructive to characterize the extremal distributions that attain the supremum in (7) for a given sample covariance matrix $\widehat{\Sigma}$ and a fixed candidate estimator X.

Theorem 5.1 (Extremal Distributions)**.** *For any $\widehat{\Sigma},X\in\mathbb{S}_{++}^p$ and $\rho>0$, the supremum in (7) is attained by the normal distribution $\mathbb{Q}^*=\mathcal{N}(0,S^*)$ with covariance matrix*

$$S^*=(\gamma^*)^2(\gamma^*I-X)^{-1}\widehat{\Sigma}(\gamma^*I-X)^{-1},$$

where γ^* is the unique solution with $\gamma^*I\succ X$ of the following algebraic equation:

$$\begin{aligned}&\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]+2\gamma^*\mathrm{Tr}\left[(\gamma^*I-X)^{-1}\widehat{\Sigma}\right]\\ &\quad-(\gamma^*)^2\mathrm{Tr}\left[(\gamma^*I-X)^{-1}\widehat{\Sigma}(\gamma^*I-X)^{-1}\right]=0.\end{aligned}\tag{21}$$

Proof. From Proposition 2.8 we know that the worst-case expectation problem (7) is equivalent to the semidefinite program (8). Note that the strictly convex objective function of (8) is bounded below by

$$\gamma\left(\rho^2-\mathrm{Tr}\left[\widehat{\Sigma}\right]\right)+\lambda_{\min}\gamma^2\mathrm{Tr}\left[(\gamma I-X)^{-1}\right],$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $\widehat{\Sigma}$. As $\lambda_{\min}$ is positive by assumption, the objective function of (8) tends to infinity as γ approaches the largest eigenvalue of X, in which case $\gamma I-X$ becomes singular. Thus, the unique optimal solution γ^* of (8) satisfies $\gamma^*I\succ X$ and solves the first-order optimality condition (21).

Now we are ready to prove that $\mathbb{Q}^*$ is both feasible and optimal in (7). By the formula for S^* in terms of γ^*, $\widehat{\Sigma}$, and S, and by using Definition 2.3 and Proposition 2.2, it is easy to verify that (21) is equivalent to

$$\mathrm{Tr}[S^*] + \mathrm{Tr}\left[\widehat{\Sigma}\right] - 2\mathrm{Tr}\left[\sqrt{\widehat{\Sigma}^{\frac{1}{2}} S^* \widehat{\Sigma}^{\frac{1}{2}}}\right] = \rho^2 \quad \Longleftrightarrow$$
$$\mathbb{W}_S\left(S^*, \widehat{\Sigma}\right) = \mathbb{W}\left(\mathbb{Q}^*, \widehat{\mathbb{P}}\right) = \rho,$$

which confirms that $\mathbb{Q}^*$ is feasible in (7). Moreover, the objective value of $\mathbb{Q}^*$ in (7) amounts to

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}^*}\left[\langle \xi\xi^\top, X\rangle\right] \\
&= \langle S^*, X\rangle = (\gamma^*)^2\left\langle (\gamma^* I - X)^{-1} \right. \\
&\quad \left. \widehat{\Sigma}(\gamma^* I - X)^{-1}, X\right\rangle \\
&= (\gamma^*)^2\left\langle (\gamma^* I - X)^{-1}\widehat{\Sigma}(\gamma^* I \right. \\
&\quad \left. - X)^{-1}, (X - \gamma^* I) + \gamma^* I\right\rangle \\
&= -(\gamma^*)^2\mathrm{Tr}\left[(\gamma^* I - X)^{-1}\widehat{\Sigma}\right] \\
&\quad + (\gamma^*)^3\mathrm{Tr}\left[(\gamma^* I - X)^{-1}\widehat{\Sigma}(\gamma^* I - X)^{-1}\right] \\
&= \gamma^*\left(\rho^2 - \mathrm{Tr}\left[\widehat{\Sigma}\right]\right) + (\gamma^*)^2 \\
&\quad \times \left\langle (\gamma^* I - X)^{-1}, \widehat{\Sigma}\right\rangle = g\left(\widehat{\Sigma}, X\right),
\end{aligned}$$

where the penultimate equality exploits (21), and the last equality follows from the optimality of γ^* in (8) and from Proposition 2.8. Thus, $\mathbb{Q}^*$ is optimal in (7).

In the absence of sparsity information (i.e., if $\mathcal{X} = \mathbb{S}_{++}^p$), the unique minimizer X^* of Problem (5) is available in closed form thanks to Theorem 3.1. In this case, the extremal distribution attaining the supremum in (7) at $X = X^*$ can also be computed in closed form even if $\widehat{\Sigma}$ is rank deficient.

Corollary 5.2 (Extremal Distribution for Optimal Estimator)**.** *Assume that $\rho > 0$, $\mathcal{X} = \mathbb{S}_{++}^p$, and $\widehat{\Sigma} \in \mathbb{S}_+^p$ admits the spectral decomposition $\widehat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^\top$ with eigenvalues λ_i and corresponding orthonormal eigenvectors v_i, $i \le p$. If (X^*, γ^*) represents the unique solution of (6) given in* Theorem 3.1*, then the supremum in (7) at $X = X^*$ is attained by the normal distribution $\mathbb{Q}^* = \mathcal{N}(0, S^*)$ with covariance matrix*

$$S^* = \sum_{i=1}^p s_i^* v_i v_i^\top,$$
$$\text{where} \quad s_i^* = \begin{cases} (\gamma^*)^2\lambda_i(\gamma^* - x_i^*)^{-2} & \text{if } \lambda_i > 0, \\ (\gamma^*)^{-1} & \text{if } \lambda_i = 0. \end{cases}$$

Proof. If $\widehat{\Sigma} > 0$, the claim follows immediately by substituting the formula for X^* from Theorem 3.1 into the formula for S^* from Theorem 5.1. If $\widehat{\Sigma} \ge 0$ is rank deficient, we consider the invertible sample covariance matrix $\widehat{\Sigma} + \varepsilon I > 0$ for some $\varepsilon > 0$, we denote by $(X^*(\varepsilon), \gamma^*(\varepsilon))$ the corresponding minimizer of Problem (6) as constructed in (16), and we let $S^*(\varepsilon)$ be the covariance matrix of the extremal distribution of Problem (7) at $X = X^*(\varepsilon)$. Using the same reasoning as in the proof of Theorem 3.1, one can show that $(X^*(\varepsilon), \gamma^*(\varepsilon))$ is continuous in $\varepsilon \in \mathbb{R}_+$ and converges to (X^*, γ^*) as ε tends to 0. Similarly, $S^*(\varepsilon)$ is continuous in $\varepsilon \in \mathbb{R}_+$ and converges to S^* as ε tends to 0. To see this, note that the eigenvalues $s_i^*(\varepsilon)$, $i \le p$, of $S^*(\varepsilon)$ satisfy

$$\begin{aligned}
&\lim_{\varepsilon\to 0^+} s_i^*(\varepsilon) \\
&= \lim_{\varepsilon\to 0^+} \frac{\gamma^*(\varepsilon)^2(\lambda_i + \varepsilon)}{\left(\gamma^*(\varepsilon) - x_i^*(\varepsilon)\right)^2} \\
&= \lim_{\varepsilon\to 0^+} \frac{4(\lambda_i + \varepsilon)}{\left(\sqrt{(\lambda_i+\varepsilon)^2\gamma^*(\varepsilon)^2 + 4(\lambda_i+\varepsilon)\gamma^*(\varepsilon)} - (\lambda_i + \varepsilon)\gamma^*(\varepsilon)\right)^2} \\
&= s_i^* \quad \forall i \le p,
\end{aligned}$$

where the first equality follows from the first part of the proof, the second equality exploits (16a), and the third equality holds as a result of the definition of s_i^*.

We are now armed to prove that $\mathbb{Q}^*$ is both feasible and optimal in (7). Indeed, using the continuity of $S^*(\varepsilon)$ and $\mathbb{W}_S(S_1, S_2)$ in their respective arguments, we find

$$\begin{aligned}
\mathbb{W}\left(\mathbb{Q}^*, \widehat{\mathbb{P}}\right) &= \mathbb{W}_S\left(S^*, \widehat{\Sigma}\right) \\
&= \lim_{\varepsilon\to 0^+} \mathbb{W}_S\left(S^*(\varepsilon), \widehat{\Sigma} + \varepsilon I\right) = \rho,
\end{aligned}$$

where the last equality follows from the construction of $S^*(\varepsilon)$ in the proof of Theorem 5.1. Thus, $\mathbb{Q}^*$ is feasible in (7). Similarly, using the continuity of $S^*(\varepsilon)$ and $X^*(\varepsilon)$ in ε, we have

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}^*}\left[\langle \xi\xi^\top, X^*\rangle\right] &= \langle S^*, X^*\rangle = \lim_{\varepsilon\to 0^+} \langle S^*(\varepsilon), X^*(\varepsilon)\rangle \\
&= \lim_{\varepsilon\to 0^+} g\left(\widehat{\Sigma} + \varepsilon I, S^*(\varepsilon)\right) = g\left(\widehat{\Sigma}, S^*\right),
\end{aligned}$$

where the second-to-last and last equalities follow from the construction of $S^*(\varepsilon)$ in the proof of Theorem 5.1 and the continuity of $g(\widehat{\Sigma}, X)$ established in Proposition 2.8, respectively. Thus, $\mathbb{Q}^*$ is optimal in (7).

6. Numerical Experiments

To assess the statistical and computational properties of the proposed Wasserstein shrinkage estimator, we compare it against two state-of-the-art precision matrix estimators from the literature.

Definition 6.1 (Linear Shrinkage Estimator)**.** Denote by $\mathrm{diag}(\widehat{\Sigma})$ the diagonal matrix of all sample variances. Then, the linear shrinkage estimator with mixing parameter $\alpha \in [0,1]$ is defined as

$$X^* = \left[(1-\alpha)\widehat{\Sigma} + \alpha\,\mathrm{diag}\left(\widehat{\Sigma}\right)\right]^{-1}.$$

The linear shrinkage estimator uses the diagonal matrix of sample variances as the shrinkage target. Thus, the sample covariances are shrunk to zero, whereas the sample variances are preserved. We emphasize that the most prevalent shrinkage target is a scaled identity matrix (Ledoit 2004a). The benefits of using $\mathrm{diag}(\widehat{\Sigma})$ instead are discussed in Schäfer and Strimmer (2005, section 2.4). This particular linear shrinkage estimator can also be interpreted as the maximum a posteriori estimator based on an inverse Wishart prior (Murphy 2013, section 4.2.6). Note that although $\widehat{\Sigma}$ is never invertible for $n < p$, $\mathrm{diag}(\widehat{\Sigma})$ is almost surely invertible whenever the true covariance matrix is invertible and $n > 1$. Thus, the linear shrinkage estimator is almost surely well defined for all $\alpha > 0$. Moreover, it can be efficiently computed in $\mathcal{O}(p^3)$ arithmetic operations.

Definition 6.2. (ℓ_1**-Regularized Maximum Likelihood Estimator**). The ℓ_1-regularized maximum likelihood estimator with penalty parameter $\beta \geq 0$ is defined as

$$X^* = \arg\min_{X \succeq 0} \left\{ -\log\det X + \left\langle \widehat{\Sigma}, X \right\rangle + \beta \sum_{i,j=1}^{p} |X_{ij}| \right\}.$$

Adding an ℓ_1-regularization term to the standard maximum likelihood estimation problem gives rise to sparse—and thus interpretable—estimators (Banerjee et al. 2008, Friedman et al. 2008). The resulting semidefinite program can be solved with general-purpose interior point solvers such as SDPT3 or with structure-exploiting methods such as the QUIC algorithm, which enjoys a quadratic convergence rate and requires $\mathcal{O}(p^3)$ arithmetic operations per iteration (Hsieh et al. 2014). In the remainder of this section we test the Wasserstein shrinkage, linear shrinkage, and ℓ_1-regularized maximum likelihood estimators on synthetic and real data sets. All experiments are implemented in MATLAB, and the corresponding codes are included in the Wasserstein Inverse Covariance Shrinkage Estimator (WISE) package available at https://www.github.com/nvietanh/wise.

Remark 6.3 (Bessel's Correction)**.** So far, we used $\mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ as the nominal distribution, where the sample covariance matrix $\widehat{\Sigma}$ was identified with the (biased) maximum likelihood estimator. In practice, it is sometimes useful to use $\widehat{\Sigma}/\kappa$ as the nominal covariance matrix, where $\kappa \in (0,1)$ is a Bessel correction that removes the bias (see, e.g., Sections 6.2.1 and 6.2.2). Under the premise that $\mathcal{X}$ is a cone, it is easy to see that if (X^*, γ^*) is optimal in (15) for a prescribed Wasserstein radius ρ and a scaled sample covariance matrix $\widehat{\Sigma}/\kappa$, then $(\kappa X^*, \kappa\gamma^*)$ is optimal in (15) for a scaled Wasserstein radius $\sqrt{\kappa}\rho$ and the original sample covariance matrix $\widehat{\Sigma}$. Thus, up to scaling, using a Bessel correction is tantamount to shrinking ρ.

6.1. Experiments with Synthetic Data

Consider a $(p = 20)$-variate Gaussian random vector ξ with zero mean. The (unknown) true covariance matrix Σ_0 of ξ is constructed as follows. We first choose a density parameter $d \in \{12.5\%, 50\%, 100\%\}$. Using the legacy MATLAB 5.0 uniform generator initialized with seed 0, we then generate a matrix $C \in \mathbb{R}^{p\times p}$ with $\lfloor d \times p^2 \rfloor$ randomly selected nonzero elements, all of which represent independent Bernoulli random variables taking the values +1 or −1 with equal probabilities. Finally, we set $\Sigma_0 = (C^\top C + 10^{-3} I)^{-1} \succ 0$.

As usual, the quality of an estimator X^* for the precision matrix Σ_0^{-1} is evaluated using Stein's loss function,

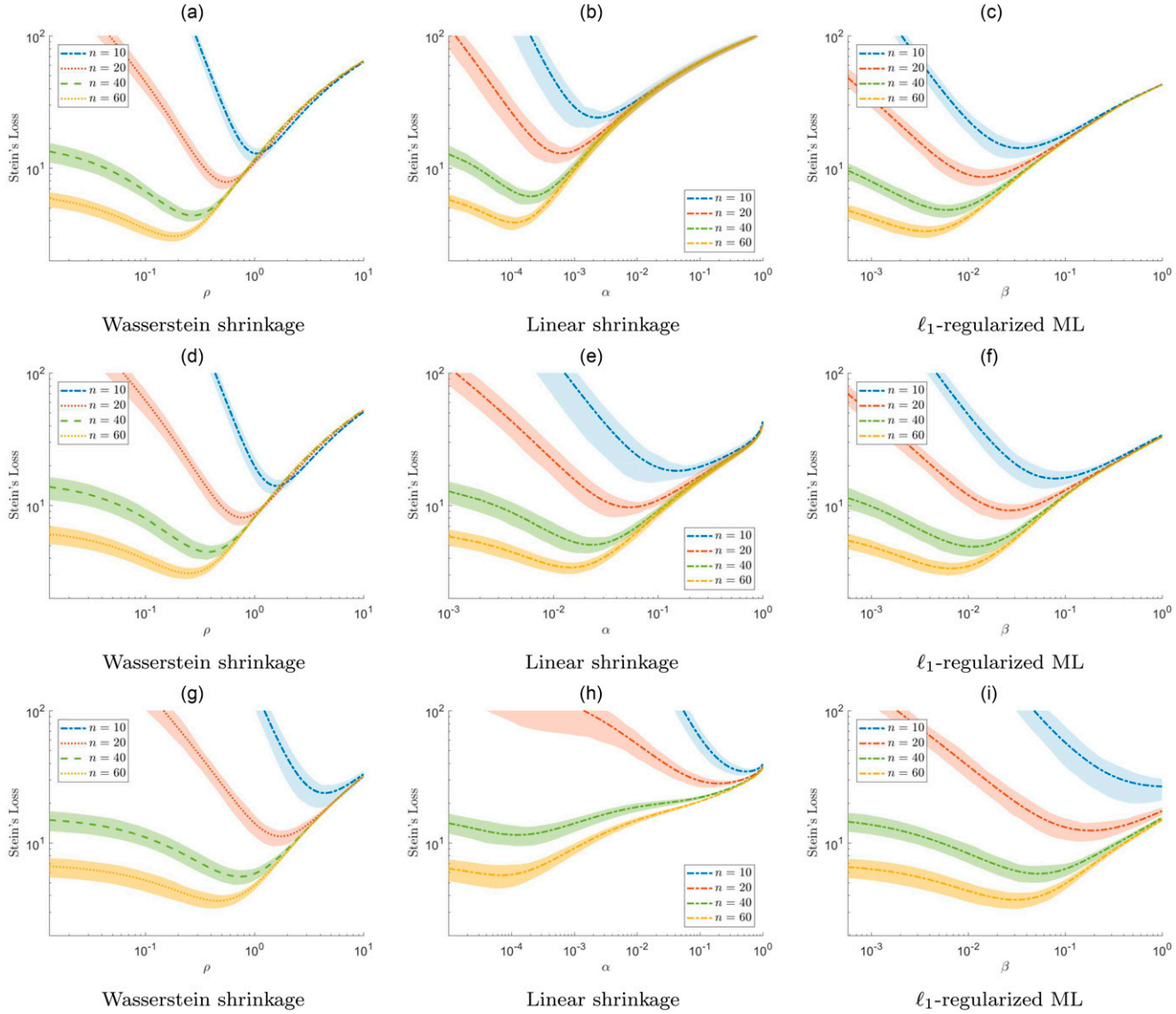
$$L(X^*, \Sigma_0) = -\log\det(X^*\Sigma_0) + \langle X^*, \Sigma_0 \rangle - p,$$

which vanishes if $X^* = \Sigma_0^{-1}$ and is strictly positive otherwise (James and Stein 1961).

All simulation experiments involve 100 independent trials. In each trial, we first draw $n \in \{10, 20, 40, 60\}$ independent samples from $\mathcal{N}(0, \Sigma_0)$, which are used to compute the sample covariance matrix $\widehat{\Sigma}$ and the corresponding precision matrix estimators. Figure 2 shows Stein's loss of the Wasserstein shrinkage estimator without structure information for $\rho \in [10^{-2}, 10^{1}]$, the linear shrinkage estimator for $\alpha \in [10^{-5}, 10^{0}]$, and the ℓ_1-regularized maximum likelihood (ML) estimator for $\beta \in [5 \times 10^{-5}, 10^{0}]$. Lines represent averages, and shaded areas capture the tubes between the empirical 20% and 80% quantiles across all 100 trials. Note that all three estimators approach $\widehat{\Sigma}^{-1}$ when their respective tuning parameters tend to 0. As $\widehat{\Sigma}$ is rank deficient for $n < p = 20$, Stein's loss thus diverges for small tuning parameters when $n = 10$.

The best Wasserstein shrinkage estimator in a given trial is defined as the one that minimizes Stein's loss over all $\rho \geq 0$. The best linear shrinkage and ℓ_1-regularized maximum likelihood estimators are defined analogously. Figure 2 reveals that the best Wasserstein shrinkage estimators dominate the best linear shrinkage and—to a lesser extent—the best ℓ_1-regularized maximum likelihood estimators in terms of Stein's loss for all considered parameter settings. The dominance is more pronounced for small sample sizes. We emphasize that Stein's loss depends explicitly on the unknown true covariance matrix Σ_0. Thus, Figure 2 is not available in practice, and the

Figure 2. (Color online) Stein's Loss of the Wasserstein Shrinkage, Linear Shrinkage, and ℓ_1-Regularized Maximum Likelihood Estimators as a Function of Their Respective Tuning Parameters for $d = 100\%$ (Panels 2(a)–(c)), $d = 50\%$ (Panels (d)–(f)), and $d = 12.5\%$ (Panels (g)–(i))



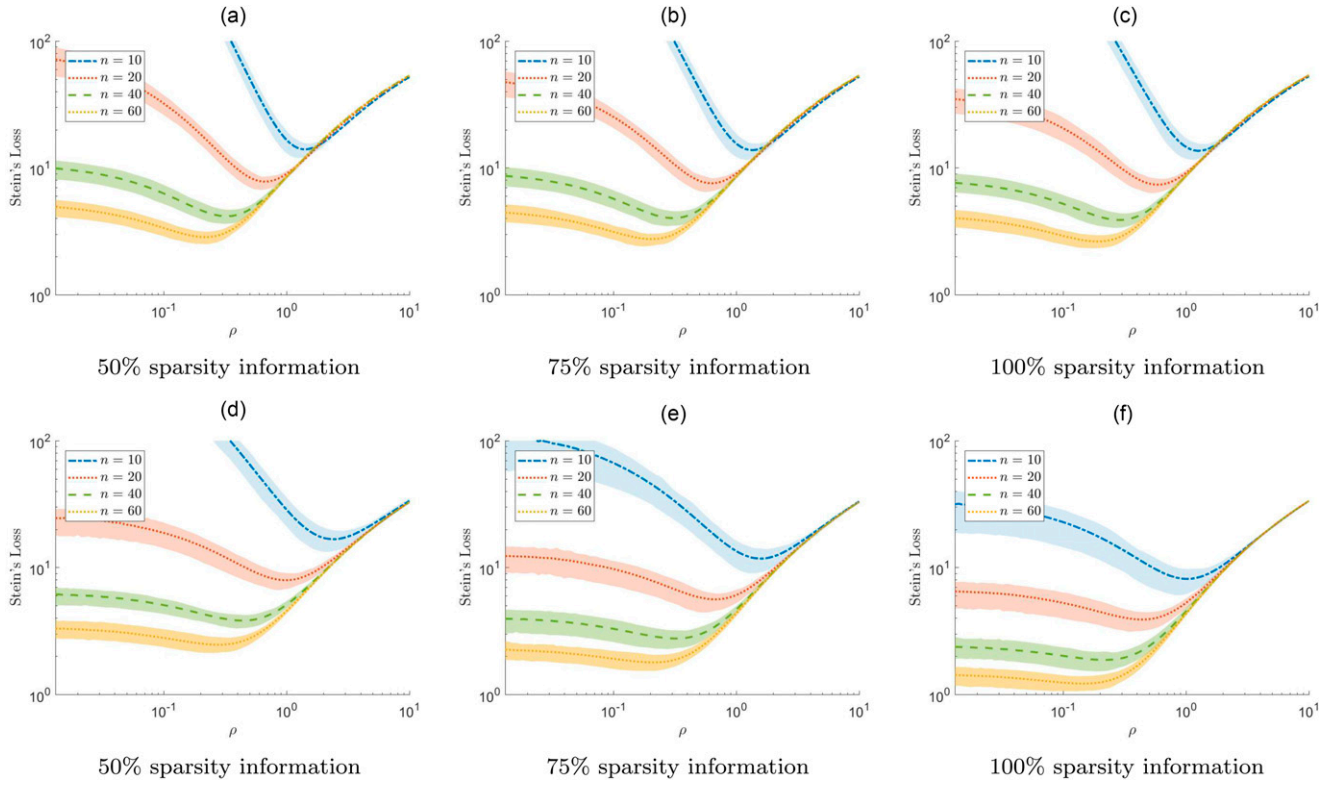
optimal tuning parameters ρ^*, α^*, and β^* cannot be computed exactly. The performance of different precision matrix estimators with *estimated* tuning parameters will be studied in Section 6.2.

For $d = 12.5\%$ and $d = 50\%$, the true precision matrix Σ_0^{-1} has many zeros, and prior knowledge of their positions could be used to improve estimator accuracy. To investigate this effect, we henceforth assume that the feasible set $\mathcal{X}$ correctly reflects a randomly selected portion of 50%, 75%, or 100% of all zeros of Σ_0^{-1}, whereas $\mathcal{X}$ contains no (neither correct nor incorrect) information about the remaining zeros. In this setting, we construct the Wasserstein shrinkage estimator by solving Problem (5) numerically.

Figure 3 shows Stein's loss of the Wasserstein shrinkage estimator with prior information for $\rho \in [10^{-2}, 10^{1}]$. Lines represent averages, and shaded areas capture the tubes between the empirical 20% and 80% quantiles across 100 trials. As expected, correct prior sparsity information improves estimator quality, and the more zeros are known, the better. Note that Σ_0^{-1} contains 21.5% zeros for $d = 12.5\%$ and 68% zeros for $d = 50\%$.

In the last experiment, we investigate the Wasserstein radius ρ^* of the best Wasserstein shrinkage estimator without sparsity information. Figure 4 visualizes the average of ρ^* across 100 independent trials as a function of the sample size n. A standard regression analysis based on the data of Figure 4

Figure 3. (Color online) Stein's Loss of the Wasserstein Shrinkage Estimator with 50%, 75%, or 100% Sparsity Information as a Function of the Wasserstein Radius ρ for $d = 50\%$ (Panels (a)–(c)) and $d = 12.5\%$ (Panels (d)–(f))

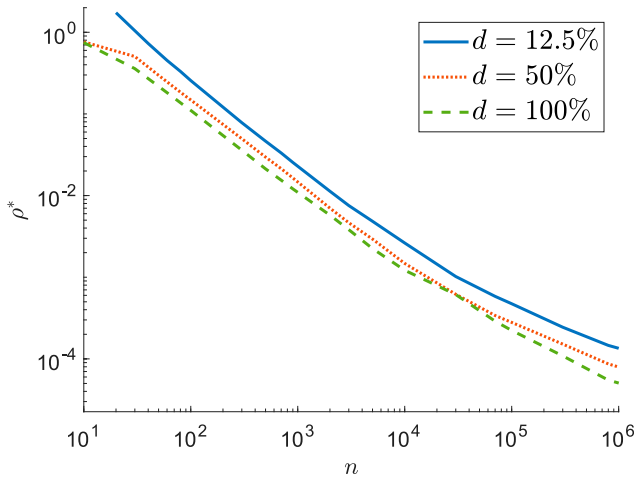


reveals that ρ^* converges to 0 approximately as $n^{-\kappa}$ with $\kappa \approx 61\%$ for $d = 12.5\%$, $\kappa \approx 66\%$ for $d = 50\%$, and $\kappa \approx 68\%$ for $d = 100\%$.

6.2. Experiments with Real Data

We now study the properties of the Wasserstein shrinkage estimator in the context of linear discriminant analysis, portfolio selection, and the inference of solar irradiation patterns.

Figure 4. (Color online) Dependence of the Best Wasserstein Radius ρ^* on the Sample Size n



6.2.1. Linear Discriminant Analysis.

Linear discriminant analysis aims to predict the class $y \in \mathcal{Y}$, $|\mathcal{Y}| < \infty$, of a feature vector $z \in \mathbb{R}^p$ under the assumption that the conditional distribution of z given y is normal with a class-dependent mean $\mu_y \in \mathbb{R}^p$ and class-*in*-dependent covariance matrix $\Sigma_0 \in \mathbb{S}_{++}^p$ (Hastie et al. 2001). If all μ_y and Σ_0 are known, the maximum likelihood classifier $\mathcal{C} : \mathbb{R}^p \to \mathcal{Y}$ assigns z to a class that maximizes the likelihood of observing y; that is,

$$\mathcal{C}(z) \in \arg\min_{y \in \mathcal{Y}} (z - \mu_y)^\top \Sigma_0^{-1} (z - \mu_y). \tag{22}$$

In practice, however, the conditional moments are typically unknown and must be inferred from finitely many training samples $(\widehat{z}_i, \widehat{y}_i)$, $i \le n$. If we estimate μ_y by the sample average

$$\widehat{\mu}_y = \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} \widehat{x}_i,$$

where $\mathcal{I}_y = \{i \in \{1, \ldots, n\} : \widehat{y}_i = y\}$ records all samples in class y, then it is natural to define the residual feature vectors as $\widehat{\xi}_i = \widehat{z}_i - \widehat{\mu}_{\widehat{y}_i}$, $i \le n$. Accounting for

Bessel's correction, the conditional distribution of $\widehat{\xi}_i$ given $\widehat{y}_i$ is normal with mean 0 and covariance matrix $(|\mathcal{I}_{\widehat{y}_i}| - 1)\,|\mathcal{I}_{\widehat{y}_i}|^{-1}\Sigma_0$. The marginal distribution of $\widehat{\xi}_i$ thus constitutes a mixture of $|\mathcal{Y}|$ normal distributions with mean 0, all of which share the same covariance matrix up to a scaling factor close to unity. As such, the residuals fail to be normally distributed. Moreover, because of their dependence on the sample means, the residuals are correlated. However, if each class accommodates many training samples, then the residuals can approximately be regarded as independent samples from $\mathcal{N}(0, \Sigma_0)$.

Irrespective of these complications, the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n - |\mathcal{Y}|} \sum_{i=1}^{n} \widehat{\xi}_i \widehat{\xi}_i^\top$$

provides an unbiased estimator for Σ_0. Indeed, by the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}\left[\widehat{\Sigma}\right] &= \frac{1}{n - |\mathcal{Y}|} \mathbb{E}^{\mathbb{P}}\left[\sum_{i=1}^{n} \mathbb{E}^{\mathbb{P}}\left[\widehat{\xi}_i \widehat{\xi}_i^\top \middle| \widehat{y}_i\right]\right] \\
&= \frac{1}{n - |\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \mathbb{E}^{\mathbb{P}}\left[\sum_{i \in \mathcal{I}_y} \frac{|\mathcal{I}_{\widehat{y}_i}| - 1}{|\mathcal{I}_{\widehat{y}_i}|} \Sigma\right] \\
&= \frac{1}{n - |\mathcal{Y}|} \sum_{y \in \mathcal{Y}} (|\mathcal{I}_y| - 1)\Sigma_0 \;=\; \Sigma_0,
\end{aligned}$$

where $\mathbb{P}$ stands for the unknown true joint distribution of the residuals and class labels. In a data-driven setting, the ideal maximum likelihood classifier (22) is replaced with

$$\widehat{\mathcal{C}}(\xi) = \arg\min_{y \in \mathcal{Y}} \left(\xi - \widehat{\mu}_y\right)^\top X^* \left(\xi - \widehat{\mu}_y\right), \tag{23}$$

which depends on the raw data through the sample averages $\widehat{\mu}_y$, $y \in \mathcal{Y}$, and some precision matrix estimator X^*. The possible choices for X^* include the Wasserstein shrinkage estimator without prior information, the linear shrinkage estimator, and the ℓ_1-regularized maximum likelihood estimator, all of which depend on the data merely through $\widehat{\Sigma}$. Note that the naïve precision matrix estimator $\widehat{\Sigma}^{-1}$ exists only for $n > p$ and is therefore disregarded. All estimators depend on a scalar parameter (the Wasserstein radius ρ, the mixing parameter α, or the penalty parameter β) that can be used to tune the performance of the classifier (23).

We test the classifier (23) equipped with different estimators X^* on two preprocessed data sets from Dettling (2004):

1. The "colon cancer" data set contains 62 gene expression profiles, each of which involves 2,000 features and is classified either as normal tissue (NT) or tumor-affected tissue (TT). The data are split into a training data set of 29 observations (9 in class NT and 20 in class TT) and a test data set of 33 observations (13 in class NT and 20 in class TT).

2. The "leukemia" data set contains 72 gene expression profiles, each of which involves 3,571 features and is classified either as acute lymphocytic leukemia (ALL) or acute myeloid leukemia (AML). The data are split into a training data set of 38 observations (27 in class ALL and 11 in class AML) and a test data set of 34 observations (20 in class ALL and 14 in class AML).

Classification is based solely on the first $p \in \{20, 40, 80, 100\}$ features of each gene expression profile. We use leave-one-out cross-validation on the training data to tune the precision matrix estimator X^* with the goal to maximize the correct classification rate of the classifier (23). To keep the computational overhead manageable, we optimize the tuning parameters over the finite search grids

$$\begin{aligned}
\rho &\in \left\{10^{\frac{j}{20}-1} : j = 0, \ldots, 60\right\}, \\
\alpha &\in \left\{10^{\frac{j}{20}-3} : j = 0, \ldots, 60\right\} \quad \text{and} \\
\beta &\in \left\{10^{\frac{j}{20}-3} : j = 0, \ldots, 60\right\}.
\end{aligned}$$

We highlight that, in the case of the ℓ_1-regularized maximum likelihood estimator, cross-validation becomes computationally prohibitive for $p > 80$ even if the state-of-the-art QUIC routine is used (Hsieh et al. 2014) to solve the underlying semidefinite programs. By contrast, the Wasserstein and linear shrinkage estimators can be computed and tuned quickly even for $p \gg 100$. Once the optimal tuning parameters are found, we fix them and recalculate X^* on the basis of the entire training data set. Finally, we substitute the resulting precision matrix estimator into the classifier (23) and evaluate its correct classification rate on the test data set. The test results are reported in Table 1. We observe that the Wasserstein shrinkage estimator frequently outperforms the linear shrinkage and ℓ_1-regularized maximum likelihood estimators, especially for higher values of p.

6.2.2. Minimum Variance Portfolio Selection. Consider the minimum variance portfolio selection problem without short sale constraints (Jagannathan and Ma 2003):

$$\begin{aligned}
&\min_{w \in \mathbb{R}^p} \; w^\top \Sigma_0 w \\
&\text{s.t. } \mathbb{1}^\top w = 1,
\end{aligned}$$

where the portfolio vector $w \in \mathbb{R}^p$ captures the percentage weights of initial capital allocated to p different assets with random returns, $\mathbb{1} \in \mathbb{R}^p$ stands for the vector of ones, and $\Sigma_0 \in \mathbb{S}_{++}^p$ denotes the covariance matrix of the asset returns. The objective

Table 1. Correct Classification Rate of the Classifier (23) Instantiated with Different Precision Matrix Estimators

	Colon cancer data set				Leukemia data set			
Estimator	$p = 20$	$p = 40$	$p = 80$	$p = 100$	$p = 20$	$p = 40$	$p = 80$	$p = 100$
Wasserstein shrinkage	**72.73**	75.76	**78.79**	**75.76**	**73.53**	67.65	**91.18**	**91.18**
Linear shrinkage	57.58	72.73	72.73	72.73	70.59	**70.59**	82.35	82.35
ℓ_1-regularized ML	**72.73**	**78.79**	**78.79**	72.73	70.59	64.71	82.35	82.35

Note. The best result in each experiment is highlighted in bold.

represents the variance of the portfolio return, which is strictly convex in w thanks to the positive definiteness of Σ_0. The unique optimal solution of this portfolio selection problem is given by $w^* = \Sigma_0^{-1}\mathbb{1}/\mathbb{1}^\top \Sigma_0^{-1}\mathbb{1}$. In practice, the unknown true precision matrix Σ_0^{-1} must be replaced with an estimator X^*, which gives rise to the estimated minimum variance portfolio $\widehat{w}^* = X^*\mathbb{1}/\mathbb{1}^\top X^*\mathbb{1}$.

A vast body of literature in finance focuses on finding accurate precision matrix estimators for portfolio construction; see, for example, Stevens (1998), DeMiguel and Nogales (2009), Ledoit (2004b), Goto and Xu (2015), and Torri et al. (2019). In the following, we compare the minimum variance portfolios based on the Wasserstein shrinkage estimator without structural information, the linear shrinkage estimator, and ℓ_1-regularized maximum likelihood estimator on two preprocessed data sets from the Fama–French online data library:[3] the "48 industry portfolios" data set (FF48) and the "100 portfolios formed on size and book-to-market" data set (FF100). Recall that the estimators depend on the data only through the sample covariance matrix $\widehat{\Sigma}$, which is computed from the residual returns relative to the sample means and thus needs to account for Bessel's correction. The data sets both consist of monthly returns for the period from January 1996 to December 2016. The first 120 observations from January 1996 to December 2005 serve as the training data set. The optimal tuning parameters that minimize the portfolio variance are estimated via leave-one-out cross-validation on the training data set using the finite search grids

$$\rho \in \left\{10^{\frac{j}{100}-2} : j = 0,\ldots,200\right\},$$
$$\alpha \in \left\{10^{\frac{j}{100}-2} : j = 0,\ldots,200\right\}, \quad \text{and}$$
$$\beta \in \left\{10^{\frac{j}{50}-4} : j = 0,\ldots,200\right\}.$$

The out-of-sample performance of the minimum variance portfolio corresponding to a particular precision matrix estimator is then evaluated using the rolling horizon method over the period from January 2006 to December 2016, where the sample covariance matrix needed as an input for the precision matrix is reestimated every three months based on the most recent 120 observations (10 years) while the tuning parameter is kept fixed. The resulting out-of-sample mean, standard deviation, and Sharpe ratio of the portfolio return are reported in Table 2. Although the ℓ_1-regularized maximum likelihood estimator yields the portfolio with the lowest standard deviation for both data sets, the Wasserstein shrinkage estimator always generates the highest mean and, maybe surprisingly, the highest Sharpe ratio.

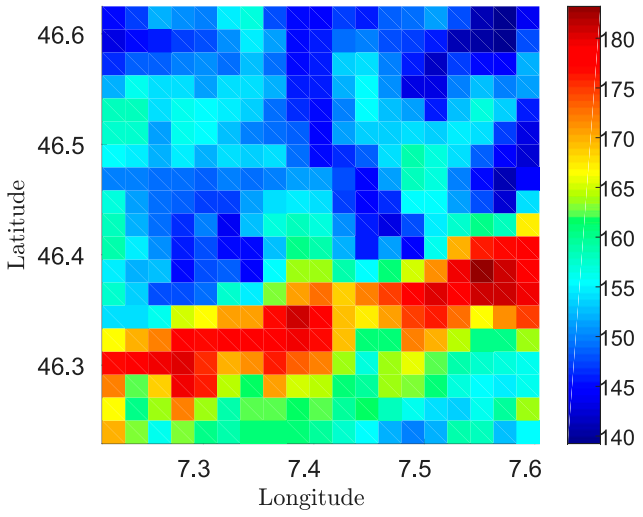
6.2.3. Inference of Solar Irradiation Patterns. In the last experiment we aim to estimate the spatial distribution of solar irradiation in Switzerland using the "surface incoming shortwave radiation" (SIS) data provided by MeteoSwiss.[4] The SIS data capture the horizontal solar irradiation intensities in watts per square meter for pixels of size 1.6 km by 2.3 km based on the effective cloud albedo, which is derived from satellite imagery. The data set spans 13 years, from 2004 to 2016, with a total number of 4,749 daily observations. We deseasonalize the time series of each pixel as follows. First, we divide the original time series by a shifted sinusoid with a yearly period, whose baseline level, phase, and amplitude are estimated via ordinary least squares regression. Next, we subtract unity. The resulting deseasonalized time series is viewed as the sample path of a zero mean Gaussian noise process. This approach relies on the assumption that the mean and the standard deviation of the original time series share the same seasonality pattern. It remains to estimate the joint distribution of the pixel-wise Gaussian white noise processes, which is fully determined by the precision matrix of the deseasonalized data. We estimate the precision matrix using the

Table 2. Standard Deviation, Mean, and Sharpe Ratio of the Minimum Variance Portfolio Based on Different Estimators

	FF48 data set			FF100 data set		
Estimator	SD	Mean	Sharpe	SD	Mean	Sharpe
Wasserstein shrinkage	3.146	**0.701**	**0.223**	3.518	**1.079**	**0.307**
Linear shrinkage	3.152	0.688	0.218	3.484	0.965	0.277
ℓ_1-regularized ML	**3.077**	0.668	0.217	**3.423**	1.010	0.295

Note. The best result in each experiment is highlighted in bold.

Figure 5. (Color online) Average Solar Irradiation Intensities (in Watts per Square Meter) for the Diablerets Region in Switzerland



Wasserstein shrinkage, linear shrinkage, and ℓ_1-regularized maximum likelihood estimators. As each pixel represents a geographical location, and as the solar irradiation intensities at two distant pixels are likely to be conditionally independent given the intensities at all other pixels, it is reasonable to assume that the precision matrix is sparse; see also Das et al. (2017) and Wiesel et al. (2010). Specifically, we assume here that the solar irradiation intensities at two pixels indexed by (i, j) and (i', j') are conditionally independent and that the corresponding entry of the precision matrix vanishes whenever $|i - i'| + |j - j'| > 3$. This sparsity information can be used to enhance the basic Wasserstein shrinkage estimator.

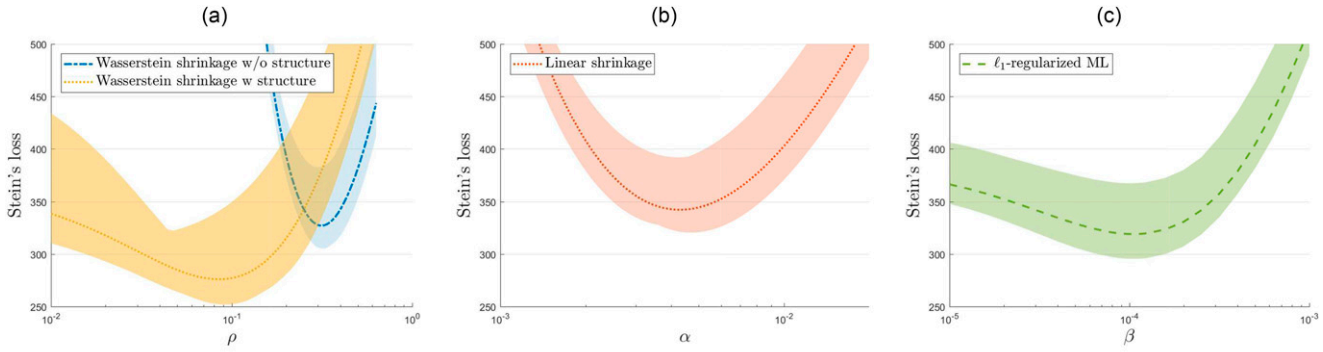
Consider now the Diablerets region of Switzerland, which is described by a spatial matrix of 20×20 pixels. Thus, the corresponding precision matrix has a dimension 400×400. The average daily solar irradiation intensities within the region of interest are visualized in Figure 5. We note that the sunshine exposure is highly variable because of the heterogeneous geographical terrain characterized by a high mountain range in the south intertwined with deep valleys in the north. In order to assess the quality of a specific precision matrix estimator, we use K fold cross-validation with $K = 13$. The kth fold comprises all observations of year k and is used to construct the estimator X_k^*. The data of the remaining 12 years, without year k, are used to compute the empirical covariance matrix $\widehat{\Sigma}_{-k}$. The estimation error of X_k^* is then measured via Stein's loss:

$$L\left(X_k^*, \widehat{\Sigma}_{-k}\right) = -\log\det\left(X_k^* \widehat{\Sigma}_{-k}\right) + \left\langle X_k^*, \widehat{\Sigma}_{-k}\right\rangle - p.$$

We emphasize that here, in contrast to the experiment with synthetic data, $\widehat{\Sigma}_{-k}$ is used as a proxy for the unknown true covariance matrix Σ. Figure 6 shows Stein's loss of the Wasserstein shrinkage estimator with and without structure information for $\rho \in [10^{-2}, 10^{0}]$, the linear shrinkage estimator for $\alpha \in [10^{-3}, 2 \times 10^{-2}]$, and the ℓ_1-regularized maximum likelihood estimator for $\beta \in [10^{-5}, 10^{-3}]$. Lines represent averages, and shaded areas capture the tubes between the best- and worst-case loss realizations across all K folds.

The Wasserstein shrinkage estimator with structure information reduces the minimum average loss by 13.5% relative to the state-of-the-art ℓ_1-regularized maximum likelihood estimator. Moreover, the average runtimes for computing the different estimators amount to 51.84 s for the Wasserstein shrinkage estimator with structural information (Algorithm 1), 0.08 s for the Wasserstein shrinkage estimator without structural information (analytical formula and bisection algorithm), 0.01 s for the linear shrinkage estimator (analytical formula) and 1,493.61 s for the ℓ_1-regularized maximum likelihood estimator (QUIC algorithm; Hsieh et al. 2014).

Figure 6. (Color online) Stein's Loss of the Wasserstein Shrinkage, Linear Shrinkage, and ℓ_1-Regularized Maximum Likelihood Estimators as a Function of Their Respective Tuning Parameters



Endnotes

[1] An elementary calculation shows that $\mathbb{E}^{\mathbb{P}^n}[\widehat{\Sigma}] = \frac{n-1}{n}\Sigma$.

[2] A proof for the continuity of $\mathbb{W}_S$ can be found in lemma A.2 of Nguyen et al. (2021).

[3] See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (accessed January 2018).

[4] See http://www.meteoschweiz.admin.ch/data/products/2014/raeumliche-daten-globalstrahlung.html (accessed January 2018; in German).

References

Banerjee O, El Ghaoui L, d'Aspremont A (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Machine Learn. Res.* 9(June):485–516.

Berge C (1963) *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity* (Dover, Mineola, NY).

Bernstein DS (2009) *Matrix Mathematics: Theory, Facts, and Formulas* (Princeton University Press, Princeton, NJ).

Bertsekas DP (2009) *Convex Optimization Theory* (Athena Scientific, Belmont, MA).

Bien J, Tibshirani RJ (2011) Sparse estimation of a covariance matrix. *Biometrika* 98(4):807–820.

Blanchet J, Murthy K (2016) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

Blanchet J, Si N (2019) Optimal uncertainty size in distributionally robust inverse covariance estimation. *Oper. Res. Lett.* 47(6):618–621.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Chun SY, Browne MW, Shapiro A (2018) Modified distribution-free goodness-of-fit test statistic. *Psychometrika* 83(1):48–66.

Dahl J, Roychowdhury V, Vandenberghe L (2005) Maximum likelihood estimation of Gaussian graphical models: Numerical implementation and topology selection. Working paper, University of California, Los Angeles, Los Angeles.

Das A, Sampson AL, Lainscsek C, Muller L, Lin W, Doyle JC, Cash SS, Halgren E, Sejnowski TJ (2017) Interpretation of the precision matrix and its application in estimating sparse brain connectivity during sleep spindles from human electrocorticography recordings. *Neural Comput.* 29(3):603–642.

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

DeMiguel V, Nogales FJ (2009) Portfolio selection with robust estimation. *Oper. Res.* 57(3):560–577.

Dettling M (2004) BagBoosting for tumor classification with gene expression data. *Bioinformatics* 20(18):3583–3593.

Dey DK, Srinivasan C (1985) Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* 13(4):1581–1591.

Du L, Li J, Stoica P (2010) Fully automatic computation of diagonal loading levels for robust adaptive beamforming. *IEEE Trans. Aerospace Electronic Systems* 46(1):449–458.

Fan J, Fan Y, Lv J (2008) High dimensional covariance matrix estimation using a factor model. *J. Econometrics* 147(1):186–197.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7(2):179–188.

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Gao R, Kleywegt A (2016) Distributionally robust stochastic optimization with Wasserstein distance. Preprint, submitted April 8, https://arxiv.org/abs/1604.02199.

Gao R, Chen X, Kleywegt A (2016) Wasserstein distributional robustness and regularization in statistical learning. Preprint, submitted December 17, https://arxiv.org/abs/1712.06050.

Givens CR, Shortt RM (1984) A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* 31(2):231–240.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4, Part 1):902–917.

Goto S, Xu Y (2015) Improving mean variance optimization through sparse hedging restrictions. *J. Financial Quant. Anal.* 50(6):1415–1441.

Haff LR (1991) The variational form of certain Bayes estimators. *Ann. Statist.* 19(3):1163–1190.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning* (Springer, New York).

Hespanha JP (2009) *Linear Systems Theory* (Princeton University Press, Princeton, NJ).

Hsieh C-J, Sustik MA, Dhillon IS, Ravikumar P (2014) QUIC: Quadratic approximation for sparse inverse covariance estimation. *J. Machine Learn. Res.* 15(83):2911–2947.

Jagannathan R, Ma T (2003) Risk reduction in large portfolios: Why imposing the wrong constraints helps. *J. Finance* 58(4):1651–1683.

James W, Stein C (1961) Estimation with quadratic loss. *Proc. 4th Berkeley Sympos. Math. Statist. Probab., Vol. 1: Contributions to the Theory of Statistics* (University of California Press, Berkeley), 361–379.

Lauritzen SL (1996) *Graphical Models* (Oxford University Press, Oxford, UK).

Ledoit O (2004a) A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* 88(2):365–411.

Ledoit O (2004b) Honey, I shrunk the sample covariance matrix. *J. Portfolio Management* 30(4):110–119.

Ledoit O (2012) Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* 40(2):1024–1060.

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

Löfberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. *2004 IEEE Internat. Conf. Robotics Automation* (IEEE, Piscataway, NJ), 284–289.

Markowitz H (1952) Portfolio selection. *J. Finance* 7(1):77–91.

Mohajerin Esfahani P, Kuhn D (2017) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.

Murphy KP (2013) *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA).

Nocedal J, Wright SJ (2006) *Numerical Optimization* (Springer, New York).

Nguyen VA, Shafieezadeh-Abadeh S, Kuhn D, Mohajerin Esfahani P (2022) Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Math. Oper. Res.* Forthcoming.

Oztoprak F, Nocedal J, Rennie S, Olsen PA (2012) Newton-like methods for sparse inverse covariance estimation. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Red Hook, NY), 755–763.

Pan VY, Chen ZQ (1999) The complexity of the matrix eigenproblem. *Proc. 31st Annual ACM Sympos. Theory Comput.* (ACM, New York), 507–516.

Perlman MD (2007) STAT 542: Multivariate statistical analysis. Lecture notes, University of Washington, Seattle. http://courses.washington.edu/stat512/542Notes2007.pdf.

Ribes A, Azaïs J-M, Planton S (2009) Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynam.* 33(5):707–722.

Rippl T, Munk A, Sturm A (2016) Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* 151(October):90–109.

Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genetics Molecular Biol.* 4(1): Article 32.

Shafieezadeh-Abadeh S, Kuhn D, Mohajerin Esfahani P (2017) Regularization via mass transportation. Preprint, submitted October 27, https://arxiv.org/abs/1710.10016.

Shafieezadeh-Abadeh S, Mohajerin Esfahani P, Kuhn D (2015) Distributionally robust logistic regression. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 28 (Curran Associates, Red Hook, NY), 1576–1584.

Stein C (1975) Estimation of a covariance matrix. Rietz Lecture, 39th Annual Meeting IMS, Institute of Mathematical Statistics, Cleveland.

Stein C (1986) Lectures on the theory of estimation of many parameters. *J. Soviet Math.* 34(1):1373–1403.

Stevens GVG (1998) On the inverse of the covariance matrix in portfolio analysis. *J. Finance* 53(5):1821–1827.

Torri G, Giacometti R, Paterlini S (2019) Sparse precision matrices for minimum variance portfolios. *Comput. Management Sci.* 16(3): 375–400.

Touloumis A (2015) Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Comput. Statist. Data Anal.* 83(March):251–261.

Tseng P, Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Math. Programming* 117(1–2): 387–423.

Tütüncü RH, Toh KC, Todd MJ (2003) Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Programming* 95(2):189–217.

van der Vaart HR (1961) On certain characteristics of the distribution of the latent roots of a symmetric random matrix under general conditions. *Ann. Math. Statist.* 32(3):864–873.

Wiesel A, Eldar Y, Hero A (2010) Covariance estimation in decomposable Gaussian graphical models. *IEEE Trans. Signal Process.* 58(3):1482–1492.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.

Won J-H, Lim J, Kim S-J, Rajaratnam B (2013) Condition number regularized covariance estimation. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* 75(3):427–450.

Yang R, Berger JO (1994) Estimation of a covariance matrix using the reference prior. *Ann. Statist.* 22(3):1195–1211.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.

Viet Anh Nguyen is a postdoctoral scholar in the Department of Management Science and Engineering, Stanford University. His research interests focus on robust decision analytics and machine learning.

Daniel Kuhn holds the chair of risk analytics and optimization at Ecole Polytechnique Fédérale de Lausanne (EPFL). His research interests revolve around robust optimization and stochastic programming.

Peyman Mohajerin Esfahani is an assistant professor in the Delft Center for Systems and Control at the Delft University of Technology. His research interests include theoretical and practical aspects of decision-making problems in uncertain and dynamic environments, with applications to control and security of large-scale and distributed systems.