Distributionally Robust Formulation and Model Selection for the Graphical Lasso: Supplementary Material

A Using a different ambiguity set

Recall that $X \in \mathbb{R}^d$ is a zero-mean random vector with covariance matrix $\Sigma \in \mathbb{S}_{\mathrm{d}}^{++}$ and measure $\mathbb{Q}_0$, and that we consider an iid random sample $X_1, \cdots, X_n \sim X$, $n > d$, with empirical measure $\mathbb{Q}_n$. Since we use the graphical loss function as in equation (3), we are interested in finding a tractable or closed-form expression for the optimization problem

$$\sup_{Q\colon\ \mathcal{D}_{c'}(Q,\mathbb{Q}_n)\leq\delta} E_Q[l(X;K)] \tag{1}$$

with $K \in \mathbb{S}_{\mathrm{d}}^{++}$. Ideally, the solution should be connected to the graphical lasso estimator, since it is one of the most commonly-used sparse inverse covariance estimators in practice. The ambiguity set in this formulation is specified by the collection of measures $\{Q \mid \mathcal{D}'_{c'}(Q, \mathbb{Q}_n) \leq \delta\}$, which we now describe. Given two probability distributions Q_1 and Q_2 on $\mathbb{R}^d$ and some transportation cost function $c' : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$ (which we will specify below), we define the *optimal transport cost* between Q_1 and Q_2 as

$$\begin{aligned} \mathcal{D}'_{c'}(Q_1, Q_2) = \inf\{E_\pi\left[c'\left(u, v\right)\right] |\pi \in \mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right), \\ \pi_u = Q_1,\ \pi_v = Q_2\} \end{aligned} \tag{2}$$

where $\mathcal{P}\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ is the set of joint probability distributions π of (u, v) supported on $\mathbb{R}^d \times \mathbb{R}^d$, and π_u and π_v denote the marginals of u and v under π, respectively. In this paper, we are interested in cost functions

$$c'(u, v) = \|u - v\|_q^\rho\,, \tag{3}$$

with $u, v \in \mathbb{R}^d$, $\rho \geq 1$, $q \in [1, \infty]$.

Now, observe that the function $l(\cdot; K) : \mathbb{R}^d \to \mathbb{R}$ is Borel measurable since it is a continuous function. Then, we use the duality result from Proposition 4 of (Blanchet et al., 2016, version 2) and obtain

$$\sup_{P\colon\ \mathcal{D}'_{c'}(Q,\mathbb{Q}_n)\leq\delta} E_Q\big[l(X;K)\big] = \inf_{\gamma\geq 0}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^{n}\left(\sup_{u\in\mathbb{R}^d}\{l(u;K) - \gamma c'(u, X_i)\}\right)\right\}. \tag{4}$$

Let $\Delta := u - X_i$. Then

$$\begin{aligned} &\sup_{u\in\mathbb{R}^d}\{l(u;K) - \gamma c(u, X_i)\} \\ &= \sup_{u\in\mathbb{R}^d}\{u^T K u - \log|K| - \gamma\left\|u - X_i\right\|_q^\rho\} \\ &= \sup_{\Delta\in\mathbb{R}^d}\{(\Delta + X_i)^T K(\Delta + X_i) - \gamma\left\|\Delta\right\|_q^\rho\} - \log|K|. \end{aligned} \tag{5}$$

Replacing this expression back in (4), it may be difficult, if not impossible, to obtain a closed form optimization problem over K. Even if such a simplification is possible, it will not provide the desired connection to the graphical lasso estimator. That is why, in this paper, as outlined in section 2, we redefine the ambiguity set to obtain a desired closed form as expressed in Theorem 2.1 in a more transparent way.

B Proofs for the paper

B.1 Proof of Theorem 2.1

Proof. Consider $K \in \mathbb{S}_{\mathrm{d}}^{++}$. Observe that the function $l(\cdot; K) : \mathbb{S}_{\mathrm{d}} \to \mathbb{R}$ is Borel measurable since it is a continuous function. Then, we use the duality result for the DRO formulation from Proposition C.2 from the appendix of this paper and obtain

$$\sup_{P\colon\, \mathcal{D}_c(P,\mathbb{P}_n)\leq\delta} E_P\big[l(W;K)\big] = \inf_{\gamma\geq 0}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^{n}\left(\sup_{W\in\mathbb{S}_{\mathrm{d}}}\{l(W;K) - \gamma c(W,W_i)\}\right)\right\}. \tag{6}$$

Let $\Delta := W - W_i$. Then,

$$\begin{aligned}
&\sup_{W\in\mathbb{S}_{\mathrm{d}}}\{l(W;K) - \gamma c(W,W_i)\}\\
&\quad= \sup_{W\in\mathbb{S}_{\mathrm{d}}}\{\mathrm{trace}(KW) - \log|K| - \gamma \left\|\mathbf{vec}(W) - \mathbf{vec}(W_i)\right\|_q^{\rho}\}\\
&\quad= \sup_{\Delta\in\mathbb{S}_{\mathrm{d}}}\{\mathrm{trace}(K(\Delta + W_i)) - \gamma\left\|\mathbf{vec}(\Delta)\right\|_q^{\rho}\} - \log|K|\\
&\quad= \sup_{\Delta\in\mathbb{S}_{\mathrm{d}}}\{\mathrm{trace}(K\Delta) - \gamma\left\|\mathbf{vec}(\Delta)\right\|_q^{\rho}\} + \mathrm{trace}(KW_i) - \log|K|\\
&\quad= \sup_{\Delta\in\mathcal{M}(K)}\{\left\|\mathbf{vec}(\Delta)\right\|_q\left\|\mathbf{vec}(K)\right\|_p - \gamma\left\|\mathbf{vec}(\Delta)\right\|_q^{\rho}\} + \mathrm{trace}(KW_i) - \log|K|
\end{aligned} \tag{7}$$

with $\mathcal{M}(K) = \{\Delta \in \mathbb{S}_{\mathrm{d}} \mid \mathrm{trace}(K\Delta) > 0,\ |\Delta_{ij}|^q = \theta|k_{ij}|^p \text{ for some } \theta > 0\}$ so that the fourth line follows from selecting a $\Delta \in \mathbb{S}_{\mathrm{d}}$ (since $K \in \mathbb{S}_{\mathrm{d}}^{++}$) such that Holder's inequality holds tightly (with $\frac{1}{p} + \frac{1}{q} = 1$). In fact, Holder's inequality holds tightly if and only if $\Delta \in \mathcal{M}(K)$ (Steele, 2004, Chapter 9), even for the limiting case $q = \infty$, $p = 1$. Observe that there exist multiple $\Delta \in \mathbb{S}_{\mathrm{d}}$ that can satisfy Holder's inequality tightly. As a consequence, we are still free to choose the magnitude of the q-norm of such $\mathbf{vec}(\Delta)$ (and this is what we will use next).

Now, the argument inside the supremum in the last line of (7) is a polynomial function on $\left\|\mathbf{vec}(\Delta)\right\|_q$. We have to analyze two cases.

Case 1: $\rho = 1$. In this case we observe that, by setting $\epsilon(\gamma, K) = \sup_{\Delta\in\mathcal{M}(K)}\{\left\|\mathbf{vec}(\Delta)\right\|_q(\left\|\mathbf{vec}(K)\right\|_p - \gamma)\}$:

- if $\gamma \geq \left\|\mathbf{vec}(K)\right\|_p$, then $\epsilon(\gamma, K) = 0$ (in particular, if $\gamma = \left\|\mathbf{vec}(K)\right\|_p$, the optimizer is $\Delta = \mathbb{0}_{d\times d}$);
- if $\gamma < \left\|\mathbf{vec}(K)\right\|_p$, then $\epsilon(\gamma, K) = \infty$;

so that, recalling (6), due to the outside infimum to be taken over $\gamma \geq 0$; and so we must have that

$$\sup_{P\colon\, \mathcal{D}_c(P,\mathbb{P}_n)\leq\delta} E_P\big[l(W;K)\big] = \inf_{\gamma\geq\left\|\mathbf{vec}(K)\right\|_p}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^{n}(\mathrm{trace}(KW_i) - \log|K|)\right\}$$

from which we immediately obtain (8).

Case 2: $\rho > 1$. By differentiation and basic calculus (e.g., using the first and second derivative test) we obtain that the maximizer

$$\Delta^* = \arg\sup_{\Delta\in\mathcal{M}(K)}\{\left\|\mathbf{vec}(\Delta)\right\|_q\left\|\mathbf{vec}(K)\right\|_p - \gamma\left\|\mathbf{vec}(\Delta)\right\|_q^{\rho}\}$$

is such that $\|\mathbf{vec}(\Delta^*)\|_q = \left(\frac{\|\mathbf{vec}(K)\|_p}{\gamma\rho}\right)^{\frac{1}{\rho-1}}$. Then,

$$\begin{aligned}\sup_{W \in \mathbb{S}_{\mathrm{d}}} \{l(W;K) - \gamma c(W, W_i)\} &= \frac{\|\mathbf{vec}(K)\|_p^{\frac{\rho}{\rho-1}}}{(\gamma\rho)^{\frac{1}{\rho-1}}} - \gamma\left(\frac{\|\mathbf{vec}(K)\|_p}{\gamma\rho}\right)^{\frac{\rho}{\rho-1}} + \mathrm{trace}(KW_i) \\ &\quad - \log|K| \\ &= \|\mathbf{vec}(K)\|_p^{\frac{\rho}{\rho-1}} \frac{\rho-1}{\rho^{\frac{\rho}{\rho-1}}\gamma^{\frac{1}{\rho-1}}} + \mathrm{trace}(KW_i) - \log|K|.\end{aligned} \tag{8}$$

Replacing this back in (6),

$$\begin{aligned}&\sup_{P\colon \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P\big[l(W;K)\big] \\ &\qquad= \inf_{\gamma\ge0}\left\{\gamma\delta + \frac{1}{n}\sum_{i=1}^{n}\left(\|\mathbf{vec}(K)\|_p^{\frac{\rho}{\rho-1}} \frac{\rho-1}{\rho^{\frac{\rho}{\rho-1}}\gamma^{\frac{1}{\rho-1}}} + \mathrm{trace}(KW_i) - \log|K|\right)\right\} \\ &\qquad= \inf_{\gamma\ge0}\left\{\gamma\delta + \|\mathbf{vec}(K)\|_p^{\frac{\rho}{\rho-1}} \frac{\rho-1}{\rho^{\frac{\rho}{\rho-1}}\gamma^{\frac{1}{\rho-1}}} + \mathrm{trace}\left(K\frac{1}{n}\sum_{i=1}^{n}W_i\right) - \log|K|\right\} \\ &\qquad= \inf_{\gamma\ge0}\left\{\gamma\delta + \|\mathbf{vec}(K)\|_p^{\frac{\rho}{\rho-1}} \frac{\rho-1}{\rho^{\frac{\rho}{\rho-1}}\gamma^{\frac{1}{\rho-1}}}\right\} + \mathrm{trace}(KA_n) - \log|K|.\end{aligned} \tag{9}$$

Now, we observe that the argument inside the infimum in the last line of (9) is a function that grows to infinity when $\gamma \to 0$ or $\gamma \to \infty$, so that the minimum is attained for some optimal γ. By using the first and second derivative tests, we obtain that the minimizer is $\gamma^* = \frac{\|\mathbf{vec}(K)\|_p}{\rho\delta^{\frac{\rho-1}{\rho}}}$. Then, replacing this back in (9) and then this in (6), we finally obtain (8) after some algebraic simplification. □

B.2 Proof of Lemma 3.1

Proof. Consider $K \in \mathbb{S}_{\mathrm{d}}^{++}$ and define $g(K) = \sup_{P\colon \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P\left[l\big(W;K\big)\right]$ for a fixed δ. We prove (12), by a direct application of Proposition 8 of (Blanchet et al., 2016, Appendix C), observing that we satisfy the three conditions for its application:

(i) g is convex on $\mathbb{S}_{\mathrm{d}}^{++}$ and finite,

(ii) there exists $b \in \mathbb{R}$ such that the sublevel set $\kappa_b = \{K \mid g(K) \le b\}$ is compact and non-empty,

(iii) $E_P[l(W;K)]$ is lower semi-continuous and convex as a function of K throughout κ_b for any $P \in \{P \mid \mathcal{D}_c(P,\mathbb{P}_n) \le \delta\}$.

First, we observe that

$$E_{\mathbb{P}_0}[l(W,K)] = E_{\mathbb{P}_0}[\mathrm{trace}(KW) - \log|K|] \le \mathrm{trace}(KE_{\mathbb{P}_0}[W]) < \infty,$$

since $E_{\mathbb{Q}_0}(\|X\|_2^2) < \infty$. Then, using Theorem 2.1, the function

$$g(K) = \sup_{P\colon \mathcal{D}_c(P,\mathbb{P}_n)\le\delta} E_P\left[l\big(W;K\big)\right] = \mathrm{trace}(KA_n) - \log|K| + \delta^{1/\rho}\,\|\mathbf{vec}(K)\|_p$$

is finite. Moreover, we also claim it is convex for all $K \in \mathbb{S}_{\mathrm{d}}^{++}$. This follows from the fact that $\mathrm{trace}(KA_n) - \log|K|$ and $\|\mathbf{vec}(K)\|_p$, $p \in [1,\infty]$ are two convex functions on $K \in \mathbb{S}_{\mathrm{d}}^{++}$, and from the fact that the nonnegative weighted sum of two convex functions is another convex function (Boyd and Vandenberghe, 2004). This proves (i).

Now, we claim that $g(K)$ is radially unbounded, i.e., $g(K) \to \infty$ as $\|\mathbf{vec}(K)\|_p \to \infty$. To see this, recall that $\text{trace}(KA_n) - \log|K|$ is a differentiable convex function in K that is minimized whenever $K^{-1} = A_n$, since A_n is invertible for $n > d$ almost surely. Then,

$$g(K) = \text{trace}(KA_n) - \log|K| + \delta^{1/\rho}\|\mathbf{vec}(K)\|_p \geq d - \log|A_n^{-1}| + \delta^{1/\rho}\|\mathbf{vec}(K)\|_p,$$

from which it immediately follows that $g(K) \to \infty$ as $\|\mathbf{vec}(K)\|_p \to \infty$.

Now, again, since $g(K)$ is also convex and continuous in $\mathbb{S}_\mathrm{d}^{++}$, we conclude that the level sets $\kappa_b = \{K \mid g(K) \leq b\}$ are compact and nonempty as long as $b > l(W; K) + \delta^{1/\rho}\|\mathbf{vec}(K)\|_p$. This proves (ii). Moreover, since $l(W; K)$ is convex and continuous on any $K \in \mathbb{S}_\mathrm{d}^{++}$, it follows that $E_P[l(W; K)]$ for any $P \in \{P \mid \mathcal{D}_c(P, \mathbb{P}_n) \leq \delta\}$ is also continuous and convex on any $K \in \mathbb{S}_\mathrm{d}^{++}$, thus (iii) follows immediately. □

B.3 Proof of Theorem 3.2

Proof. Consider $K \in \mathbb{S}_\mathrm{d}^{++}$. Setting $h(U; K) = U - K^{-1}$, it is clear that we satisfy the conditions for applying Proposition C.1 in the Appendix, and so we obtain

$$R_n(K) = \sup_{\Lambda \in \mathbb{S}_\mathrm{d}} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \sup_{U \in \mathbb{S}_\mathrm{d}} \{\text{trace}(\Lambda^\top(U - K^{-1})) - \|\mathbf{vec}(U) - \mathbf{vec}(W_i)\|_q^\rho\} \right\} \tag{10}$$

Now, letting $\Delta := U - W_i^\top$

$$\begin{aligned}
\sup_{U \in \mathbb{S}_\mathrm{d}} \{\text{trace}(\Lambda^\top(U - K^{-1})) - &\|\mathbf{vec}(U) - \mathbf{vec}(W_i)\|_q^\rho\} \\
&= \sup_{\Delta \in \mathbb{S}_\mathrm{d}} \{\text{trace}(\Lambda^\top(\Delta + W_i - K^{-1})) - \|\mathbf{vec}(\Delta)\|_q^\rho\} \\
&= \sup_{\Delta \in \mathbb{S}_\mathrm{d}} \{\text{trace}(\Lambda^\top\Delta) - \|\mathbf{vec}(\Delta)\|_q^\rho\} + \text{trace}(\Lambda^\top(W_i - K^{-1})) \qquad (11)\\
&= \sup_{\Delta \in \mathcal{M}(\Lambda)} \{\|\mathbf{vec}(\Lambda)\|_p \|\mathbf{vec}(\Delta)\|_q - \|\mathbf{vec}(\Delta)\|_q^\rho\} \\
&\quad + \text{trace}(\Lambda^\top(W_i - K^{-1}))
\end{aligned}$$

with $\mathcal{M}(\Lambda)$ as in the proof of Theorem 2.1, so that the third line follows from selecting a $\Delta \in \mathbb{S}_\mathrm{d}$ such that Holder's inequality holds tightly (with $\frac{1}{p} + \frac{1}{q} = 1$), whose existence has been explained in the proof of Theorem 2.1. Thus, we are still free to choose the magnitude of the q-norm of such $\mathbf{vec}(\Delta)$ (and this is what we will use next).

Now, the argument inside the supremum in the last line of (11) is a polynomial function on $\|\mathbf{vec}(\Delta)\|_q$. We have to analyze two cases.

Case 1: $\rho = 1$. In this case we observe that, by setting $\epsilon(\Lambda) = \sup_{\Delta \in \mathbb{S}_\mathrm{d}}\{\|\mathbf{vec}(\Delta)\|_q (\|\mathbf{vec}(\Lambda)\|_p - 1)\}$:

- if $\|\mathbf{vec}(\Lambda)\|_p \leq 1$, then $\epsilon(\Lambda) = 0$ (in particular, if $\|\mathbf{vec}(\Lambda)\|_p < 1$, the optimizer is $\Delta = \mathbb{0}_{d \times d}$);
- if $\|\mathbf{vec}(\Lambda)\|_p > 1$, then $\epsilon(\Lambda) = \infty$;

so that, recalling (10), we see that if $\|\mathbf{vec}(\Lambda)\|_p > 1$, then $R_n(K) = -\infty$. Then, we obtain that

$$\begin{aligned}
R_n(K) &= \sup_{\Lambda \in \mathbb{S}_\mathrm{d} : \|\mathbf{vec}(\Lambda)\|_p \leq 1} \left\{ -\frac{1}{n}\sum_{i=1}^{n} \text{trace}(\Lambda^\top(W_i - K^{-1}) \right\} \\
&= \sup_{\Lambda \in \mathbb{S}_\mathrm{d} : \|\mathbf{vec}(\Lambda)\|_p \leq 1} \left\{ -\text{trace}(\Lambda^\top(A_n - K^{-1}) \right\} \\
&= \sup_{\Lambda \in \mathbb{S}_\mathrm{d} : \|\mathbf{vec}(\Lambda)\|_p \leq 1} \left\{ \mathbf{vec}(\Lambda)^\top \mathbf{vec}(A_n - K^{-1}) \right\} \\
&= \left\|\mathbf{vec}(A_n - K^{-1})\right\|_q
\end{aligned}$$

where the third line results from the fact that Λ is a free variable so we can flip its sign, and the last line follows from the the analysis of conjugate norms and the fact that $\Lambda, A_n - K^{-1} \in \mathbb{S}_{\mathrm{d}}$. We thus obtained (16).

Case 2: $\rho > 1$. By differentiation and basic calculus (e.g., using the first and second derivative test) we obtain that the maximizer

$$\Delta^* = \arg \sup_{\Delta \in \mathbb{S}_{\mathrm{d}}} \left\{ \left\|\mathbf{vec}(\Delta)\right\|_q \left\|\mathbf{vec}(\Lambda)\right\|_p - \gamma \left\|\mathbf{vec}(\Delta)\right\|_q^\rho \right\}$$

is such that $\left\|\mathbf{vec}(\Delta^*)\right\|_q = \left(\frac{\|\mathbf{vec}(K)\|_p}{\rho}\right)^{\frac{1}{\rho-1}}$. Then, replacing this back in (10),

$$\begin{aligned} R_n(K) &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left(\left\|\mathbf{vec}(\Lambda)\right\|_p^{\frac{\rho}{\rho-1}} \frac{\rho - 1}{\rho^{\frac{\rho}{\rho-1}}} + \operatorname{trace}(\Lambda^\top (W_i - K^{-1})) \right) \right\} \\ &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ - \left\|\mathbf{vec}(\Lambda)\right\|_p^{\frac{\rho}{\rho-1}} \frac{\rho - 1}{\rho^{\frac{\rho}{\rho-1}}} - \operatorname{trace}(\Lambda^\top (A_n - K^{-1})) \right\} \\ &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ \operatorname{trace}(\Lambda^\top (A_n - K^{-1})) - \left\|\mathbf{vec}(\Lambda)\right\|_p^{\frac{\rho}{\rho-1}} \frac{\rho - 1}{\rho^{\frac{\rho}{\rho-1}}} \right\} \\ &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ \left\|\mathbf{vec}(\Lambda)\right\|_p \left\|\mathbf{vec}(A_n - K^{-1})\right\|_q - \left\|\mathbf{vec}(\Lambda)\right\|_p^{\frac{\rho}{\rho-1}} \frac{\rho - 1}{\rho^{\frac{\rho}{\rho-1}}} \right\}. \end{aligned}$$

Again, by differentiation and basic calculus, we obtain that the maximizer Λ^* is such that $\left\|\mathbf{vec}(\Lambda^*)\right\|_p = \rho \left\|\mathbf{vec}(A_n - K^{-1})\right\|_q^{\rho-1}$. Replacing this value back in our previous expression, we get that $R_n(K) = \left\|\mathbf{vec}(A_n - K^{-1})\right\|_q^{\rho}$, and thus we showed (16). □

C Applicability of the dual representations of the RWP function and the DRO formulation

The dual representations of the RWP function and the DRO formulation for the case in which the space of probability measures is $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is studied in the paper (Blanchet et al., 2016). In this paper, we are interested in the case $\mathcal{P}(\mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}})$. In other words, we consider the samples to be in $\mathbb{S}_{\mathrm{d}}$ instead of $\mathbb{R}^d$. We want to emphasize that the derivations of these dual representations rely on the dual formulation of the so called "problem of moments" or a specific class of "Chebyshev-type inequalities" referenced in the work by Isii (1962). The derivation by Isii is actually more general in the sense that is applied to more general probability spaces than the ones used in this paper and in (Blanchet et al., 2016) (in fact, it is stated for general spaces of non-negative measures).

Throughout this section, we consider an integrable function $h : \mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}} \to \mathbb{S}_{\mathrm{d}}$, and a lower semi-continuous function $c : \mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}} \to [0, \infty)$ such that $c(U, U) = 0$ for any $U \in \mathbb{S}_{\mathrm{d}}$ and such that the set

$$\Omega := \left\{ (U', W') \in \mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}} \mid c(U', W') < \infty \right\}$$

is Borel measurable and non-empty. Also consider an iid random sample $W_1, \cdots, W_n \sim W$ with W coming from a distribution on $\mathcal{P}(\mathbb{S}_{\mathrm{d}})$.

Now, let us focus first on the RWP function in the following proposition which parallels (Blanchet et al., 2016, Proposition 3).

Proposition C.1. *Consider $K \in \mathbb{S}_{\mathrm{d}}^{++}$. Let $h(\cdot, K)$ be Borel measurable. Also, suppose that $\mathbb{0}_{d\times d}$ lies in the interior of the convex hull of $\{h(U', C) \mid U' \in \mathbb{S}_{\mathrm{d}}\}$. Then,*

$$R_n(K) = \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{U \in \mathbb{S}_{\mathrm{d}}} \{\mathrm{trace}(\Lambda^\top h(U; K)) - c(U, W_i)\} \right\}.$$

Proof. Consider the proof of (Blanchet et al., 2016, Proposition 3). If we:

- set the estimating equation by $E[h(W; K)] = \mathbb{0}_{d\times d}$,
- set
$$R_n(K) = \inf \{E_\pi[c(U, W)] \mid E_\pi[h(U; K)] = \mathbb{0}_{n\times n}, \pi_W = \mathbb{P}_n, \pi \in \mathcal{P}(\mathbb{S}_{\mathrm{d}} \times \mathbb{S}_{\mathrm{d}})\},$$
with π_W denoting the marginal distribution of W,
- consider the previously defined Ω,

then we obtain that, following the rest of this proof (and using (Isii, 1962, Theorem 1) with its special case):

$$\begin{aligned} R_n(K) &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{U \in \mathbb{S}_{\mathrm{d}}} \{\mathbf{vec}(\Lambda)^\top \mathbf{vec}(h(U; K)) - c(U, W_i)\} \right\} \\ &= \sup_{\Lambda \in \mathbb{S}_{\mathrm{d}}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{U \in \mathbb{S}_{\mathrm{d}}} \{\mathrm{trace}(\Lambda^\top h(U; K)) - c(U, W_i)\} \right\}, \end{aligned}$$

thus obtaining the dual representation of the RWP function. □

The following proposition for the dual representation of the DRO formulation parallels (Blanchet et al., 2016, Proposition 1).

Proposition C.2. *For $\gamma \geq 0$ and loss functions $l(U'; K)$ that are upper semi-continuous in $U' \in \mathbb{S}_{\mathrm{d}}$ for each $K \in \mathbb{S}_{\mathrm{d}}^{++}$, let*

$$\phi_\gamma(W_i; K) = \sup_{U \in \mathbb{S}_{\mathrm{d}}} \{l(U; K) - \gamma c(U, W_i))\}. \tag{12}$$

Then

$$\sup_{P:\ \mathcal{D}_c(P, \mathbb{P}_n) \leq \delta} E_P\big[l(W; K)\big] = \min_{\gamma \geq 0} \left\{ \gamma\delta + \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(W_i; K) \right\}.$$

Proof. The proof for the dual representation of the DRO for our domain of symmetric matrices is also very similar to the one described in Proposition 4 of (Blanchet et al., 2016, version 2), just by following appropriate similar changes as we did for the proof of C.1. □

D Figures, tables and additional analysis for the numerical results (Section 4 of the paper)

D.1 Matthews correlation coefficient analysis

Let true positives (TP) be the number of nonzero off-diagonal entries of Ω that are correctly identified, false negatives (FN) be the number of its nonzero off-diagonal entries that are incorrectly identified as zeros, false positives (FP) be the number of its zero off-diagonal entries that are incorrectly identified as nonzeros, and true negatives (TN) be the number of its zero off-diagonal entries that are correctly identified. Given the estimated precision matrix $\hat{K}$, the Matthews correlation coefficient (MCC) (Powers, 2011) is defined as:

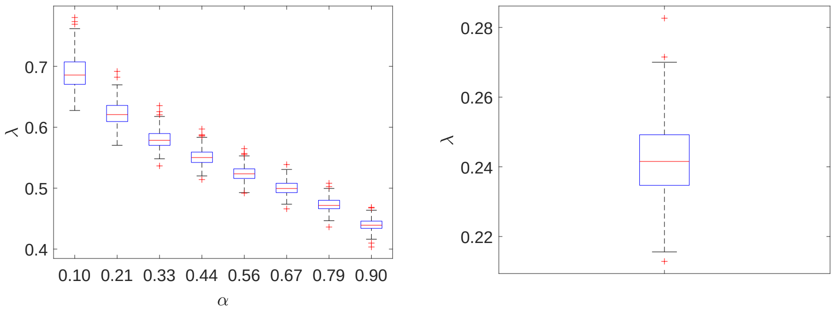
$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FP)}}, \tag{13}$$

and, whenever the denominator is zero, it can be arbitrarily set to one. It can be shown that $MCC \in [-1, +1]$, and $+1$ is interpreted as a perfect prediction (of both zero and nonzero values), 0 is interpreted as prediction no better than a random one, and -1 is interpreted as indicating total disagreement between prediction and observation.

MCC has been argued to be one of the most informative coefficients for assessing the performance of binary classification (in this case, classifying if an entry of the precision matrix is zero or nonzero) since it summarizes all information from the TP, TN, FP and FN quantities (Chicco, 2017; Powers, 2011), in contrast to other measures like TPR and FPR.

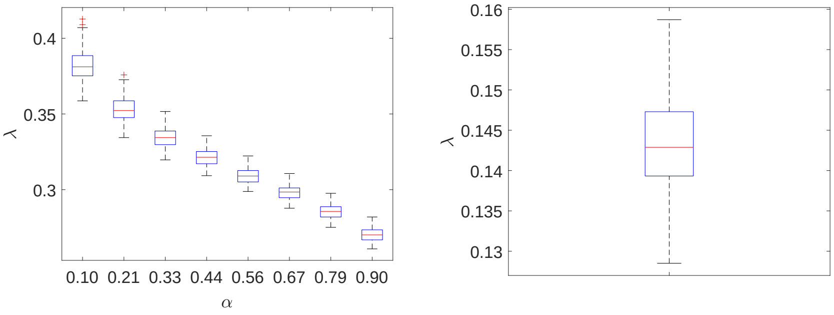
D.2 Regularization parameters

All the plots related to the RS criterion for choosing λ have in their x-axis the values
$\alpha \in \{0.10, 0.21, 0.33, 0.44, 0.56, 0.67, 0.79, 0.90\}$. We study the cases for sample sizes $n \in \{75, 200, 1000\}$.



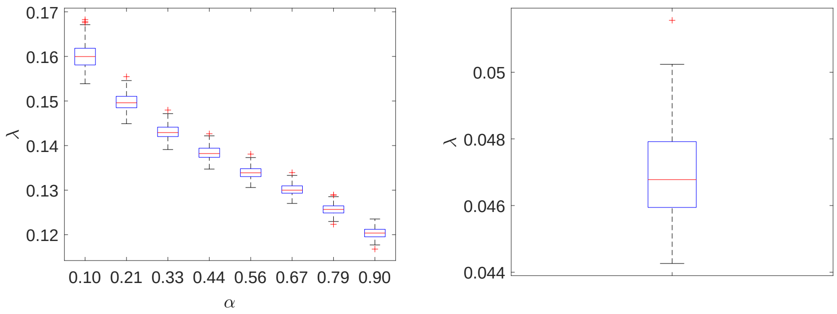
(a) RS (b) CV

Figure 1: $n = 75$



(a) RS (b) CV

Figure 2: $n = 200$



(a) RS (b) CV

Figure 3: $n = 1000$

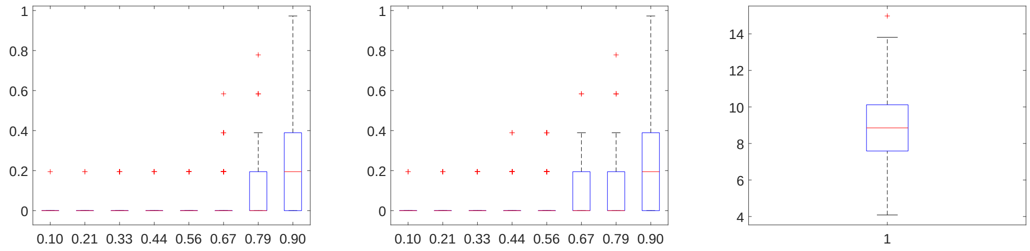
D.3 Performance figures for different choices of the regularization parameter

All the plots related to the RS and RWP criteria for choosing λ have in their x-axis the values $\alpha \in \{0.10, 0.21, 0.33, 0.44, 0.56, 0.67, 0.79, 0.90\}$. We study the cases for sample sizes $n \in \{75, 200, 1000\}$.

In this supplemental section, we introduce two additional performance criteria, each one having the intuitive interpretation of measuring the closeness of the estimated inverse covariance matrix $\hat{K}_\lambda$ and the true precision matrix Ω:

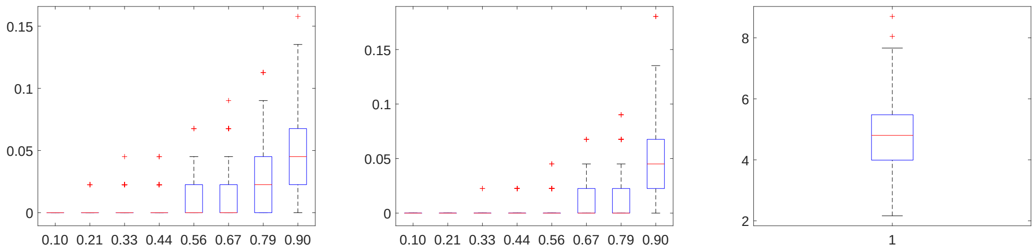
1. *Stein's loss*: $\text{trace}(\Omega^{-1}\hat{K}_\lambda) - \log|\Omega^{-1}\hat{K}_\lambda| - d$, which is zero whenever $\hat{K}_\lambda = \Omega$.
2. *Difference of inverse covariances*: $\frac{\|\hat{K}_\lambda - \Omega\|_F}{d}$, with $\|\cdot\|_F$ being the Frobenius norm.

D.3.1 $n = 75$



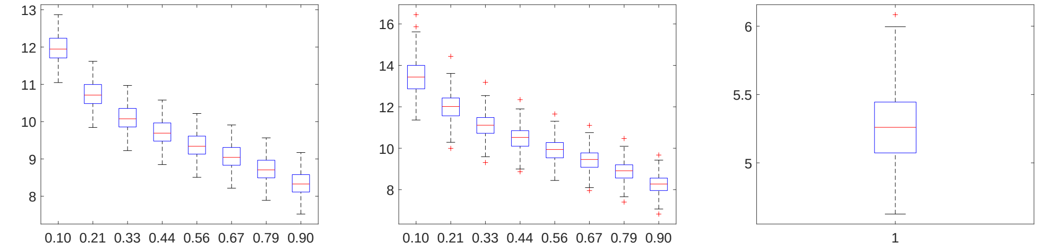
(a) RWP (b) RS (c) CV

Figure 4: True positive rate (%)



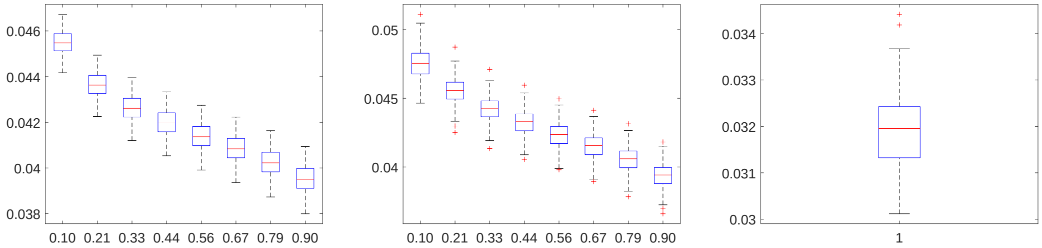
(a) RWP (b) RS (c) CV

Figure 5: False detection rate (%)



(a) RWP (b) RS (c) CV

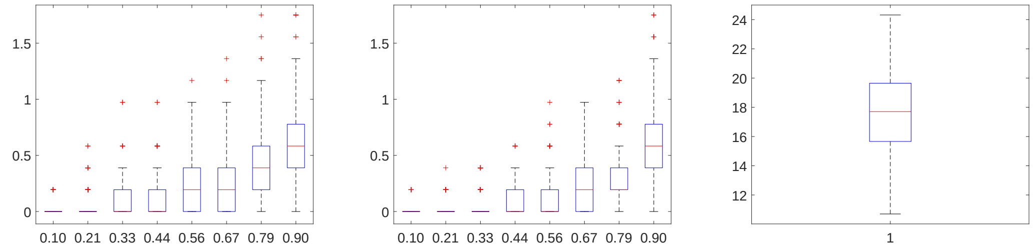
Figure 6: Stein's loss



(a) RWP (b) RS (c) CV

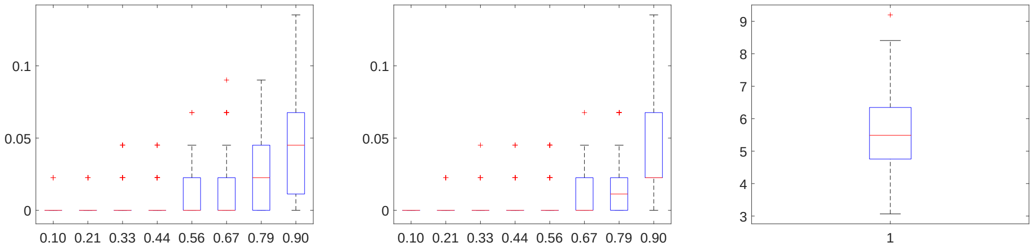
Figure 7: Difference of inverse covariances

D.3.2 $n = 200$



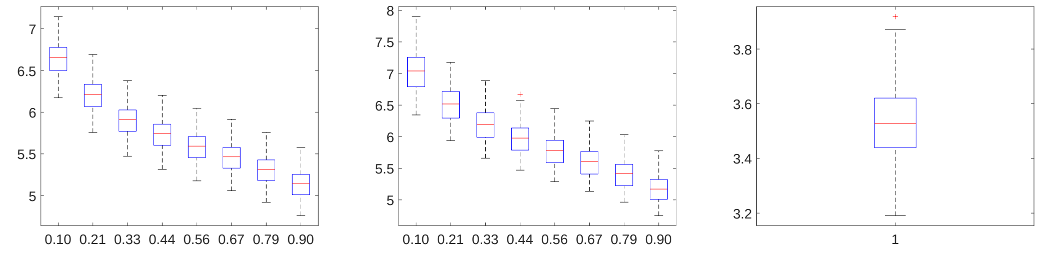
(a) RWP (b) RS (c) CV

Figure 8: True positive rate (%)



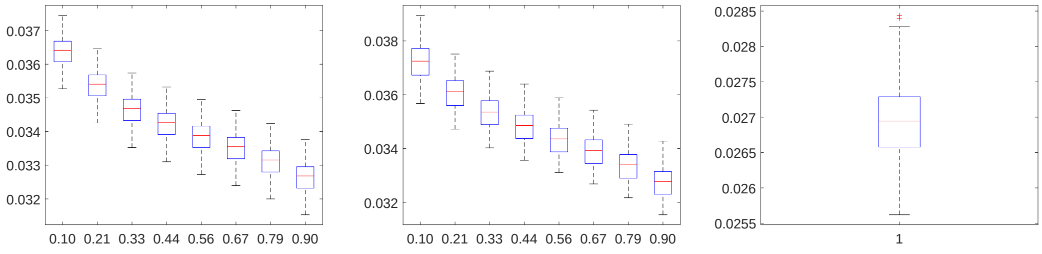
(a) RWP (b) RS (c) CV

Figure 9: False detection rate (%)



(a) RWP (b) RS (c) CV

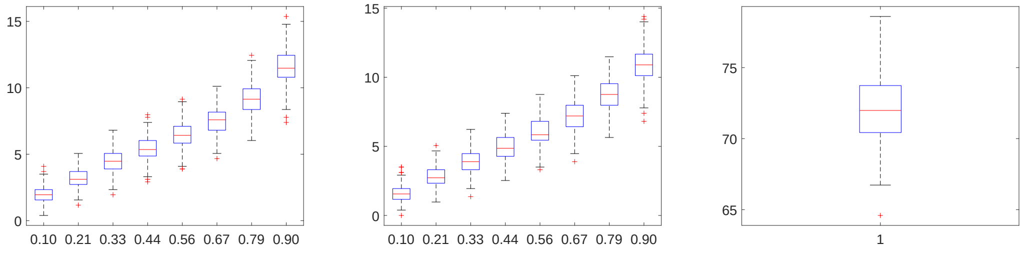
Figure 10: Stein's loss



(a) RWP (b) RS (c) CV

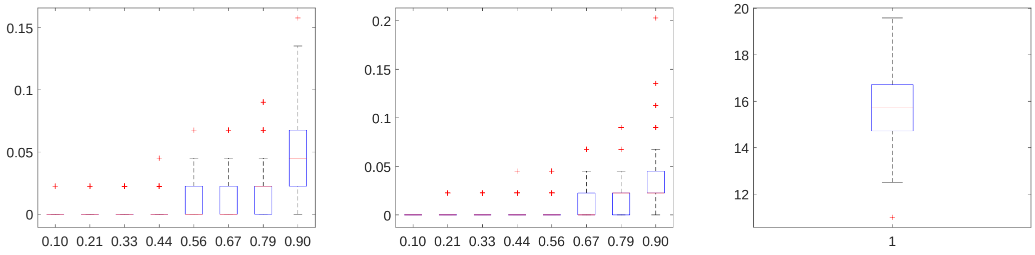
Figure 11: Difference of inverse covariances

D.3.3 $n = 1000$



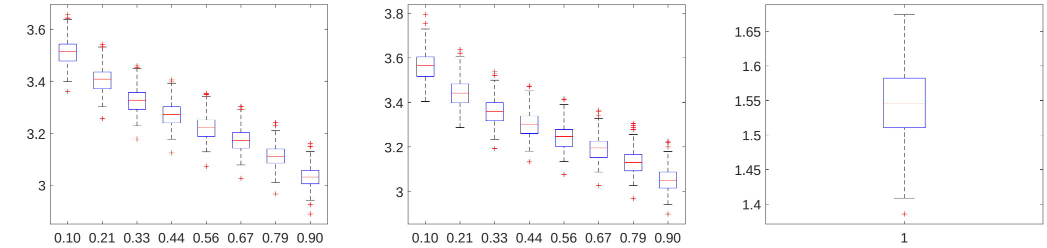
(a) RWP (b) RS (c) CV

Figure 12: True positive rate (%)



(a) RWP (b) RS (c) CV

Figure 13: False detection rate (%)

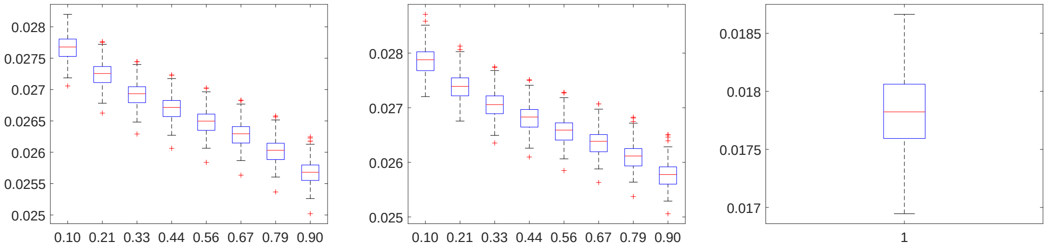


(a) RWP

(b) RS

(c) CV

Figure 14: Stein's loss



(a) RWP

(b) RS

(c) CV

Figure 15: Difference of inverse covariances

References

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.

D. Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, Dec 2017. doi:10.1186/s13040-017-0155-3.

K. Isii. On sharpness of tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962. doi:10.1007/BF02868641.

D. M. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2:37–63, 2011.

J. M. Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004. doi:10.1017/CBO9780511817106.