Distributionally Robust Formulation and Model Selection for the Graphical Lasso

Pedro Cisneros-Velarde
University of California,
Santa Barbara

Sang-Yun Oh
University of California,
Santa Barbara
Lawrence Berkeley National Lab

Alexander Petersen
University of California,
Santa Barbara

Abstract

Building on a recent framework for distributionally robust optimization, we consider inverse covariance matrix estimation for multivariate data. A novel notion of Wasserstein ambiguity set is provided that is specifically tailored to this problem, leading to a tractable class of regularized estimators. Penalized likelihood estimators for Gaussian data, specifically the graphical lasso estimator, are special cases. Consequently, a direction connection is made between the radius of the Wasserstein ambiguity and the regularization parameter, so that the level of robustness of the estimator is shown to correspond to the level of confidence with which the ambiguity set contains a distribution with the population covariance. A unique feature of the formulation is that the radius can be expressed in closed-form as a function of the ordinary sample covariance matrix. Taking advantage of this finding, a simple algorithm is developed to determine a regularization parameter for graphical lasso, using only the bootstrapped sample covariance matrices, rendering computationally expensive repeated evaluation of the graphical lasso algorithm unnecessary. Alternatively, the distributionally robust formulation can also quantify the robustness of the corresponding estimator if one uses an off-the-shelf method such as cross-validation. Finally, a numerical study is performed to analyze the robustness of the proposed method relative to other automated tuning procedures used in practice.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 Introduction

In statistics and machine learning, the covariance matrix Σ of a random vector $X \in \mathbb{R}^d$ is a fundamental quantity for characterizing marginal pairwise dependencies between variables. Furthermore, the inverse covariance matrix $\Omega = \Sigma^{-1}$ provides information about the *conditional* linear dependency structure between the variables. For example, in the case that X is Gaussian, $\Omega_{jk} = 0$ if and only if the jth and kth variables of X are conditionally independent given the rest. Such relationships are of interest in many applications such as environmental science, biology, and neuroscience (Guillot et al., 2015; Huang et al., 2010; Krumsiek et al., 2011), and have given rise to various statistical and machine learning methods for inverse covariance estimation.

Given an independent sample $X_i \sim X$, $i = 1, \ldots, n$, the sample covariance $A_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, which is the maximum likelihood estimator if X is Gaussian, can be a poor estimator of Σ unless d/n is very small. Driven by a so-called high dimensional setting where $n \ll d$, where A_n is not invertible, regularized estimation of the precision matrix has gained significant interest (Cai et al., 2011; Edwards, 2010; Friedman et al., 2007; Khare et al., 2015; Won et al., 2012; Yuan and Lin, 2007). Such regularization procedures are useful even when A_n is a stable estimate (i.e., positive definite with small condition number), since the inverse covariance estimate A_n^{-1} is dense and will not reflect the sparsity of corresponding nonzero elements in Ω.

Distributionally Robust Optimization: Let $\mathbb{S}_{\mathrm{d}}$ be the set of $d \times d$ symmetric matrices, and $\mathbb{S}_{\mathrm{d}}^{++} \subset \mathbb{S}_{\mathrm{d}}$ be the subset of positive definite symmetric matrices. Given a loss function $l(X; K)$ for $K \in \mathbb{S}_{\mathrm{d}}^{++}$ and $X \in \mathbb{R}^d$, a classical approach would be to estimate Ω by minimizing the empirical loss $\sum_{i=1}^n l(X_i; K)$ over $K \in \mathbb{S}_{\mathrm{d}}^{++}$, perhaps including a regularization or penalty term. The error in this estimate arises from

the discrepancy between the true data-generating distribution and the observed training samples, and can be assessed by various tools such as concentration bounds or rates of convergence. In contrast, *distributionally robust optimization* (DRO) is a technique that explicitly incorporates uncertainty about the distribution of the X_i into the estimation procedure. For an introduction on the general topic of DRO, we refer to the works (Shafieezadeh-Abadeh et al., 2015; Blanchet and Murthy, 2016; Shafieezadeh-Abadeh et al., 2019) and the references therein. In the context of inverse covariance estimation, a distributionally robust estimate of Ω is obtained by solving

$$\inf_{K \in \mathbb{S}_{\mathrm{d}}^{++}} \sup_{P \in \mathcal{S}} E_P[l(K; X)], \tag{1}$$

where $\mathcal{S}$, known as an *ambiguity set*, is a collection of probability measures on $\mathbb{R}^d$. As pointed out in (Blanchet and Murthy, 2016), a natural choice for $\mathcal{S}$ is the neighborhood $\{P \mid D(P, \mu) \leq \delta\}$, with μ being a chosen baseline model, δ being some tolerance level which defines the uncertainty size of the ambiguity set, and D being some discrepancy metric between two probability measures. In a practical setting, we have access to some samples (or data points) from the unknown distribution, and thus, a good candidate for the baseline model is the empirical measure.

Very recent work by Nguyen et al. (2018), also analyzed by Blanchet and Si (2019), used the DRO framework to construct a new regularized (dense) inverse covariance estimator. Working under the assumption that X is Gaussian, the authors construct an ambiguity set $\mathcal{S}$ of Gaussian distributions that, up to a certain tolerance level, are consistent with the observed data. Recent work on DRO in other machine learning problems has revealed explicit connections to well-known regularized estimators, specifically regularized logistic regression (Shafieezadeh-Abadeh et al., 2015) and the square-root lasso for linear models (Blanchet et al., 2016); however, such a connection to regularized *sparse* inverse covariance estimators that are used in practice has yet to be made.

The Graphical Lasso: One of the most common methods to recover the sparsity pattern in Ω is to add an l_1-regularization term to the Gaussian likelihood function, motivated by the consideration of the *Gaussian Graphical Model* (GGM). A sparse estimate of $\Omega = \Sigma^{-1}$ is produced by minimizing

$$\mathcal{L}_\lambda(K) = \frac{1}{n} \sum_{i=1}^{n} X_i^\top K X_i - \log |K| + \lambda \sum_{i=1}^{d} \sum_{j=1}^{d} |k_{ij}|, \tag{2}$$

where k_{ij} is the (i, j) entry of K and $\lambda > 0$ is a user-specified regularization parameter (Banerjee et al., 2008; Friedman et al., 2007; Yuan and Lin, 2007). Although several algorithms exist to solve this objective function (Friedman et al., 2007; Rolfs et al., 2012; Hsieh et al., 2014), the minimizer of (2) is often referred to as *graphical lasso* estimator (Friedman et al., 2007). The first two terms of (2) are related to Stein's loss (James and Stein, 1961) when evaluated at the empirical measure, and also correspond to the negative log-likelihood up to an additive constant if X is Gaussian. The performance of the graphical lasso estimator in high-dimensional settings has been investigated (Rothman et al., 2008; Jankova and van de Geer, 2018), as well as modifications and extensions that implement some notion of *robustness*, i.e., for making it robust to outliers or relaxing the normality assumptions in the data (Khare et al., 2015; Lam and Fan, 2009; Loh and Tan, 2018; Xue and Zou, 2012; Yang and Lozano, 2015).

Besides its theoretical relevance, the graphical lasso and its extensions also enjoy many practical advantages. For example, it has been used as a network inference tool. In these applications, the precision matrix can indicate which nodes in a network are conditionally independent given information from remaining nodes, thus giving an indication of network functionality. This has been important in neuroscience applications when studying the inference of brain connectivity (Yang et al., 2015; Smith et al., 2011; Huang et al., 2010). Applications in gene regulatory networks and metabolomics have also been reported (Menéndez et al., 2010; Sulaimanov et al., 2018; Krumsiek et al., 2011).

The performance of the graphical lasso estimator hinges critically on the choice of λ. While there have been studies on how to properly tune λ to obtain a consistent estimator or to establish correct detection of nonzero elements in the precision matrix (Rothman et al., 2008; Banerjee et al., 2008; Mazumder and Hastie, 2012), in practice, this selection is often made through automated methods like cross-validation.

Contributions: In this paper, we propose a distributionally robust reformulation of the graphical lasso estimator in (2). Following Shafieezadeh-Abadeh et al. (2015); Blanchet et al. (2016); Esfahani and Kuhn (2018), we utilize the Wasserstein metric to quantify distributional uncertainty for the construction of the ambiguity set. The following points summarize our main contributions.

- We formulate a class of DRO problems for inverse covariance estimation, leading to a tractable class of ℓ_p-norm regularized estimators. As the graphical lasso estimator (2) is a special case, this provides us with a new interpretation of this pop-

ular technique. This DRO formulation is made possible by a novel type of ambiguity set, now defined as a collection of measures on matrices. This nontrivial adaptation is necessary due to the fact that a direct generalization of other DRO approaches using vector-valued data (e.g., Shafieezadeh-Abadeh et al. (2019)) does not result in a closed-form regularization problem, and thus does not provide the desired connection to the graphical lasso.

- We use this formulation to suggest a criterion for the selection of the regularization parameter in the estimation problem in the classical regime $n > d$. This criterion follows the *Robust Wasserstein Profile* (RWP) inference recently introduced by Blanchet et al. (2016), which makes no assumption on the normality of the data, and which we tailor to our specific problem. The proposed criterion expresses the regularization parameter as an explicit function of the sample covariance A_n, unlike other instances where RWP has been implemented which rely on stochastic dominance arguments.

- We formulate a novel *robust selection* (RobSel) algorithm for regularization parameter choice. Focusing on the graphical lasso, we provide numerical results that compare the performance of cross-validation and our proposed algorithm for the selection of the regularization term.

The paper is organized as follows. In Section 2 we describe our main theoretical result: the distributionally robust formulation of the regularized inverse covariance (log-likelihood) estimation, from which graphical lasso is a particular instance. In Section 3 we propose a criterion for choosing the regularization parameter inspired by this formulation and outline the bootstrap-based RobSel algorithm for its computation. In Section 4 we present some numerical results comparing the proposed criterion of Section 3 with cross-validation. Finally, we state some concluding remarks and future research directions in Section 5. All proofs of theoretical results can be found in the supplementary material.

2 A Distributionally Robust Formulation of the Graphical lasso

First, we provide preliminary details on notation. Given a matrix $A \in \mathbb{R}^{d\times d}$, a_{jk} denotes its (j,k) entry and $\mathbf{vec}(A) \in \mathbb{R}^{d^2}$ denotes its vectorized form, which we assume to be in a row major fashion. For matrices denoted by Greek letters, its entries are simply denoted by appropriate subscripts, i.e. Σ_{jk}. The operator $|\cdot|$, when applied to a matrix, denotes its determinant; when applied to a scalar or a vector, it denotes the absolute value or entry-wise absolute value, respectively. The ℓ_p-norm of a vector is denoted by $\|\cdot\|_p$. We use the symbol $\Rightarrow$ to denote convergence in distribution.

Recall that $X \in \mathbb{R}^d$ is a zero-mean random vector with covariance matrix $\Sigma \in \mathbb{S}_{\mathrm{d}}^{++}$. Let $\mathbb{Q}_0$ be the probability law for X and $\Omega = \Sigma^{-1}$ be the precision matrix. Define the *graphical loss function* as

$$\begin{aligned} l(X;K) &= X^\top K X - \log|K| \\ &= \mathrm{trace}(KXX^\top) - \log|K|. \end{aligned} \tag{3}$$

Then $E_{\mathbb{Q}_0}[l(K;X)] = \mathrm{trace}(K\Sigma) - \log|K|$ is a convex function of K over the convex cone $\mathbb{S}_{\mathrm{d}}^{++}$. Using the first-order optimality criterion, we observe that $K = \Omega$ sets the gradient $\frac{\partial}{\partial K} E_{\mathbb{Q}_0}[l(X,K)] = \Sigma - K^{-1}$ equal to the zero matrix (see (Boyd and Vandenberghe, 2004, Appendix A) for details on this differentiation). Hence,

$$\arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_{\mathbb{Q}_0}[l(X;K)] = \Omega.$$

so that (3) is a consistent loss function.

Now, if we consider an iid random sample $X_1, \cdots, X_n \sim X$, $n > d$, with empirical measure $\mathbb{Q}_n$, then

$$\begin{aligned} \arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_{\mathbb{Q}_n}[l(X;K)] &= \arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} \frac{1}{n}\sum_{i=1}^n l(X_i;K) \\ &= A_n^{-1} \end{aligned}$$

with $A_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. Thus, as described in the Introduction, a natural approach would be to implement the DRO procedure outlined in Esfahani and Kuhn (2018) by building an ambiguity set based on perturbations of $\mathbb{Q}_n$, leading to the DRO estimate given by (1). However, this approach does not convert (1) into a regularized estimation problem as desired, since the inner supremum cannot be explicitly given in closed-form. For more details, see section A in the supplementary material.

As an alternative, let $\mathbb{P}_0$ represent the measure of the random matrix $W = XX^\top$ on $\mathbb{S}_{\mathrm{d}}$ induced by $\mathbb{Q}_0$ and, similarly let $\mathbb{P}_n$ be empirical measure of the sample $W_i = X_i X_i^\top$, $i = 1, \ldots, n$. Redefining the graphical loss function $l : \mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}}^{++}$ as

$$l(W;K) = \mathrm{trace}(KW) - \log|K|, \tag{4}$$

then

$$\begin{aligned} \Omega &= \arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_{\mathbb{P}_0}[l(W;K)] \\ A_n^{-1} &= \arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_{\mathbb{P}_n}[l(W;K)]. \end{aligned}$$

This observation leads to a tractable DRO formulation by constructing ambiguity sets built around the empirical measure $\mathbb{P}_n$. The DRO formulation for inverse covariance estimation becomes

$$\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} \sup_{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P[l(W;K)]. \tag{5}$$

The ambiguity set in this formulation is specified by the collection of measures $\{P \mid \mathcal{D}_c(P,\mathbb{P}_n)\le\delta\}$, which we now describe. Given two probability distributions P_1 and P_2 on $\mathbb{S}_{\mathrm{d}}$ and some transportation cost function $c:\mathbb{S}_{\mathrm{d}}\times\mathbb{S}_{\mathrm{d}}\to[0,\infty)$ (which we will specify below), we define the *optimal transport cost* between P_1 and P_2 as

$$\begin{aligned}\mathcal{D}_c(P_1,P_2) = \inf\{E_\pi\left[c\left(U,V\right)\right]|\pi\in\mathcal{P}\left(\mathbb{S}_{\mathrm{d}}\times\mathbb{S}_{\mathrm{d}}\right),\\ \pi_{_U}=P_1,\ \pi_{_V}=P_2\}\end{aligned} \tag{6}$$

where $\mathcal{P}\left(\mathbb{S}_{\mathrm{d}}\times\mathbb{S}_{\mathrm{d}}\right)$ is the set of joint probability distributions π of (U,V) supported on $\mathbb{S}_{\mathrm{d}}\times\mathbb{S}_{\mathrm{d}}$, and $\pi_{_U}$ and $\pi_{_V}$ denote the marginals of U and V under π, respectively. In this paper, we are interested in cost functions

$$c(U,V) = \|\mathbf{vec}(U)-\mathbf{vec}(V)\|_q^\rho, \tag{7}$$

with $U,V\in\mathbb{S}_{\mathrm{d}}$, $\rho\ge1$, $q\in[1,\infty]$. As pointed out by Blanchet et al. (2016), the resulting optimal transport cost $\mathcal{D}_c^{1/\rho}$ is the Wasserstein distance of order ρ. Our first theoretical result demonstrates that the optimization in (5) corresponds to a class of regularized estimators under the graphical loss function (4).

Theorem 2.1 (DRO formulation of regularized inverse covariance estimation)**.** *Consider the cost function in* (7) *for a fixed* $\rho\ge1$*. Then,*

$$\begin{aligned}&\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} \sup_{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P\big[l(W;K)\big]\\ &= \min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} \Big\{\mathrm{trace}(KA_n) - \log|K| + \delta^{1/\rho}\left\|\mathbf{vec}(K)\right\|_p\Big\},\end{aligned} \tag{8}$$

where $\frac{1}{p}+\frac{1}{q}=1$.

Theorem 2.1 is a remarkable theoretical result that provides a mapping between the regularization parameter and the uncertainty size δ of the ambiguity set in the DRO formulation. Then, the regularization problem reduces to determining a good criterion for choosing δ, which we explore in Section 3. Moreover, we obtain the *graphical lasso* formulation (2) by setting $q=\infty$ in (7). From (8), a smaller ambiguity set implies less robustness being introduced in the estimation problem by reducing the importance of the regularization term. Conversely, a larger regularization term increases the number of nuisance distributions inside the ambiguity set, and thus the robustness.

Remark 2.2. *The ambiguity set used in* (8) *makes no assumptions on the normality of the distribution of the samples* $\{X_1,\dots,X_n\}$*. Then,* (8) *tells us that adding a penalization to the precision matrix gives a robustness in terms of the distributions that the samples may have, which do not necessarily have to be Gaussian for this formulation to hold. Furthermore, it holds independent of the relationship between* n *and* d*.*

3 Selection of the regularization parameter

This section follows closely the line of thought recently introduced by Blanchet et al. (2016) in the analysis of regularized estimators under the DRO formulation. Specifically, we will demonstrate that the ambiguity set $\{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta\}$ represents a confidence region for $\Omega=\Sigma^{-1}$, and use the techniques of Blanchet et al. (2016) to explicitly connect the amibiguity size δ with a confidence level. As previously stated, $l(W;K)$ is a differentiable function on $K\in\mathbb{S}_{\mathrm{d}}^{++}$ with $\frac{\partial}{\partial K}l(W;K) = W-K^{-1}$, so that

$$E_{\mathbb{P}_0}\left[\frac{\partial}{\partial K}l(W;K)\Big|_{K=\Omega}\right] = \mathbb{0}_{d\times d}. \tag{9}$$

Hence, even though the loss function $l(W;K)$ has been inspired from the log-likehood estimation of the covariance matrix Σ for samples of Gaussian random vectors, equation (9) is transparent to any underlying distribution of the data. For any $K\in\mathbb{S}_{\mathrm{d}}^{++}$, define the set

$$\begin{aligned}\mathcal{O}(K) := \Big\{P\in\mathcal{P}(\mathbb{S}_{\mathrm{d}})\Big|\\ E_P\left[\frac{\partial}{\partial K'}l(W;K')\Big|_{K'=K}\right] = \mathbb{0}_{d\times d}\Big\},\end{aligned} \tag{10}$$

corresponding to all probability measures with covariance K^{-1}, i.e. for which K is an optimal loss minimization parameter; here, $\mathcal{P}(\mathbb{S}_{\mathrm{d}})$ denotes the set of all probability distributions supported on $\mathbb{S}_{\mathrm{d}}$. Thus, $\mathcal{O}(\Omega)$ contains all probability measures with covariance matrix agreeing with that of X.

Implicitly, the Wasserstein ambiguity set $\{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta\}$ is linked to the collection of covariance matrices

$$\begin{aligned}\mathcal{C}_n(\delta) :=&\{K\in\mathbb{S}_{\mathrm{d}}^{++}|\text{ there exists } P\in\mathcal{O}(K)\\ &\cap\{P\mid\mathcal{D}_c(P,\mathbb{P}_n)\le\delta\}\}\\ =&\bigcup_{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta}\ \arg\min_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_P[l(W;K)].\end{aligned} \tag{11}$$

We refer to $\mathcal{C}_n(\delta)$ as the set of *plausible* selections for Ω.

Lemma 3.1 (Interchangeability in the DRO formulation)**.** *Consider the setting of Theorem 2.1. Then, for*

$n > d$, the following holds with probability one:

$$\begin{aligned}&\inf_{K\in\mathbb{S}_{\mathrm{d}}^{++}}\ \sup_{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P\left[l(W;K)\right]\\ &=\sup_{P:\ \mathcal{D}_c(P,\mathbb{P}_n)\le\delta}\ \inf_{K\in\mathbb{S}_{\mathrm{d}}^{++}} E_P\left[l(W;K)\right].\end{aligned}\tag{12}$$

Lemma 3.1 states that any estimator obtained by minimizing the left-hand side of (12) must be in $\mathcal{C}_n(\delta)$, otherwise the right-hand side of (12) would be strictly greater than the left. Thus, in line with the goal of providing a robust estimator, the idea is to choose δ so that $\mathcal{C}_n(\delta)$ also contains the true inverse covariance matrix Ω with high confidence.

As $\mathbb{P}_0$ is the weak limit of $\mathbb{P}_n$, we will eventually have that $\Omega \in \mathcal{C}_n(\delta)$ with high probability for any δ, so that $\mathcal{C}_n(\delta)$ is a confidence region for Ω. From this observation, we can choose the uncertainty size δ optimally by the criterion

$$\delta = \inf\left\{\delta > 0 \mid \mathbb{P}_0(\Omega \in \mathcal{C}_n(\delta)) \ge 1-\alpha\right\},\tag{13}$$

i.e., for a specified confidence level $1-\alpha$, we choose δ so that $\mathcal{C}_n(\delta)$ is a $(1-\alpha)$-confidence region for Ω.

To continue our anlaysis, we make use of the so-called *Robust Wasserstein Profile (RWP) function* R_n introduced by Blanchet et al. (2016),

$$\begin{aligned}R_n(K) &= \inf\left\{\mathcal{D}_c(P,\mathbb{P}_n) \mid P \in \mathcal{O}(K)\right\}\\ &= \inf\Big\{\mathcal{D}_c(P,\mathbb{P}_n)\Big|\\ &\qquad E_P\left[\frac{\partial}{\partial K'} l(W;K')\Big|_{K'=K}\right] = \mathbb{0}_{d\times d}\Big\},\end{aligned}\tag{14}$$

for $K \in \mathbb{S}_{\mathrm{d}}^{++}$, which has the geometric interpretation of being the minimum distance between the empirical distribution and any distribution that satisfies the optimality condition for the precision matrix K. Then, using the equivalence of events $\{\Omega \in \mathcal{C}_n(\delta)\} = \{\mathcal{O}(\Omega) \cap \{P \mid \mathcal{D}_c(P,\mathbb{P}_n) \le \delta\} \ne \varnothing\} = \{R_n(\Omega) \le \delta\}$, (13) becomes equivalent to

$$\delta = \arg\inf\left\{\delta > 0 \mid \mathbb{P}_0(R_n(\Omega) \le \delta) \ge 1-\alpha\right\},\tag{15}$$

i.e., the optimal selection of δ is the $1-\alpha$ quantile of $R_n(\Omega)$. Indeed, the set $\{P \mid \mathcal{D}_c(P,\mathbb{P}_n) \le R_n(\Omega)\}$ is the smallest ambiguity set around the empirical measure $\mathbb{P}_n$ such that there exists a distribution for which Ω is an optimal loss minimization parameter. In contrast to previously reported applications of the RWP function on linear regression and logistic regression (Blanchet et al., 2016), our problem allows for a (finite sample) closed form expression of this function. This is due to the fact that we have recast the covariance Σ as the mean of the random matrix $XX^\top$, so that the following result gives a nontrivial generalization of (Blanchet et al., 2016, Example 3).

Theorem 3.2 (RWP function)**.** *Consider the cost function in* (7) *for a fixed $\rho \ge 1$. For $K \in \mathbb{S}_{\mathrm{d}}^{++}$, consider $R_n(K)$ as in* (14)*. Then,*

$$R_n(K) = \left\|\mathbf{vec}(A_n - K^{-1})\right\|_q^\rho.\tag{16}$$

We now establish important convergence guarantees on the RWP function in the following corollary.

Corollary 3.3 (Asymptotic behavior of the RWP function)**.** *Suppose that the conditions of Theorem 3.2 hold, and that $E_{\mathbb{Q}_0}(\|X\|_2^4) < \infty$. Let $H \in \mathbb{S}_{\mathrm{d}}$ be a matrix of jointly Gaussian random variables with zero mean and such that $\mathrm{Cov}(h_{ij}, h_{k\ell}) = E[w_{ij}w_{k\ell}] - \Sigma_{ij}\Sigma_{k\ell} = E[x_ix_jx_kx_\ell] - \Sigma_{ij}\Sigma_{k\ell}$. Then,*

$$n^{\rho/2} R_n(\Omega) \Rightarrow \|\mathbf{vec}(H)\|_q^\rho.\tag{17}$$

Proof. By the central limit theorem, we observe that $\sqrt{n}(A_n - \Sigma) \Rightarrow H$, and by the continuous mapping theorem, we get that

$$n^{\rho/2} R_n(\Omega) = \left\|\sqrt{n}\mathbf{vec}(A_n - \Sigma)\right\|_q^\rho \Rightarrow \|\mathbf{vec}(H)\|_q^\rho.$$

□

Remark 3.4. *Turning our attention back to Theorem 2.1, a robust selection for the ambiguity size or regularization parameter $\lambda = \delta^{1/\rho}$, as obtained from Theorem 3.2, is*

$$\delta^{1/\rho} = \inf\left\{\delta > 0 \mid \mathbb{P}_0(\|\mathbf{vec}(A_n - \Sigma)\|_q \le \delta) \ge 1-\alpha\right\}\tag{18}$$

As a result, this robust selection for λ results in a class of estimators, given by minimizers of the right-hand side of (8)*, that are invariant to the choice of ρ in* (7)*. Thus, for simplicity, we will set $\rho = 1$ in the remainder of the paper.*

Remark 3.5. *Let $r_{1-\alpha}$ be the $(1-\alpha)$ quantile from the distribution of the right-hand side of* (17)*. Then, for any fixed α, the robust selection δ in* (18) *satisfies $n^{1/2}\delta \to r_{1-\alpha}$, so that the optimal decay rate of $n^{-1/2}$ for λ is automatically chosen by the RWP function.*

As solving (18) requires knowledge of Σ, we now outline the robust selection (RobSel) algorithm for data-adaptive choice of the regularization parameter δ for our inverse covariance estimation with an ℓ_p penalization parameter. The special case $p = 1$ corresponds to the graphical lasso in (2), in which case we will also use the notation $\delta = \lambda$. The asymptotic result in Corollary 3.3 invokes a central limit theorem, and thus motivates the approximation of the RWP function through bootstrapping, which we further explain and evaluate its numerical performance in the next section. Let $\alpha \in (0,1)$ be a prespecified confidence level

Algorithm RobSel algorithm for estimation of the regularization parameter λ

1: For $b = 1, \ldots, B$, obtain a bootstrap sample $X^*_{1b}, \ldots, X^*_{nb}$ by sampling uniformly and with replacement from the data, and compute the bootstrap RWP function $R^*_{n,b} = \left\| A^*_{n,b} - A_n \right\|_q$, with the empirical covariance $A^*_{n,b}$ computed from the bootstrap sample.
2: Set λ to be the bootstrap order statistic $R^*_{n,((B+1)(1-\alpha))}$.

and B a large integer such that $(B+1)(1-\alpha)$ is also an integer.

RobSel can potentially provide considerable computational savings over cross-validation in practice. Computing sample covariance matrices for each of B bootstrap samples has cost $O(Bnd^2)$. On the other hand, it is known that each iteration of graphical lasso can cost $O(d^3)$ in the worst case (Mazumder and Hastie, 2012); therefore, performing an F-fold cross-validation to search over L-grid of regularization parameters, each taking T-iterations of graphical lasso, would cost $O(FLTd^3)$.

4 Numerical results and analysis

The true precision matrix $\Omega \in \mathbb{S}^{++}_{\mathrm{d}}$ used to generate simulated data has been constructed as follows. First, generate an adjacency matrix of an undirected Erdős-Renyi graph with equal edge probability of 0.1 and without self-loops. Then, the weight of each edge is sampled uniformly between $[0.5, 1]$, and the sign of each non-zero weight is positive or negative with equal probability of 0.5. Finally, the diagonal entries of this weighted adjacency matrix are set to 1 and the matrix is made diagonally dominant by following a procedure described in (Peng et al., 2009), which ensures that the resulting matrix Ω is positive definite. Throughout this numerical study section, a randomly generated sparse matrix Ω (edge probability 0.1 and $d = 100$) is fixed. Using this Ω, a total of $N = 200$ datasets (of varying size n) were generated as independent observations from a multivariate zero-mean Gaussian distribution, i.e., $\mathcal{N}(\mathbb{0}_d, \Omega^{-1})$.

Consider the problem of choosing the regularization parameter λ (equivalently, the ambiguity size parameter δ) to obtain graphical lasso estimates $\hat{K}_\lambda$ of Ω using the simulated datasets. An R software package, `glasso`, from CRAN was used throughout our numerical experiments. Below, we compare two different criteria for choosing λ. The first criterion is *Robust Selection (RS)*, which follows our proposed RobSel algorithm with $B = 200$ sets of bootstrap samples. We present here results mainly for $n > d$, but additional results in the high-dimensional regime $n < d$ can be found in the supplementary material. The second criterion is a 5-*fold cross-validation (CV)* procedure[1]. The performance on the validation set is the evaluation of the graphical loss function under the empirical measure of the samples on the training set.

Recall the elements in the confusion matrix to be true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). We compare model selection performance of λ chosen by the two different approaches: λ_{RS} and λ_{CV}. The following comparison metrics are used:

- *True positive rate (TPR) and false detection rate (FDR)*: $TPR = \frac{TP}{TP+FN}$ is the proportion of nonzero entries of Ω that are correctly identified in $\hat{K}_\lambda$, and $FDR = \frac{FP}{FP+TP}$ is the proportion of zero entries of Ω that are incorrectly identified as nonzeros in $\hat{K}_\lambda$.
- *Matthew's Correlation Coefficient (MCC)*: MCC summarizes all counts in the confusion matrix in contrast to other measures like TPR and FDR. More details about MCC is given in supplemental subsection D.1.

In the remainder of this section, we compare the model selection performance (FDR, TPR, MCC) from our simulation results. Additional comparison metrics can be found in the supplemental subsection D.3.

As mentioned in Remark 3.5, supplemental subsection D.2 shows that λ_{RS} decreases as n increases as is also observed to be true with λ_{CV}. Furthermore, across the tested range of α, the regularization λ_{RS} are all larger than λ_{CV} for any n. Then, our distributionally robust representation (8) allows us to observe that even for small values of n, CV always chooses a λ that corresponds to smaller ambiguity sets than RS.

To assess the accuracy of RobSel in estimating $\lambda = \delta$ for a given α, we approximated the right-hand side of

[1] For each dataset, we set up the grid for the choices of λ in the CV algorithm as follows. First, we obtained the sample covariance using the whole dataset and obtained the maximum absolute value of its entries, which we denote by $s_{\max}$. Then, we created a grid of ten values in the interval $(0, s_{\max}]$ and ran the graphical lasso over these values. Then, we determined the maximum value $s^*_{\max}$ from this grid such that the estimated inverse covariance matrix has in its off-diagonal upper triangular part: 1) at least one non-zero element for $n \leq 200$, or 2) at least 5% non-zero entries for $n > 200$. Finally, a grid of 100 values in the interval $(0, s^*_{\max}]$ was used as the grid for λ in the CV algorithm. We always had the case that the obtained λ was less than $s^*_{\max}$.

(18) using the $N = 200$ data sets and the true covariance Σ, giving the "true" value λ_{RWP}. Figures in supplemental subsection D.3 show that the performance obtained by λ_{RWP} is similar to the one obtained by λ_{RS} for all comparison metrics. This finite sample behavior of RS indicates that the RobSel bootstrap algorithm reliably approximates the desired robustness level corresponding to the choice of α. These plots also indicate that the RS criterion is more conservative than CV in achieving a lower FDR across different sample sizes, due to providing larger values for λ (see supplemental subsection D.2). More specifically, Fig. 1 shows that RS gives a better performance than CV in terms of FDR even for smaller values of n and this performance improves even more as n increases.

Moreover, the trade-off between the preference for robustness and the preference for a higher density estimation of nonzero entries in the precision matrix can be observed in terms of the Matthews correlation coefficient (MCC), as shown in Fig. 2. Higher values in the curve of MCC implies values of λ that describe a better classification of the entries of Ω as either zero or nonzero (see supplemental subsection D.1 for more details). We observe that cross-validation chooses λ_{CV} that are smaller than λ corresponding to the highest achievable MCC and overestimates the number of nonzero entries in Ω. On the other handRS chooses λ_{RS} that underestimates the number of nonzero entries in Ω induced by its robust nature. Then, it is up to the experimenter to know which method to choose depending on their false discovery rate tolerance. Remarkably, for large numbers of n, RS seems to be much closer to the highest achievable MCC performance than CV, and it does this by maintaining a lower FDR than CV while increasing its TPR.

Our results from the MCC analysis and supplemental subsection D.3 also indicate that we should aim for higher values of α if we want a performance closer to CV in terms of TPR when using the RS criterion, with the advantage of still maintaining a better performance than CV in terms of the FDR. In contrast, if we want more conservative results, we should aim for lower values of α. This is a good property of RS: it allows the use of an intuitive single parameter $\alpha \in (0, 1)$ to adjust the importance of the regularization term. Furthermore, as mentioned in section 3, an added practical benefit of RS is that it provides a candidate for λ with potential computational savings over CV.

5 Conclusion

We provide a recharacterization of the popular graphical lasso estimator in (2) as the optimal solution to a distributionally robust optimization problem. To the best of our knowledge, this is the first work to make such a connection for sparse inverse covariance estimation. The DRO form of the estimator leads to a reinterpretation of the regularization parameter as the radius of the distributional amibiguity set, which can be chosen based on a desired level of robustness. We propose the RobSel method for the selection of the regularization parameter and compare it to cross-validation for the graphical lasso. In our numerical experiments, RobSel gives a better false detection rate than cross-validation, and, as the sample size increases, other performance metrics like the true positive rate for the two are similar. Moreover, RobSel is a computationally simpler procedure, notably only performing the graphical lasso algorithm once at the final step rather than repeatedly as is necessary for cross-validation. Future work includes theoretical justification for robust selection of the regularization parameter for the graphical lasso in the high-dimensional setting.

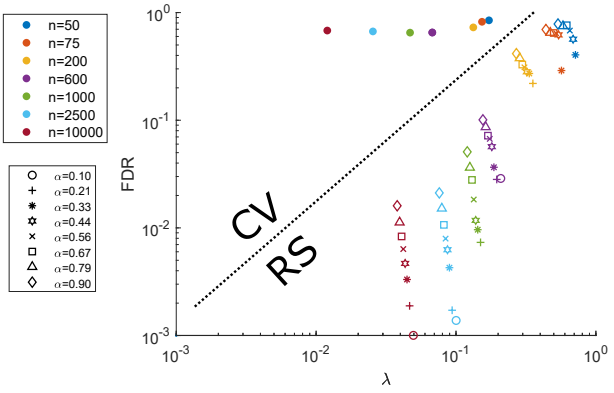

Figure 1: False Detection Rate (FDR) for seven different sample sizes, with both axes in logarithmic scale. For each sample size, the average FDR is plotted for both criteria, cross-validation (CV) and RobSel (RS). For RS, a point is plotted with a different symbol for each different value of the parameter α (some points may not be plotted for lower values of α, since those values gave no true positive detected, and so FDR was not well-defined). The dotted line observed in the figure simply emphasizes the consistent gap (across simulation parameters) between RS and CV in terms of the average FDR.

References

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008. ISSN 1532-4435.

J. Blanchet and K. Murthy. Quantifying distribu-

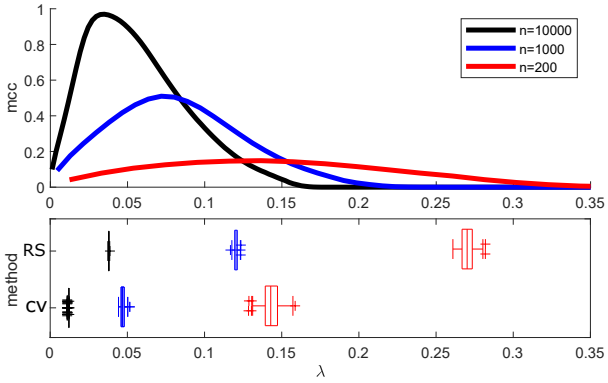

Figure 2: Matthews correlation coefficient (MCC) for three different sample sizes n. The curves in the upper plot are the average MCC obtained over $N = 200$ datasets as a function of the regularization parameter λ. The lower plot are boxplots for the obtained values of λ from the CV and RS (with parameter $\alpha = 0.9$) methods over the N datasets for the different choices of n.

tional model risk via optimal transport. 2016. doi:arXiv:1604.01446v2. URL `https://arxiv.org/pdf/1604.01446.pdf`.

J. Blanchet and N. Si. Optimal uncertainty size in distributionally robust inverse covariance estimation. 2019. doi:arXiv:1901.07693. URL `://arxiv.org/pdf/1901.07693.pdf`.

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. 2016. doi:arXiv:1610.05627v2. URL `https://arxiv.org/pdf/1610.05627.pdf`.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 0521833787.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, jun 2011. doi:10.1198/jasa.2011.tm10155.

D. Edwards. *Introduction to Graphical Modelling*. Springer-Verlag New York, 2010. ISBN 978-0-387-95054-9. doi:10.1007/978-1-4612-0493-0.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2): 115–166, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007. ISSN 1465-4644. doi:10.1093/biostatistics/kxm045.

D. Guillot, B. Rajaratnam, and J. Emile-Geay. Statistical paleoclimate reconstructions via markov random fields. *Ann. Appl. Stat.*, 9(1):324–352, 03 2015. doi:10.1214/14-AOAS794.

C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Quic: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, 15:2911–2947, 2014.

S. Huang, J. Li, L. Sun, J. Ye, A. Fleisher, T. Wu, K. Chen, and E. Reiman. Learning brain connectivity of Alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935 – 949, 2010. ISSN 1053-8119.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif., 1961. University of California Press.

J. Jankova and S. van de Geer. Inference in high-dimensional graphical models. 2018. doi:arXiv:1801.08512. URL `https://arxiv.org/pdf/1801.08512.pdf`.

K. Khare, S.-Y. Oh, and B. Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4): 803–825, sep 2015. doi:10.1111/rssb.12088.

J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21, Jan 2011. ISSN 1752-0509. doi:10.1186/1752-0509-5-21.

C. Lam and J. Fan. Sparsity and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, 2009.

P.-L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Statist.*, 12(1):1429–1467, 2018. doi:10.1214/18-EJS1427.

R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.*, 13(1):781–794, 2012. ISSN 1532-4435.

P. Menéndez, Y. Kourmpetis, C. ter Braak, and F. van Eeuwijk. Gene regulatory networks from multifactorial perturbations using graphical lasso: Application to the dream4 challenge. *PLOS ONE*, 5(12):1–8, 12 2010. doi:10.1371/journal.pone.0014147.

V. A. Nguyen, D. Kuhn, and P. M. Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. 2018. doi:arXiv:1805.07194. URL `https://arxiv.org/pdf/1805.07194.pdf`.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104 (486):735–746, 2009. doi:10.1198/jasa.2009.0126.

B. Rolfs, B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1574–1582. Curran Associates, Inc., 2012.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515, 2008. doi:10.1214/08-EJS176.

S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584. 2015.

S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20, 2019.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54(2):875 – 891, 2011. ISSN 1053-8119.

N. Sulaimanov, S. Kumar, H. Koeppl, F. Burdet, M. Pagni, and M. Ibberson. Inferring gene expression networks with hubs using a degree weighted Lasso approach. *Bioinformatics*, 35(6):987–994, 08 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty716.

J.-H. Won, J. Lim, S.-J. Kim, and B. Rajaratnam. Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):427–450, dec 2012. doi:10.1111/j.1467-9868.2012.01049.x.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional paranormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

E. Yang and A. C. Lozano. Robust Gaussian graphical modeling with the trimmed graphical lasso. In *Advances in Neural Information Processing Systems 28*, pages 2602–2610. 2015.

S. Yang, Q. Sun, S. Ji, P. Wonka, I. Davidson, and J. Ye. Structural graphical lasso for learning mouse brain connectivity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. doi:10.1145/2783258.2783391.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1): 19–35, 2007. doi:10.1093/biomet/asm018.