Distributionally Robust Stochastic Optimization with Dependence Structure

Rui Gao · Anton J. Kleywegt

the date of receipt and acceptance should be inserted later

Abstract Distributionally robust stochastic optimization (DRSO) is a framework for decision-making problems under certainty, which finds solutions that perform well for a chosen set of probability distributions. Many different approaches for specifying a set of distributions have been proposed. The choice matters, because it affects the results, and the relative performance of different choices depend on the characteristics of the problems. In this paper, we consider problems in which different random variables exhibit some form of dependence, but the exact values of the parameters that represent the dependence are not known. We consider various sets of distributions that incorporate the dependence structure, and we study the corresponding DRSO problems.

In the first part of the paper, we consider problems with linear dependence between random variables. We consider sets of distributions that are within a specified Wasserstein distance of a nominal distribution, and that satisfy a second-order moment constraint. We obtain a tractable dual reformulation of the corresponding DRSO problem. This approach is compared with the traditional moment-based DRSO, which considers all distributions whose first- and second-order moments satisfy certain constraints, and with the Wasserstein-based DRSO, which considers all distributions that are within a specified Wasserstein distance of a nominal distribution (with no moment constraints). Numerical experiments suggest that our new formulation has superior out-of-sample performance.

In the second part of the paper, we consider problems with various types of rank dependence between random variables, including rank dependence measured by Spearman's footrule distance between empirical rankings, comonotonic distributions, box uncertainty for individual observations, and Wasserstein distance between copulas associated with continuous distributions. We also obtain a dual reformulation of the DRSO problem. A desirable byproduct of the formulation is that it also avoids an issue associated with the one-sided moment constraints in moment-based DRSO problems.

Keywords Distributionally robust optimization · Data-driven · Copula · Portfolio optimization

Mathematics Subject Classification (2010) 90C15 · 91G10

1 Introduction

Stochastic optimization is an approach to optimization under uncertainty with a well-developed foundation of theory and practical applications. A core issue in stochastic optimization is that often the underlying probability distribution is not known, or the notion of multiple realizations from a single underlying probability distribution may be a questionable description of reality. Distributionally robust stochastic optimization (DRSO) is an approach to optimization under uncertainty in which, instead of assuming that there is an underlying probability distribution that is known to the optimizer, one finds a decision $x \in X$ that provides the best hedge against a set of probability distributions, by solving the

Rui Gao
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA
E-mail: rgao32@gatech.edu

Anton J. Kleywegt
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA
E-mail: anton@isye.gatech.edu

following problem:

$$\inf_{x\in X}\sup_{\boldsymbol{\mu}\in\mathfrak{M}}\mathbb{E}_{\boldsymbol{\mu}}[\Psi(x,\boldsymbol{\xi})], \tag{1}$$

where the cost function $\Psi : X \times \Xi \mapsto \mathbb{R}$ depends on a random quantity $\boldsymbol{\xi}$ which takes values in $\Xi \subset \mathbb{R}^K$, and $\mathfrak{M}$ is a subset of the set $\mathcal{P}(\Xi)$ of all Borel probability distributions on Ξ.

The set $\mathfrak{M}$ of probability distributions is chosen to make the resulting decisions robust against future variations in $\boldsymbol{\xi}$. Two approaches to choose the set $\mathfrak{M}$ have been studied in some depth. The moment-based approach considers distributions whose moments (such as mean and covariance) satisfy certain conditions [21,19,4,30]. The distance-based approach considers distributions that are close, in the sense of a chosen distance, to a nominal distribution $\boldsymbol{\nu}$, such as an empirical distribution. Popular choices of the distances are ϕ-divergences [3,2], which include Kullback-Leibler divergence [14], Burg entropy [27], and total variation distance [26] as special cases, Prokhorov metric [9], and Wasserstein metric [10,12].

In some practical settings, the decision maker may be aware that the random variables exhibit some dependence structure, even if the parameter values that specify the dependence are not known. For example, the decision maker may be aware of an approximate linear dependence, measured by Pearson's product-moment correlation coefficient [11,18], or some form of rank dependence, such as measured by Spearman's ρ [25] or Kendall's τ [16]. In this paper, we are interested in DRSO problems that take into account one of these types of dependence structure.

1.1 Linear-correlationally robust stochastic optimization

We first consider linear dependence. For instance, the moment uncertainty set in [4] is defined by

$$\begin{aligned}\mathfrak{M} \;:=\; \Big\{\boldsymbol{\mu}\in\mathcal{P}(\Xi)\;:&\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)^\top\Sigma_0^{-1}(\boldsymbol{\xi}-m_0)]\;\le\;\gamma_0,\\ &\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top]\;\preceq\;\Sigma_0\Big\}\end{aligned} \tag{2}$$

where m_0 is a specified "center" vector (such as a sample mean), $\gamma_0 \ge 0$ can be viewed as the squared radius of the confidence region of the mean vector, and $\Sigma_0 \succeq 0$ is often chosen to be the sample covariance matrix $\hat{\Sigma}$ inflated by some constant $\gamma \ge 1$. The resulting moment-based DRSO problem is given by

$$\begin{aligned}\inf_{x\in X}\sup_{\boldsymbol{\mu}\in\mathcal{P}(\Xi)}\Big\{\mathbb{E}_{\boldsymbol{\mu}}[\Psi(x,\boldsymbol{\xi})]\;:&\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)^\top\Sigma_0^{-1}(\boldsymbol{\xi}-m_0)]\;\le\;\gamma_0,\\ &\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top]\;\preceq\;\Sigma_0\Big\}\end{aligned} \tag{3}$$

Similar to other moment-based approaches, this approach is based on the curious assumption that certain conditions on the moments are known but that nothing else about the relevant distribution is known. More often in applications, either one has data from repeated observations of the quantity $\boldsymbol{\xi}$, or one has no data, and in both cases the moment conditions do not describe exactly what is known about $\boldsymbol{\xi}$ or its distribution. Moreover, such sets $\mathfrak{M}$ often lead to unrealistic worst-case distributions which make the resulting decisions x overly conservative [27,13]. Another characteristic of worst-case distributions in the moment uncertainty set in (2) is given below.

Example 1 (Degeneracy of moment uncertainty set) Suppose that $\gamma_0 = 0$, and that Ψ satisfies either of the following conditions:

(i) For given x, $\Psi(x,\cdot)$ is a concave function whose domain contains m_0.
(ii) For given x, $\Psi(x,\cdot)$ is an indicator function of a set $A(x) \subset \Xi$ which contains m_0.

Then for the given x, $\boldsymbol{\delta}_{m_0}$ is a worst-case distribution, independent of the value of Σ_0. □

Therefore, under these conditions, the second-order moment constraint $\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top] \preceq \Sigma_0$ has no effect on the problem.

To overcome some of the drawbacks of formulation (3), we consider sets $\mathfrak{M}_1$ of distributions that combine the second-order moment constraint and the Wasserstein distance constraint, as follows:

$$\mathfrak{M}_1 \;:=\; \left\{\boldsymbol{\mu}\in\mathcal{P}(\Xi)\;:\;W_p(\boldsymbol{\mu},\boldsymbol{\nu})\;\le\;R_0,\;\mathbb{E}_{\boldsymbol{\mu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top]\;\preceq\;\Sigma_0\right\}, \tag{4}$$

where $R_0 > 0$, $m_0 \in \mathbb{R}^K$, $\Sigma_0 \succeq 0$, $\boldsymbol{\nu}$ is some nominal distribution, and $W_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the Wasserstein metric of order $p \geq 1$ between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, defined as

$$\begin{aligned} & W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \\ := & \min_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)} \left\{ \int_{\Xi^2} d^p(\xi, \zeta) \boldsymbol{\gamma}(\xi, \zeta) \ : \ \boldsymbol{\gamma} \text{ has marginal distributions } \boldsymbol{\mu}, \boldsymbol{\nu} \right\}, \end{aligned}$$

where d denotes a chosen metric on Ξ. The choices of m_0 and Σ_0 are similar to the moment-based approach. The set $\mathfrak{M}_1$ contains all distributions that are close to the nominal distribution in terms of the Wasserstein metric and that satisfy the linear correlation structure expressed in terms of the centered second-order moment constraint. It has been shown recently that DRSO with Wasserstein metrics has advantages over DRSO with ϕ-divergences [10,12].

Given $\mathfrak{M}_1$, we define the following linear-correlationally robust stochastic optimization problem:

$$\inf_{x \in X} \sup_{\boldsymbol{\mu} \in \mathfrak{M}_1} \mathbb{E}_{\boldsymbol{\mu}}[\Psi(x, \boldsymbol{\xi})]. \tag{Linear-CRSO}$$

In the first part of the paper we derive a tractable reformulation of this problem, and we examine its performance.

1.2 Rank-correlationally robust stochastic optimization

In the second part of the paper we consider a different dependence structure, motivated by notions of rank dependence. First, we give its definition, and then we illustrate its modeling flexibility via some examples. Given the nominal distribution $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, let $F_k : \mathbb{R} \mapsto [0,1]$ denote its k-th marginal cumulative distribution function, $k = 1, \ldots, K$, and let $F : \mathbb{R}^K \mapsto [0,1]^K$ be given by

$$F(\zeta) \ := \ (F_1(\zeta_1), \ldots, F_K(\zeta_K)).$$

Note that F is the vector of marginal cumulative distribution functions, and not the joint cumulative distribution function of $\boldsymbol{\nu}$. Let d be a metric on $[0,1]^K$. We define a semimetric d_F on Ξ by

$$\mathsf{d}_F(\xi, \zeta) \ := \ \liminf_{\xi^n \to \xi, \ \zeta^n \to \zeta} \mathsf{d}\left(F(\xi^n), F(\zeta^n)\right)$$

Note that d_F is lower semicontinuous. Let $q \in [1, \infty]$. For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\Xi)$, we define

$$\begin{aligned} & \mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \ := \\ & \begin{cases} \left(\inf_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)} \left\{ \int_{\Xi^2} \mathsf{d}_F^q(\xi, \zeta) \boldsymbol{\gamma}(d\xi, d\zeta) \ : \ \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mu}, \boldsymbol{\nu} \right\} \right)^{1/q}, & \text{if } 1 \leq q < \infty, \\ \inf_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)} \left\{ \boldsymbol{\gamma}\text{-ess sup}_{(\xi,\zeta) \in \Xi^2} \ \mathsf{d}_F(\xi, \zeta) \ : \ \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mu}, \boldsymbol{\nu} \right\}, & \text{if } q = \infty. \end{cases} \end{aligned} \tag{5}$$

This is a transport metric that generalizes the Wasserstein metric in which the transportation cost is given by $\mathsf{d}_F(\cdot, \cdot)$.

We consider the set $\mathfrak{M}_2$ of probability distributions, where

$$\mathfrak{M}_2 \ := \ \left\{ \boldsymbol{\mu} \in \mathcal{P}(\Xi) \ : \ W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq R_0, \ \mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq r_0 \right\}, \tag{6}$$

where $R_0, r_0 > 0$. The constraints suggest that $\boldsymbol{\mu}$ is close to $\boldsymbol{\nu}$ in the sense of both Wasserstein metric W_p and transport metric W_q. Next we provide some examples to illustrate the meaning of different versions of W_q.

Example 2 (Empirical ranking) Suppose $\boldsymbol{\nu} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$ and $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\delta}_{\xi^i}$. In this case, the minimization problem involved in the definition of the transport metric W_q becomes an assignment problem. Index the points $\hat{\xi}^1, \ldots, \hat{\xi}^N$ and $\xi^1, \ldots, \xi^N$ in such a way that an optimal assignment is given by $(\hat{\xi}^1, \xi^1), \ldots, (\hat{\xi}^N, \xi^N)$. Thus the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq r_0$ is equivalent to

$$\begin{aligned} & \left[\frac{1}{N} \sum_{i=1}^N \mathsf{d}_F^q \left(\xi^i, \hat{\xi}^i \right) \right]^{1/q} \leq r_0 \\ \Leftrightarrow \quad & \left[\frac{1}{N} \sum_{i=1}^N \liminf_{\xi^{i,n} \to \xi^i, \ \hat{\xi}^{i,n} \to \hat{\xi}^i} \mathsf{d} \left(F(\xi^{i,n}), F(\hat{\xi}^{i,n}) \right) \right]^{1/q} \leq r_0 \end{aligned} \tag{7}$$

For a set of N numbers $\{\xi_k^{i,n}\}_{i=1}^N$, the mapping

$$\xi_k^{i,n} \mapsto F_k(\xi_k^{i,n})$$

is non-decreasing and assigns each $\xi_k^{i,n}$ a value in $\{0, 1/N, 2/N, \ldots, 1\}$. Thus the vector $\Big(F_k(\xi_k^{1,n}), \ldots, F_k(\xi_k^{N,n})\Big)$ can be viewed as rankings of $\{\xi_k^{i,n}\}_{i=1}^N$ that take values in $\{\frac{i}{N} : 0 \le i \le N\}$ (same ranking for different elements are allowed). Therefore, the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \le r_0$ controls the difference between the rankings of the components of points in the supports of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. The following examples consider various special cases that explain this in more detail. □

Example 3 (Spearman's footrule distance) In the above example, when $q = 1$ and $\mathsf{d}(u, v) = \|u - v\|_1$, the above constraint (7) becomes

$$\begin{aligned} & \frac{1}{N}\sum_{i=1}^N\sum_{k=1}^K \liminf_{\xi_k^{i,n}\to\xi_k^i,\ \hat{\xi}_k^{i,n}\to\hat{\xi}_k^i} \left|F_k(\xi_k^{i,n}) - F_k(\hat{\xi}_k^{i,n})\right| \le r_0 \\ \Leftrightarrow \quad & \sum_{k=1}^K\sum_{i=1}^N \liminf_{\xi_k^{i,n}\to\xi_k^i,\ \hat{\xi}_k^{i,n}\to\hat{\xi}_k^i} \left|NF_k(\xi_k^{i,n}) - NF_k(\hat{\xi}_k^{i,n})\right| \le N^2 r_0 \end{aligned} \tag{8}$$

Note that $\sum_{i=1}^N \left|NF_k(\xi_k^{i,n}) - NF_k(\hat{\xi}_k^{i,n})\right|$ is a measure of the distance between the rankings of $\Big(F_k(\xi_k^{i,n}) : i = 1, \ldots, N\Big)$ and $\Big(F_k(\hat{\xi}_k^{i,n}) : i = 1, \ldots, N\Big)$, and is called Spearman's footrule in [5]. □

Example 4 (Comonotonicity) Let $K = 2$. Suppose that $\boldsymbol{\nu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$ is comonotonic, that is, for any two observations $\hat{\xi}^i = (\hat{\xi}_1^i, \hat{\xi}_2^i)$ and $\hat{\xi}^j = (\hat{\xi}_1^j, \hat{\xi}_2^j)$, it holds that $\hat{\xi}_1^i \le \hat{\xi}_1^j$ if and only if $\hat{\xi}_2^i \le \hat{\xi}_2^j$. Without loss of generality, assume that $\hat{\xi}^i$ are sorted in increasing order. In addition, assume that they have different values component-wise, that is, $\hat{\xi}_k^1 < \hat{\xi}_k^2 < \cdots < \hat{\xi}_k^N$ for all k. As before, consider $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\xi^i}$. Let $q = 1$. Then (8) is equivalent to

$$\sum_{k=1}^K\sum_{i=1}^N \min\left\{\liminf_{\xi_k^{i,n}\to\xi_k^i} |NF_k(\xi_k^{i,n}) - (i-1)|,\ \liminf_{\xi_k^{i,n}\to\xi_k^i} |NF_k(\xi_k^{i,n}) - i)|\right\} \ \le \ N^2 r_0,$$

which is an approximate measure of the deviation of $\boldsymbol{\mu}$ from comonotonicity. □

Example 5 (Box uncertainty set for each individual data) Let $q = \infty$. Then constraint (7) becomes

$$\liminf_{\xi^{i,n}\to\xi^i,\ \hat{\xi}^{i,n}\to\hat{\xi}^i} \mathsf{d}_F\Big(\Big(F_1(\xi_1^{i,n}), \ldots, F_K(\xi_K^{i,n})\Big), \Big(F_1(\hat{\xi}_1^{i,n}), \ldots, F_K(\hat{\xi}_K^{i,n})\Big)\Big) \ \le \ r_0, \quad \forall 1 \le i \le N.$$

Thus each individual data point is constrained to be in a certain region. In particular, when $\mathsf{d}(u, v) = ||u - v||_\infty$, then the above constraint becomes

$$\begin{aligned} \xi_k^i \ &\le \ \hat{\xi}_k^j, \quad \forall\ i, j \text{ such that } F_k^-(\hat{\xi}_k^j) - F_k(\hat{\xi}_k^i) \ge r_0, \\ \xi_k^i \ &\ge \ \hat{\xi}_k^j, \quad \forall\ i, j \text{ such that } F_k^-(\hat{\xi}_k^i) - F_k(\hat{\xi}_k^j) \ge r_0, \end{aligned}$$

where $F_k^-(\hat{\xi}_k) := \lim_{\xi_k \uparrow \hat{\xi}_k} F_k(\xi_k)$ denotes the left limit of F_k at $\hat{\xi}_k$. Thus, each ξ_k^i is constrained in some interval containing $\hat{\xi}_k^i$. In particular, if $r_0 = 0$, then ξ_k^i belongs to $[\hat{\xi}_k^{i-}, \hat{\xi}_k^{i+}]$, where $\hat{\xi}_k^{i-}$ (resp. $\hat{\xi}_k^{i+}$) is the largest (resp. smallest) data value among $\{\hat{\xi}_k^i\}_{i=1}^N$ that is strictly smaller (resp. greater) than $\hat{\xi}_k^i$. □

Example 6 (Copula) Let $\boldsymbol{\mu}$ be any continuous distribution on $\mathbb{R}^K$ with cumulative distribution function H. Sklar's theorem [24] states that there exists a joint distribution $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ on $[0, 1]^K$ with uniform marginals on $[0, 1]$, such that H can be expressed in terms of its marginal cumulative distribution functions $F_k^{\boldsymbol{\mu}}$ and $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ as follows:

$$H(\xi_1, \ldots, \xi_K) \ = \ \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}\left(F_1^{\boldsymbol{\mu}}(\xi_1), \ldots, F_K^{\boldsymbol{\mu}}(\xi_K)\right), \quad \forall\ \xi \in \mathbb{R}^K. \tag{9}$$

Such $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ is called a *copula*. Suppose that $F_k^{\boldsymbol{\mu}} = F_k$ for all k. Then using change-of-variables, $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu})$ can be written as

$$\begin{aligned}
&\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \\
&= \left(\min_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)} \left\{ \int_{\Xi^2} \mathsf{d}^q(F(\xi), F(\zeta)) \boldsymbol{\gamma}(d\xi, d\zeta) \ : \ \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mu}, \boldsymbol{\nu} \right\} \right)^{1/q} \\
&= \left(\min_{\boldsymbol{\pi} \in \mathcal{P}([0,1]^{2K})} \left\{ \int_{[0,1]^{2K}} \mathsf{d}^q(u, v) \boldsymbol{\pi}(du, dv) \ : \ \boldsymbol{\pi} \text{ has marginals } \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}} \right\} \right)^{1/q}.
\end{aligned}$$

Therefore, the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq r_0$ suggests that the copulas of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are close to each other in terms of Wasserstein distance. □

Based on these examples, we observe that for data-driven problems in which the distribution is finite-supported, the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq r_0$ controls the difference between rankings of the distributions, and more generally, it controls the difference between copulas of the distributions. The corresponding decision problem is given by

$$\inf_{x \in X} \sup_{\boldsymbol{\mu} \in \mathfrak{M}_2} \mathbb{E}_{\boldsymbol{\mu}}[\Psi(x, \boldsymbol{\xi})]. \tag{Rank-CRSO}$$

We study this problem in the second part of the paper.

1.3 Related literature

The study of distributionally robust stochastic optimization can be traced back to [28,8]. We have already mentioned several papers using moment-based and distance-based approaches. In addition, it is known that DRSO problems with $\mathfrak{M}$ given by a ϕ-divergence ball can easily incorporate moment constraints, simply by adding the constraints to the reformulation of the problem [27]. However, for DRSO problems with $\mathfrak{M}$ given by a Wasserstein ball, it has not been shown before whether moment constraints can be incorporated without losing tractability. In addition, since the Wasserstein metric includes the total variation distance as a special case (see, e.g., Remark 4 in [12]), our result for problem (Linear-CRSO) implies the result in [14], in which $\mathfrak{M}$ is given by a total variation distance constraint and a moment constraint. To the best of our knowledge, our paper is the first to study the effect of dependence structure (linear dependence and rank dependence) on stochastic optimization when only inexact information on the dependence structure is available. A related topic is to study the worst-case performance of stochastic optimization problems when marginal distributions are known and no information on the dependence structure is available, see [1,7,6].

1.4 Our contribution

We summarize our results as follows:

- Motivated by the poor performance of moment-based DRSO problems, we propose a new formulation (Linear-CRSO), that takes into account the distribution's second-order moment information as well as Wasserstein distance to the nominal distribution. Based on a constructive proof, we derive a tractable dual reformulation of this problem in Theorem 1.
- We investigate how the degree of correlation affects the performance of three DRSO approaches: DRSO with $\mathfrak{M}$ given by a Wasserstein ball, DRSO with $\mathfrak{M}$ given by moment constraints, and DRSO with $\mathfrak{M}$ given by both a Wasserstein ball and moment constraints (Linear-CRSO). Numerical results on a portfolio optimization problem indicate that the new formulation outperforms the others in all (low-, medium-, high-) correlation regimes.
- We also propose a new formulation (Rank-CRSO) which, with appropriately chosen parameters, controls different dependence structures, including Spearman's footrule distance between empirical rankings, comonotonicity, box uncertainty for individual data points, and in general Wasserstein distance between copulas. We also derive the dual reformulation for (Rank-CRSO) in Theorems 2 and 3.

2 Linear-correlationally robust stochastic optimization

In this section, we study linear correlationally robust stochastic optimization problem (Linear-CRSO). We will derive a convex programming dual reformulation of (Linear-CRSO) in Section 2.1, and apply it to a portfolio optimization problem in 2.2.

Since we focus only on the inner maximization problem, we suppress x in $\Psi(x,\xi)$. Throughout this section, we assume $1 \le p < \infty$, Ψ is upper semi-continuous, and satisfies the growth rate condition

$$\kappa \;:=\; \limsup_{d(\xi,\zeta_0)\to\infty} \frac{\max\big(\Psi(\xi)-\Psi(\zeta_0),0\big)}{d^p(\xi,\zeta_0)} < \infty,$$

for some $\zeta_0 \in \Xi$.

2.1 Dual reformulation of (Linear-CRSO)

The main result is given in the following Theorem 1.

Theorem 1 *Assume $\mathbb{E}_{\boldsymbol{\nu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top] \prec \Sigma_0$. Then* (Linear-CRSO) *has a strong dual problem*

$$\min_{\substack{\lambda\ge 0\\ \Lambda\succeq 0}} \left\{ \lambda R_0^p + \langle \Lambda, \Sigma_0\rangle + \int_\Xi \sup_{\xi\in\Xi} \big[\Psi(\xi) - \lambda d^p(\xi,\zeta) - (\xi-m_0)^\top \Lambda(\xi-m_0)\big]\boldsymbol{\nu}(d\zeta)\right\}.$$

In the dual problem, λ and Λ are the Lagrangian multiplier of primal Wasserstein constraint and second-order moment constraint. The dual objective is a convex function of λ and Λ. The measurability of the integrand is guaranteed by Lemma 1 below. Denote by $(\Xi, \mathscr{B}_{\boldsymbol{\nu}}(\Xi), \boldsymbol{\nu})$ the completion of measure space $(\Xi, \mathscr{B}(\Xi), \boldsymbol{\nu})$ (see, e.g., Lemma 1.25 in [15]). A function $f : \mathbb{R}^m \times \Xi \to \bar{\mathbb{R}}$ is called a *normal integrand* [20], if the associated epigraphical multifunction $\zeta \mapsto \text{epi } f(\cdot,\zeta)$ is closed valued and measurable.

Lemma 1 *The function $\Phi : \mathbb{R}\times\mathbb{R}^{K\times K}\times\Xi$ defined by*

$$\Phi(\lambda,\Lambda,\zeta) := \sup_{\xi\in\Xi}[\Psi(\xi) - (\xi-m_0)^\top\Lambda(\xi-m_0) - \lambda d^p(\xi,\zeta)]$$

is a normal integrand with respect to $\mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R}^{K\times K})\otimes\mathscr{B}_{\boldsymbol{\nu}}(\Xi)$.

Proof Define a function $g : \Xi\times\mathbb{R}\times\mathbb{R}^{K\times K}\times\Xi \to \bar{\mathbb{R}}$ by

$$g(\xi,\lambda,\Lambda,\zeta) = \Psi(\xi) - (\xi-m_0)^\top\Lambda(\xi-m_0) - \lambda d^p(\xi,\zeta).$$

Then for every $\zeta\in\Xi$, $-g(\cdot,\cdot,\cdot,\zeta)$ is lower semi-continuous, thus g is $\mathscr{B}(\Xi)\otimes\mathscr{B}(\mathbb{R})\otimes\mathscr{B}(\mathbb{R}^{K\times K})\otimes\mathscr{B}_{\boldsymbol{\nu}}(\Xi)$-measurable. Hence by joint measurability criterion (see, e.g., Corollary 14.34 in [20]), g is a normal integrand, thereby the function Φ is also a normal integrand (Theorem 7.38 in [23]). □

Proof (Proof of Theorem 1) We divide the proof into four steps.

Step 1. We first show weak duality. Observe that for any random vector $(\boldsymbol{\xi},\boldsymbol{\zeta})$ with joint distribution $\boldsymbol{\gamma}\in\mathcal{P}(\Xi^2)$ and marginals $\boldsymbol{\mu},\boldsymbol{\nu}\in\mathcal{P}$, by property of conditional expectation, it holds that

$$\int_\Xi \Psi(\xi)\boldsymbol{\mu}(d\xi) = \int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}(d\xi,d\zeta) = \int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta),$$

where $\boldsymbol{\gamma}_\zeta$ represents the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\zeta}=\zeta$. Thus we can write problem (Linear-CRSO) as

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1}\int_\Xi \Psi(\xi)\boldsymbol{\mu}(d\xi)\\
=&\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\Big\{\int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) : \int_{\Xi^2} d^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\le R_0^p,\\
&\qquad\qquad \int_{\Xi^2}(\xi-m_0)(\xi-m_0)^\top\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\preceq\Sigma_0\Big\}.
\end{aligned}$$

The Lagrangian weak duality yields that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1}\int_{\Xi}\Psi(\xi)\boldsymbol{\mu}(d\xi)\\
\leq&\inf_{\lambda\geq 0,\Lambda\succeq 0}\Big\{\lambda R_0^p+\langle\Lambda,\Sigma_0\rangle+\\
&\quad\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\Big\{\int_{\Xi^2}\big[\Psi(\xi)-(\xi-m_0)^\top\Lambda(\xi-m_0)-\lambda d^p(\xi,\zeta)\big]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\Big\}\Big\}\\
\leq&\inf_{\lambda\geq 0,\Lambda\succeq 0}\Big\{\lambda R_0^p+\langle\Lambda,\Sigma_0\rangle\\
&\qquad\qquad+\int_{\Xi}\sup_{\xi\in\Xi}[\Psi(\xi)-(\xi-m_0)^\top\Lambda(\xi-m_0)-\lambda d^p(\xi,\zeta)]\boldsymbol{\nu}(d\zeta)\Big\}.
\end{aligned}$$

Step 2. We next show the existence of a dual minimizer. Let $h(\lambda,\Lambda)$ be the dual objective function. To begin with, let v_d be the dual optimal value, and we claim that there exists $L>0$ such that

$$v_d=\inf_{0\leq\lambda\leq L,\Lambda\succeq 0}h(\lambda,\Lambda). \tag{10}$$

Indeed, the growth rate assumption on Ψ implies that there exists $M>0$ such that $\Psi(\xi)-\Psi(\zeta_0)\leq Md^p(\xi,\zeta_0)$ for all $\xi\in\Xi$. By choosing $\lambda=M$ and $\Lambda=0$, we have that

$$\begin{aligned}
v_d&\leq MR_0^p+\int_{\Xi}\sup_{\xi\in\Xi}[\Psi(\zeta_0)+M(d^p(\xi,\zeta_0)-d^p(\xi,\zeta))]\boldsymbol{\nu}(d\zeta)\\
&\leq MR_0^p+\Psi(\zeta_0)+M\int_{\Xi}d^p(\zeta_0,\zeta)\boldsymbol{\nu}(d\zeta)<\infty.
\end{aligned}$$

On the other hand, for any feasible solution (λ,Λ), it holds that

$$\begin{aligned}
h(\lambda,\Lambda)\geq&\lambda R_0^p+\langle\Lambda,\Sigma_0\rangle+\int_{\Xi}\big[\Psi(\zeta)-(\zeta-m_0)^\top\Lambda(\zeta-m_0)\big]\boldsymbol{\nu}(d\zeta)\\
=&\lambda R_0^p+\Big\langle\Lambda,\Sigma_0-\int_{\Xi}(\zeta-m_0)(\zeta-m_0)^\top\boldsymbol{\nu}(d\zeta)\Big\rangle+\int_{\Xi}\Psi(\zeta)\boldsymbol{\nu}(d\zeta)\\
\geq&\lambda R_0^p+\int_{\Xi}\Psi(\zeta)\boldsymbol{\nu}(d\zeta),
\end{aligned}$$

which tends to ∞ as $\lambda\to\infty$. Hence the claim holds by choosing sufficiently large L.

Now let $(\lambda^{(m)},\Lambda^{(m)})_m$ be a minimizing sequence of problem (10). Since $(\lambda^{(m)})_m\subset[0,L]$, Bolzano-Weierstrass theorem implies that it has a convergent subsequence, whose limit is denoted by λ_*. Fixing $\lambda=\lambda_*$, the weak Lagrangian dual of the problem

$$\inf_{\Lambda\succeq 0}h(\lambda_*,\Lambda) \tag{11}$$

is given by

$$\begin{aligned}
&\inf_{\Lambda\succeq 0}\Big\{\lambda_* R_0^p + \langle \Lambda, \Sigma_0\rangle \\
&\qquad\qquad + \int_\Xi \sup_{\xi\in\Xi}[\Psi(\xi) - (\xi-m_0)^\top \Lambda(\xi-m_0) - \lambda_* d^p(\xi,\zeta)]\boldsymbol{\nu}(d\zeta)\Big\} \\
\geq &\inf_{\Lambda\succeq 0}\Big\{\lambda_* R_0^p + \langle \Lambda, \Sigma_0\rangle \\
&\quad + \sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)} \int_{\Xi^2}[\Psi(\xi) - (\xi-m_0)^\top \Lambda(\xi-m_0) - \lambda_* d^p(\xi,\zeta)]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\Big\} \\
\geq &\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)} \inf_{\Lambda\succeq 0}\Big\{\lambda_* R_0^p + \langle \Lambda, \Sigma_0\rangle \\
&\qquad\qquad + \int_{\Xi^2}[\Psi(\xi) - (\xi-m_0)^\top \Lambda(\xi-m_0) - \lambda_* d^p(\xi,\zeta)]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\Big\} \\
= &\sup_{\boldsymbol{\gamma}\in\mathcal{P}(\Xi^2)}\Big\{\lambda_* R_0^p + \int_\Xi [\Psi(\xi) - \lambda_* d^p(\xi,\zeta)]\boldsymbol{\gamma}(d\xi,d\zeta) : \\
&\qquad\qquad \int_\Xi (\xi-m_0)(\xi-m_0)^\top \boldsymbol{\gamma}(d\xi,d\zeta) \preceq \Sigma_0\Big\}.
\end{aligned} \tag{12}$$

Note that $\mathbb{E}_{\boldsymbol{\nu}}[(\boldsymbol{\xi}-m_0)(\boldsymbol{\xi}-m_0)^\top] \prec \Sigma_0$, problem (12) satisfies the Slater condition, i.e., it is strictly feasible at $\boldsymbol{\gamma}_0$ defined by $\boldsymbol{\gamma}_0(A) := \boldsymbol{\nu}\{\zeta : (\zeta,\zeta)\in A\}$ for all Borel set $A\subset\Xi^2$, thus strong duality results for moment problem (cf. [22,4]) implies the existence of a dual minimizer Λ_* of problem (11). Therefore we have shown the existence of a dual minimizer (λ_*, Λ_*).

Step 3. We then establish the first-order optimality condition of the dual problem. By Lemma 1, the function

$$\Phi(\lambda,\Lambda,\zeta) = \sup_{\xi\in\Xi}[\Psi(\xi) - (\xi-m_0)^\top\Lambda(\xi-m_0) - \lambda d^p(\xi,\zeta)]$$

is a normal integrand. Moreover, for all $\zeta\in\Xi$, $\Phi(\cdot,\cdot,\zeta)$ is a convex function on $\mathbb{R}\times\mathbb{R}^{K\times K}$, and the growth rate condition implies that the set of maximizers is non-empty and compact for all $\lambda>\kappa$. Then by generalized Moreau-Rockafellar theorem (see, e.g., Theorem 7.47 in [23]), for any $(\lambda,\Lambda)\in \operatorname{dom} h\cap((\kappa,\infty)\times S_+^K)$, it holds that

$$\partial h(\lambda,\Lambda) = \begin{bmatrix} R_0^p \\ \Sigma_0\end{bmatrix} - \int_\Xi \partial\Phi(\lambda,\Lambda,\zeta)\boldsymbol{\nu}(d\zeta) + \mathcal{N}(\lambda,\Lambda),$$

where $\mathcal{N}(\lambda,\Lambda)$ stands for the normal cone at (λ,Λ) to the feasible region $\mathbb{R}_+\times S_+^K$. It follows from Theorem 2.4.18 in [29] that for any $(\lambda,\Lambda)\in\operatorname{dom} h\cap((\kappa,\infty)\times S_+^K)$,

$$\begin{aligned}
\partial\Phi(\lambda,\Lambda,\zeta) = \operatorname{conv}\Big\{&\begin{bmatrix} d^p(\xi,\zeta) \\ (\xi-m_0)^\top\Lambda(\xi-m_0)\end{bmatrix} : \\
&\xi\in\arg\max_{\xi\in\Xi}[\Psi(\xi) - (\xi-m_0)^\top\Lambda(\xi-m_0) - \lambda d^p(\xi,\zeta)]\Big\}.
\end{aligned}$$

Set

$$T_\lambda(\zeta) := \arg\max_{\xi\in\Xi}\ [\Psi(\xi) - (\xi-m_0)^\top\Lambda_*(\xi-m_0) - \lambda d^p(\xi,\zeta)].$$

The first-order optimality condition $0\in\partial h(\lambda_*,\Lambda_*)$ implies that there exists $\Sigma_*\in S_+^K$ with $\Sigma_*\preceq\Sigma_0$ and $\Lambda_*(\Sigma_0-\Sigma_*)=0$, such that if $\lambda_*>\kappa$, it holds that

$$\begin{bmatrix} R_0^p \\ \Sigma_*\end{bmatrix} \in \int_\Xi \operatorname{conv}\left\{\begin{bmatrix} d^p(\xi(\zeta),\zeta) \\ (\xi(\zeta)-m_0)^\top\Lambda_*(\xi(\zeta)-m_0)\end{bmatrix} :\ \xi(\zeta)\in T_{\lambda_*}(\zeta)\right\}\boldsymbol{\nu}(d\zeta), \tag{13}$$

and if $\lambda_*=\kappa$, for any $\lambda>\kappa$, there exists $R_\lambda\leq R_0$ such that

$$\begin{bmatrix} R_\lambda^p \\ \Sigma_*\end{bmatrix} \in \int_\Xi \operatorname{conv}\left\{\begin{bmatrix} d^p(\xi(\zeta),\zeta) \\ (\xi(\zeta)-m_0)^\top\Lambda_*(\xi(\zeta)-m_0)\end{bmatrix} :\ \xi(\zeta)\in T_{\lambda}(\zeta)\right\}\boldsymbol{\nu}(d\zeta). \tag{14}$$

Step 4. Finally, we construct a primal (approximate) optimal solution. Let us first consider the case $\lambda_* > \kappa$. Similar to the argument in Step 4 of proof for Theorem 2, (13) suggests that there exists a probability kernel $\{\boldsymbol{\gamma}_\zeta^*\}_{\zeta\in\Xi}$ such that each $\boldsymbol{\gamma}_\zeta^*$ is a probability distribution on $T_{\lambda_*}(\zeta)$, and satisfies

$$\int_{\Xi^2} d^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta) = R_0^p, \tag{15a}$$

$$\int_{\Xi^2} (\xi-m_0)(\xi-m_0)^\top\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta) = \Sigma_*, \tag{15b}$$

$$\Sigma_* \preceq \Sigma_0, \tag{15c}$$

$$\Lambda_*(\Sigma_0-\Sigma_*) = 0. \tag{15d}$$

Now define a probability measure $\boldsymbol{\mu}_*$ by

$$\boldsymbol{\mu}_*(A) := \int_\Xi \boldsymbol{\gamma}_\zeta^*(A)\boldsymbol{\nu}(d\zeta),\ \forall A\in\mathscr{B}(\Xi).$$

Then $\boldsymbol{\mu}_*$ is a primal feasible solution due to (15a). In addition,

$$\begin{aligned}
&\int_\Xi \Psi(\xi)\boldsymbol{\mu}_*(d\xi)\\
=&\int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta)\\
=&\int_{\Xi^2}[\Psi(\xi)-(\xi-m_0)^\top\Lambda_*(\xi-m_0)-\lambda_* d^p(\xi,\zeta)]\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta)+\lambda_* R_0^p+\langle\Lambda_*,\Sigma_*\rangle\\
=&\int_\Xi \max_{\xi\in\Xi}[\Psi(\xi)-(\xi-m_0)^\top\Lambda_*(\xi-m_0)-\lambda_* d^p(\xi,\zeta)]\boldsymbol{\nu}(d\zeta)+\lambda_* R_0^p+\langle\Lambda_*,\Sigma_0\rangle\\
\geq& v_d,
\end{aligned}$$

where the second and the third equalities follows from (15a)-(15d).

We then consider $\lambda_* = \kappa$. (14) suggests that for any $\lambda > \kappa$, there exists a probability kernel $\{\boldsymbol{\gamma}_\zeta^\lambda\}_{\zeta\in\Xi}$ such that each $\boldsymbol{\gamma}_\zeta^\lambda$ is a probability distribution on $T_\lambda(\zeta)$, and satisfies

$$\int_{\Xi^2} d^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta^\lambda(d\xi)\boldsymbol{\nu}(d\zeta) = R_\lambda^p, \tag{16a}$$

$$\int_{\Xi^2} (\xi-m_0)(\xi-m_0)^\top\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta) = \Sigma_*, \tag{16b}$$

$$\Sigma_* \preceq \Sigma_0, \tag{16c}$$

$$\Lambda_*(\Sigma_0-\Sigma_*) = 0. \tag{16d}$$

Now define a sequence of probabilty measures $\{\boldsymbol{\mu}^\lambda\}$ by

$$\boldsymbol{\mu}^\lambda(A) := \int_\Xi \boldsymbol{\gamma}_\zeta^\lambda(A)\boldsymbol{\nu}(d\zeta),\ \forall A\in\mathscr{B}(\Xi).$$

Then $\boldsymbol{\mu}^\lambda$ is a primal feasible solution due to (16a). In addition, from (16a)-(16d)

$$\begin{aligned}
&\int_\Xi \Psi(\xi)\boldsymbol{\mu}^\lambda(d\xi)\\
=&\int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta^\lambda(d\xi)\boldsymbol{\nu}(d\zeta)\\
=&\int_{\Xi^2}[\Psi(\xi)-(\xi-m_0)^\top\Lambda_*(\xi-m_0)-\lambda d^p(\xi,\zeta)]\boldsymbol{\gamma}_\zeta^\lambda(d\xi)\boldsymbol{\nu}(d\zeta)+\lambda R_0^p+\langle\Lambda_*,\Sigma_*\rangle\\
=&\int_\Xi \max_{\xi\in\Xi}[\Psi(\xi)-(\xi-m_0)^\top\Lambda_*(\xi-m_0)-\lambda d^p(\xi,\zeta)]\boldsymbol{\nu}(d\zeta)+\lambda R_0^p+\langle\Lambda_*,\Sigma_0\rangle,
\end{aligned}$$

which goes to $h(\kappa,\Lambda_*)$ as $\lambda\to\kappa$. Therefore, combined with Step 1, we have shown the strong duality. □

Remark 1 When Ψ is a concave function and Ξ is convex, $T_{\lambda_*}(\zeta)$ and $T_\lambda(\zeta)$ are convex sets, and thus the convex combination in (13)(14) belong to $T_{\lambda_*}(\zeta)$ and $T_\lambda(\zeta)$ respectively. Thus we can replace the convex hull by a single point in $T_{\lambda_*}(\zeta)$ or $T_\lambda(\zeta)$. It follows that when Ψ is concave and $\boldsymbol{\nu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$, it is sufficient to restrict $\mathfrak{M}_1$ on its subset $\mathfrak{M}_1 \cap \left\{\frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\xi^i} : \{\xi^i\}_{i=1}^N \subset \Xi\right\}$, the set of distributions in $\mathfrak{M}_1$ that are supported on at most N points:

$$\max_{\xi^i}\left\{\frac{1}{N}\sum_{i=1}^N \Psi(\xi^i):\ \frac{1}{N}\sum_{i=1}^N d^p(\xi^i,\hat{\xi}^i)\le R_0^p,\ \frac{1}{N}\sum_{i=1}^N(\xi^i-m_0)(\xi^i-m_0)^\top \preceq \Sigma_0\right\}.$$

□

For piecewise linear convex objective function $\Psi(\xi) = \max_{1\le j\le J} a_j^\top \xi + b_j$ and empirical nominal distribution, we have the following result.

Corollary 1 *Suppose* $\Psi(\xi) = \max_{1\le j\le J} a_j^\top \xi + b_j$, $\boldsymbol{\nu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$, $p = 1$ *and* $d(\xi,\zeta) = ||\xi-\zeta||$. *Then the DRSO problem admits a semi-definite program reformulation:*

$$\begin{aligned}\min_{\substack{y^i\in\mathbb{R}\\ \lambda\ge 0,\Lambda\succeq 0}}\quad & \lambda R_0 + \langle \Lambda, \Sigma_0\rangle + \frac{1}{N}\sum_{i=1}^N y_i\\ s.t.\quad & \begin{bmatrix}\Lambda & -\frac{1}{2}a_j+\frac{1}{2}\zeta^i-\Lambda m_0\\ (-\frac{1}{2}a_j+\frac{1}{2}\zeta^i-\Lambda m_0)^\top & -b_j-{\zeta^i}^\top\hat{\xi}^i+m_0^\top\Lambda m_0+y_i\end{bmatrix}\succeq 0,\\ & \qquad\qquad \forall 1\le j\le J,\ \forall 1\le i\le N,\\ & ||\zeta^i||_*\le\lambda,\quad \forall 1\le i\le N,\end{aligned}$$

where $||\cdot||_*$ *denotes the dual norm of* $||\cdot||$.

Proof Observe that $||\xi-\hat{\xi}^i|| = \sup_{||\zeta||_*\le 1}\zeta^\top(\xi-\hat{\xi}^i)$, then by convex programming duality, we have that

$$\begin{aligned}&\max_{\xi\in\Xi}\left\{a_j^\top\xi+b_j-\lambda||\xi-\hat{\xi}^i||-(\xi-m_0)^\top\Lambda(\xi-m_0)\right\}\\ =&\max_{\xi\in\Xi}\inf_{||\zeta||_*\le 1}\left\{a_j^\top\xi+b_j-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\right\}\\ =&\inf_{||\zeta||_*\le 1}\max_{\xi\in\Xi}\Big\{\sum_j\alpha_{ij}(a_j^\top\xi+b_j)-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\Big\}.\end{aligned}$$

Hence the constraint

$$y_i\ge\max_{\xi\in\Xi}\left\{a_j^\top\xi+b_j-\lambda||\xi-\hat{\xi}^i||-(\xi-m_0)^\top\Lambda(\xi-m_0)\right\}$$

can be written as

$$\exists||\zeta||_*\le 1,\ s.t.\ y_i\ge\max_{\xi\in\Xi}\left\{a_j^\top\xi+b_j-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\right\},$$

which is further equivalent to

$$\exists||\zeta||_*\le 1,\ s.t.\ \begin{bmatrix}\Lambda & -\frac{1}{2}a_j+\frac{1}{2}\lambda\zeta-\Lambda m_0\\ (-\frac{1}{2}a_j+\frac{1}{2}\lambda\zeta-\Lambda m_0)^\top & -b_j-\lambda\zeta^\top\hat{\xi}^i+m_0^\top\Lambda m_0+y_i\end{bmatrix}\succeq 0.$$

Replacing $\lambda\zeta$ by ζ^i we obtain the result. □

2.2 Application in portfolio optimization

In this section, we study the effect of degree of correlation on the performance of (Linear-CRSO), through a mean-risk portfolio optimization problem adapted from [10]. In this problem, the random returns of K risky assets is captured by a random vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)$, the decision variable $x \in X = \{x \in \mathbb{R}^K_+ : \sum_{k=1}^K x_k = 1\}$ represents the portfolio weights without short-selling, and the goal is to minimize a weighted combination of the mean and conditional value-at-risk of the negative of portfolio return $-x^\top \boldsymbol{\xi}$:

$$\min_{x\in X} \sup_{\boldsymbol{\mu}\in\mathfrak{M}_1} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[-x^\top \boldsymbol{\xi}] + c \cdot \mathrm{CVaR}^{\boldsymbol{\mu}}_\alpha(-x^\top \boldsymbol{\xi}) \right\},$$

where c is some constant. It can be written as (see [10] for a derivation)

$$\inf_{x\in X, \tau\in\mathbb{R}} \sup_{\boldsymbol{\mu}\in\mathfrak{M}_1} \left\{ \max_{1\le j\le J} a_j x^\top \boldsymbol{\xi} + b_j \tau \right\}, \tag{17}$$

where $J = 2$, $a_1 = -1$, $a_2 = -1 - c/\alpha$, $b_1 = c$ and $b_2 = c(1 - 1/\alpha)$. Suppose the nominal distribution $\boldsymbol{\nu}$ is chosen as the empirical distribution, then we can obtain a semi-definite programming reformulation using Corollary 1. In defining $\mathfrak{M}_1$, we set m_0 to be the sample mean vector, and Σ_0 to be sample covariance matrix $\hat{\Sigma}$ inflated by a factor $\gamma \ge 1$.

The parameters for the numerical experiments are given as follows. We set $K = 10$, $c = 10$, $\alpha = 20\%$. We consider Wasserstein distance of order $p = 1$ and Ξ to be the Euclidean space $\mathbb{R}^K$. Assume $\boldsymbol{\xi}$ is joint Gaussian distributed, and each $\boldsymbol{\xi}_k$ has mean $0.03k$ and standard deviation $0.025k$, $k = 1, \ldots, K$. The correlation between $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_j$ is set to be $\rho^{|k-j|}$ for $\rho = 0.5, 0.6, 0.7, 0.8, 0.9, 0.99$. Note that in $0.9^5 \simeq 0.69$ and $0.6^5 \simeq 0.08$, we can view $\rho = 0.99, 0.9$ as high correlation regime, $\rho = 0.8, 0.7$ as medium correlation regime, and $\rho = 0.6, 0.5$ as low correlation regime. We consider small dataset regime $N = 40$, for which DRSO approach should be more suitable than SAA (sample average approximation) method. We run the simulation with 200 repetitions.

The tuning parameters (Wasserstein radius R_0 and inflation factor γ) are selected using hold-out cross validation. In each repetition, the N samples are randomly partitioned into a training dataset (70% of the data) and a validation dataset (the remaining 30%). For different tuning parameters, we use the training dataset to solve problem (17) and use the validation dataset to estimate the out-of-sample performance of different parameter values and select the ones with the best performance. Then we resolve problem (17) with the best tuning parameters using all N samples and obtain the optimal solutions for the three uncertainty sets. Finally we examine the out-of-sample performance of these solutions using an independent testing dataset consisting of 10^3 samples.

Figure 1 shows the box plots of the optimal values in 200 repetitions of four different approaches: Sample Average Approximation, DRSO with Wasserstein uncertainty set, DRSO with moment uncertainty set and (Linear-CRSO), and the solid curves represent the average performance of these approaches. We observe that our new formulation (Linear-CRSO) performs consistently the best in all regimes; DRSO with moment uncertainty set (3) performs well in high-correlation regime; and DRSO with Wasserstein uncertainty set performs well in medium- and low-correlation regime.

To provide a rough explanation of the results, we begin with describing the behavior of the worst-case distributions in three approaches. Observe that the objective of problem (17) is a weighted combination of mean and CVaR and is piecewise-linear convex in ξ. To achieve the inner maximization, the worst-case distributions of Wasserstein uncertainty set and (Linear-CRSO) tend to perturb the data points with small returns (large value of $-x^\top \xi$) in the direction of minimizing $-x^\top \xi$. Meanwhile, for moment uncertainty set, according to Proposition 3 in [4], the worst-case distribution also tends to spread out the probability mass towards this direction until the centered second-order moment constraint becomes tight.

Now let us consider the situations in different correlation regimes. In high-correlation regime, the optimal portfolio under true underlying distribution puts almost all the weight on the first asset, which has the smallest weighted combination of mean and CVaR among all K assets. Since the expected returns are almost comonotonic, DRSO with moment uncertainty set becomes finding a worst-case distribution among all univariate distributions with given mean and variance. This is a relatively small set of distributions, and considering the underlying distribution is Gaussian, the solution yielding from moment uncertainty set is not overly conservative, and can effectively identify a portfolio close to the true optimal one. On the other hand, for Wasserstein uncertainty set, the worst-case distribution perturbs the data with small returns in the direction of minimizing $-x^\top \xi$. Note that in the presence of high correlation, data points are concentrating on a subspace of $\mathbb{R}^K$ whose dimension is much less than K, and they can

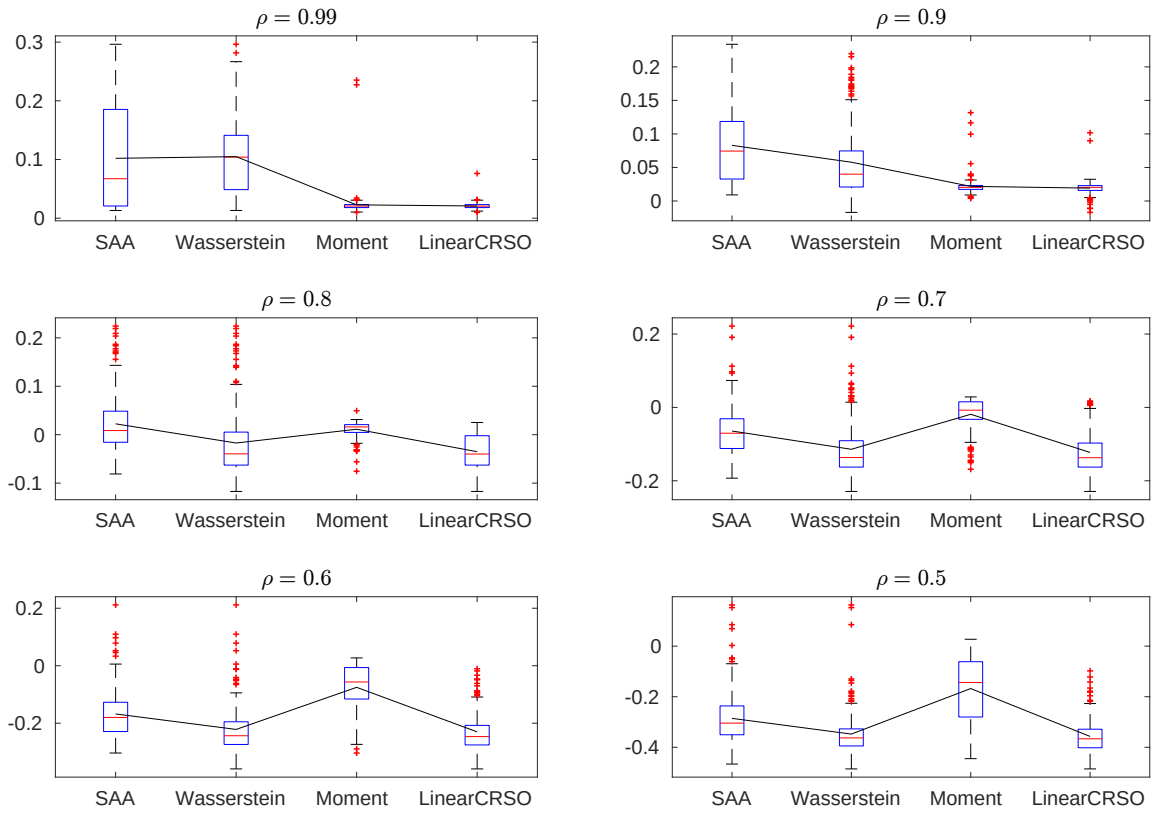

Fig. 1 Out-of-sample performance under different degrees of correlation

be easily perturbed out of this subspace. Therefore the worst-case distribution may not lie in the same subspace as the empirical distribution. Hence the Wasserstein uncertainty set may hedge against some distributions unlikely to happen, which makes the decision over-conservative. Indeed, the solution yielding from Wasserstein uncertainty set tends to equally allocate the weights to all assets. In medium- and low-correlation regime, the data points are not so concentrating, so perturbation towards any direction does not affect the correlation structure too much, hence Wasserstein constraint alone performs well, and almost as good as (Linear-CRSO) in low-correlation regime. But moment uncertainty set is too conservative since it contains too many distributions. As a hybrid approach, (Linear-CRSO) take advantages of Wasserstein and moment uncertainty sets.

3 Rank-correlationally robust stochastic optimization

In this section, we study rank-correlationally robust stochastic optimization problem (Rank-CRSO). Dual reformulation of the problem is derived in 3.1, and comparison of (Rank-CRSO) with other approaches are in 3.2 and 3.3. As in previous section, we suppress x in $\Psi(x,\xi)$. Throughout this section, we assume $p \in [1,\infty)$, $q \in [1,\infty]$, Ψ is upper semi-continuous, and satisfies the growth rate condition

$$\limsup_{d(\xi,\zeta_0)\to\infty} \frac{\max\big(\Psi(\xi)-\Psi(\zeta_0),0\big)}{d^p(\xi,\zeta_0)} = 0,$$

for some $\zeta_0 \in \Xi$.

3.1 Dual reformulation of (Rank-CRSO)

We consider $q \in [1,\infty)$ in Theorem 2 and $q=\infty$ in Theorem 3.

Theorem 2 *Suppose* $1 \le q < \infty$.

(i) Problem (Rank-CRSO) *can be reformulated as*

$$\min_{\alpha,\beta\geq 0}\left\{\alpha R_0^p+\beta r_0^q+\int_{\Xi}\max_{\xi\in\Xi}\left[\Psi(\xi)-\alpha d^p(\xi,\zeta)-\beta \mathsf{d}_F^q(\xi,\zeta)\right]\boldsymbol{\nu}(d\zeta)\right\}. \tag{18}$$

(ii) For data-driven problem, i.e., $\boldsymbol{\nu}=\frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$*, the following program*

$$\max_{\{\xi^i\}\subset\Xi}\left\{\frac{1}{N}\sum_{i=1}^N\Psi(\xi^i):\ \frac{1}{N}\sum_{i=1}^N d^p(\xi^i,\hat{\xi}^i)\leq R_0^p,\ \frac{1}{N}\sum_{i=1}^N \mathsf{d}_F^q(\xi^i,\hat{\xi}^i)\leq r_0^q\right\} \tag{19}$$

is a $(1-O(\frac{1}{N}))$*-approximation of the inner maximization of* (Rank-CRSO).

Remark 2 In Section 1.2, we have showed various examples regrading different notions of distance by assuming the relevant distribution $\boldsymbol{\mu}$ is supported on (at most) N points. Thanks to statement (ii) (and also Theorem 3(ii) below), for data-driven problems, this assumption is almost not restrictive at all, since the worst-case distribution of the approximation problem (19) is supported on at most N points. □

Proof (Proof of Theorem 2) We first prove (i). The idea is similar to the proof of Theorem 1. The measurability of the integrand in the dual problem follows similarly to Lemma 1, so we omit the proof. Using Lagrangian weak duality we have weak duality

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathfrak{M}_2}\left\{\int_{\Xi}\Psi(\xi)\boldsymbol{\mu}(d\xi)\right\}\\
=&\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\inf_{\alpha,\beta\geq 0}\left\{\int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta(d\zeta)\boldsymbol{\nu}(d\zeta)+\alpha R_0^p-\alpha\int_{\Xi^2}d^q(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right.\\
&\qquad\left.+\beta r_0^q-\beta\int_{\Xi^2}\mathsf{d}_F^q(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}\\
\leq&\inf_{\alpha,\beta\geq 0}\left\{\alpha R_0^p+\beta r_0^q+\int_{\Xi}\max_{\xi\in\Xi}\left[\Psi(\xi)-\alpha d^p(\xi,\zeta)-\beta\mathsf{d}_F^q(\xi,\zeta)\right]\boldsymbol{\nu}(d\zeta)\right\}.
\end{aligned}$$

We next show that the dual problem admits a minimizer. Let v_d be the optimal value of the dual problem. We claim that there exists $L>0$ such that

$$\begin{aligned}
&v_d=\\
&\inf_{0\leq\alpha,\beta\leq L}\left\{\alpha R_0^p+\beta r_0^q+\int_{\Xi}\max_{\xi\in\Xi}\left[\Psi(\xi)-\alpha d^p(\xi,\zeta)-\beta\mathsf{d}_F^q(\xi,\zeta)\right]\boldsymbol{\nu}(d\zeta)\right\}.
\end{aligned}\tag{20}$$

Indeed, according to the growth-rate assumption on Ψ, there exists $M>0$ and $\zeta_0\in\Xi$ such that $\Psi(\xi)-\Psi(\zeta_0)\leq Md^p(\xi,\zeta_0)$ for all $\xi\in\Xi$, thus by choosing $\alpha=M$ and $\beta=0$, we obtain that

$$\begin{aligned}
v_d\leq& MR_0^p+\int_{\Xi}\max_{\xi\in\Xi}[\Psi(\xi)+M(d^p(\xi,\zeta_0)-d^p(\xi,\zeta))]\boldsymbol{\nu}(d\zeta)\\
\leq& MR_0^p+\Psi(\xi)+M\int_{\Xi}d^p(\zeta,\zeta_0)\boldsymbol{\nu}(d\zeta)<\infty.
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
&\alpha R_0^p+\beta r_0^q+\int_{\Xi}\max_{\xi\in\Xi}\left[\Psi(\xi)-\alpha d^p(\xi,\zeta)-\beta\mathsf{d}_F^q(\xi,\zeta)\right]\boldsymbol{\nu}(d\zeta)\\
\geq&\alpha R_0^p+\beta r_0^q+\int_{\Xi}\Psi(\zeta)\boldsymbol{\nu}(d\zeta),
\end{aligned}$$

which tends to infinity as $\max(\alpha,\beta)\to\infty$. Hence the claim holds. Then by Bolzano-Weierstrass theorem there exists a minimizer.

We then establish the first-order optimality condition of problem (20) at a minimizer (α_*,β_*). Let $h(\alpha,\beta)$ be the objective function and set

$$\Phi(\alpha,\beta,\zeta)=\Psi(\xi)-\alpha d^p(\xi,\zeta)-\beta\mathsf{d}_F^q(\xi,\zeta).$$

For any $(\alpha,\beta)\in\operatorname{dom}h$ we compute the differential of h and Φ with respect to α,β:

$$\partial h(\alpha,\beta)=(R_0^p,r_0^q)^\top-\int_{\Xi}\partial_{\alpha,\beta}\Phi_n(\alpha,\beta,v)\boldsymbol{\nu}(d\zeta)+\mathcal{N}(\alpha,\beta),$$

where $\mathcal{N}(\alpha,\beta)$ stands for the normal cone at (α,β) to the feasible region $\mathbb{R}^2_+$, and

$$\partial_{\alpha,\beta}\Phi(\alpha,\beta,\zeta)=\mathrm{conv}\Big\{\big(d^p(\xi,\zeta),\mathsf{d}_F^q(\xi,\zeta)\big)^\top:\ \xi\in T(\zeta)\Big\},$$

where

$$T(\zeta):=\arg\max_{\xi\in\Xi}\big[\Psi(\xi)-\alpha_* d^p(\xi,\zeta)-\beta_*\mathsf{d}_F^q(\xi,\zeta)\big].$$

The first-order optimality condition $0\in\partial h(\alpha_*,\beta_*,0)$ implies that there exists $0\le R_*\le R_0$ with $\alpha_*(R_0-R_*)=0$ and $0\le r_*\le r_0$ with $\beta_*(r_0-r_*)=0$, such that

$$(R_*^p,r_*^q)\in\int_\Xi \mathrm{conv}\Big\{\big(d^p(\xi(\zeta),\zeta),\mathsf{d}_F^q(\xi(\zeta),\zeta)\big):\xi(\zeta)\in T(\zeta)\Big\}\boldsymbol{\nu}(d\zeta). \tag{21}$$

Finally we construct a primal optimal solution. (21) suggests that there is a probability kernel $\{\boldsymbol{\gamma}_\zeta^*\}_{\zeta\in\Xi}$ such that each $\boldsymbol{\gamma}_\zeta^*$ is a probability distribution on $T(\zeta)$ and satisfies

$$\int_{\Xi^2} d^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta)=R_*^p, \tag{22a}$$

$$R_*\le R_0, \tag{22b}$$

$$\alpha_*(R_0-R_*)=0, \tag{22c}$$

$$\int_{\Xi^2} \mathsf{d}_F^q(\xi,\zeta)\boldsymbol{\gamma}_\zeta^*(d\xi)\boldsymbol{\nu}(d\zeta)=r_*^q, \tag{22d}$$

$$r_*\le r_0, \tag{22e}$$

$$\beta_*(r_0-r_*)=0. \tag{22f}$$

We can verify that the probability measure $\boldsymbol{\mu}_*$ by

$$\boldsymbol{\mu}_*(A):=\int_\Xi \boldsymbol{\gamma}_\zeta^*(A)\boldsymbol{\nu}(d\zeta),\ \forall A\in\mathscr{B}(\Xi).$$

is a primal feasible and optimal, and thus we have prove (i).

Now we prove (ii). Let (α_*,β_*) be the optimal dual solution, then for data-driven problem, the dual optimal values equals

$$\alpha_* R_0^p+\beta_* r_0^q+\frac{1}{N}\sum_{i=1}^N\max_{\xi\in\Xi}\big[\Psi(\xi)-\alpha_* d^p(\xi,\hat\xi^i)-\beta_*\mathsf{d}_F^q(\xi,\hat\xi^i)\big], \tag{23}$$

Caratheodory's theorem implies that for each $\hat\xi^i$, we can choose $\boldsymbol{\gamma}_{\hat\xi^i}^*$ defined by (22) such that its support contains at most three points. Hence, the third term in problem (23) is equivalent to

$$\max_{p^{ij}\ge0,\xi^{ij}\in\Xi}\Bigg\{\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^3 p^{ij}\big[\Psi(\xi^{ij})-\alpha_* d^p(\xi^{ij},\hat\xi^i)-\beta_*\mathsf{d}_F^q(\xi^{ij},\hat\xi^i)\big]:\ \sum_j p^{ij}=1,\ \forall i\Bigg\},$$

which is also equivalent to

$$\max_{p^{ij}\ge0,\xi^{ij}\in\Xi}\Bigg\{\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^3 p^{ij}\Psi(\xi^{ij}):\ \sum_j p^{ij}=1,\ \forall i,\ \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^3 p^{ij}\big[\alpha_* d^p(\xi^{ij},\hat\xi^i)+\beta_*\mathsf{d}_F^q(\xi^{ij},\hat\xi^i)\big]\le\alpha_* R_0^p+\beta_* r_0^q\Bigg\},$$

Let us fix $\{\xi^{ij}\}_{ij}$ to be the optimal value $\{\xi_*^{ij}\}_{ij}$ in the program above and consider finding the best $\{p^{ij}\}_{ij}$. This is continuous knapsack problem which can be solved by a greedy algorithm. More specifically, those (i,j) with large value of $\Psi(\xi^{ij})$ are preferably to be set to one, until the knapsack constraints is violated, when we may set such p^{ij} to be a fractional value to make the knapsack constraint tight. Hence,

there exists an optimal solution $\{p_*^{ij}\}_{ij}$ with at most two fractional points. Now we modify it by making the fractional values to be zero, and denote the modified solution by $\{\tilde{p}^{ij}\}_{ij}$. Then $\{\tilde{p}^{ij}\}_{ij}$ is also feasible, and yields a feasible distribution of (19)

$$\frac{1}{N}\sum_{i=1}^{N}\Big[\sum_{j=1}^{3}\tilde{p}^{ij}\boldsymbol{\delta}_{\xi_*^{ij}}+(1-\sum_{j=1}^{3}\tilde{p}^{ij})\boldsymbol{\delta}_{\hat{\xi}^i}\Big].$$

The growth-rate assumption on Ψ implies that ξ_*^{ij} are uniformly bounded by some constant M. Hence the objective value of this feasible distribution differ from the optimal value by at most $O(\frac{1}{N})$. □

Theorem 3 *Suppose $q=\infty$.*

(i) Problem (Rank-CRSO) *can be reformulated as*

$$\min_{\alpha\geq 0}\left\{\alpha R_0^p+\int_{\Xi}\max_{\xi\in\Xi}\{\Psi(\xi)-\alpha d^p(\xi,\zeta):\ \mathsf{d}_F(\xi,\zeta)\leq r_0\}\boldsymbol{\nu}(d\zeta)\right\}.$$

(ii) For data-driven problem, i.e., $\boldsymbol{\nu}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\delta}_{\hat{\xi}^i}$, the following program

$$\begin{aligned}v_N:=\sup_{\xi^i\in\Xi}\Big\{\frac{1}{N}\sum_{i=1}^{N}\Psi(\xi^i):\ &\frac{1}{N}\sum_{i=1}^{N}d^p(\xi^i,\hat{\xi}^i)\leq R_0^p,\\ &\mathsf{d}(F(\xi^i),F(\hat{\xi}^i))\leq r_0,\ \forall 1\leq i\leq N\Big\}\end{aligned}\tag{24}$$

is a $(1-O(\frac{1}{N}))$-approximation of the inner maximization of (Rank-CRSO). *In particular when $\mathsf{d}(u,v)=||u,v||_\infty$, the second constraint in* (24) *can be written as linear constraint*

$$\begin{aligned}\xi_k^i\ &\leq\ \hat{\xi}_k^j,\quad \forall\ i,j \text{ such that } F_k^-(\hat{\xi}_k^j)-F_k(\hat{\xi}_k^i)\geq r_0,\\ \xi_k^i\ &\geq\ \hat{\xi}_k^j,\quad \forall\ i,j \text{ such that } F_k^-(\hat{\xi}_k^i)-F_k(\hat{\xi}_k^j)\geq r_0,\end{aligned}\tag{25}$$

where $F_k^-(\hat{\xi}_k):=\lim_{\xi_k\uparrow\hat{\xi}_k}F_k(\xi_k)$ denotes the left limit of F_k at $\hat{\xi}_k$. If, in addition, Ξ is convex and Ψ is a concave function, then the approximation is exact.

Proof We first prove (i). For each $\zeta\in\Xi$, set $B_\zeta(r_0):=\{\xi\in\Xi:\ \mathsf{d}_F(\xi,\zeta)\leq r_0\}$. It follows that $B_\zeta(r_0)$ is closed. Using the same reasoning as in the proof of Theorem 2, we can obtain the strong duality

$$\begin{aligned}&\sup_{\boldsymbol{\mu}\in\mathfrak{M}_2}\left\{\int_{\Xi}\Psi(\xi)\boldsymbol{\mu}(d\xi)\right\}\\ =&\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\inf_{\alpha,\beta\geq 0}\Big\{\int_{\Xi}\Psi(\xi)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)+\alpha R_0^p-\alpha\int_{\Xi}d^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta):\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{supp }\gamma_\zeta\subset B_\zeta(r_0),\ \forall\zeta\in\Xi\Big\}\\ =&\min_{\alpha\geq 0}\left\{\alpha R_0^p+\max_{\boldsymbol{\gamma}_\zeta\in\mathcal{P}(B_\zeta(r_0))}\int_{\Xi}[\Psi(\xi)-\alpha d^p(u,v)]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}.\end{aligned}$$

Thus we obtain (i).

The first part of statement (ii) follows essentially the same as in the proof of Theorem 2(ii). When Ξ is convex and d is induced from $||\cdot||_\infty$, the reformulation of the second constraint follows by definition of d_F, and in this case, $B_\zeta(r_0)$ is a cube and thus convex. If, in addition, Ψ is concave, then using the same reasoning as in Remark 1 we obtain the result. □

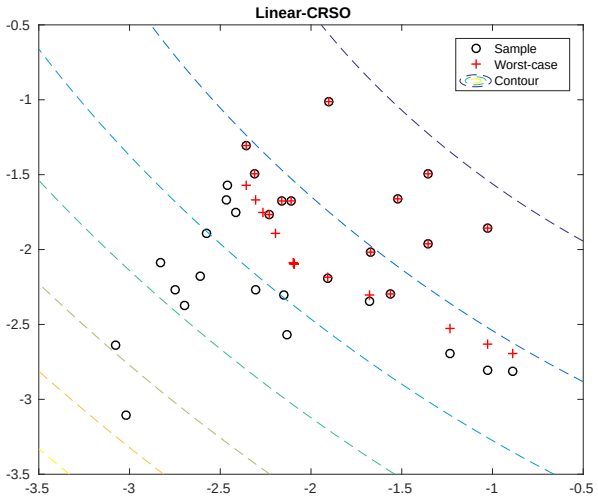

Fig. 2 One-sided second-order moment constraint may be not tight

3.2 Comparison of (Linear-CRSO) and (Rank-CRSO)

It may appear that formulation (Linear-CRSO) and (Rank-CRSO) are in different nature and should have different scopes of application, in the sense that (Linear-CRSO) considers linear dependence, and (Rank-CRSO) controls rank dependence as $\{F_k(\hat{\xi}_k^i)\}$ is an ordinal scaling of the data $\{\hat{\xi}_k^i\}$. However, we point out this is not the case, since our new formulation controls not only ordinal association, but to some extent, also the dependence without scaling. This is because for relevant distribution $\boldsymbol{\mu} \in \mathfrak{M}_2$, instead of using its own marginal distributions to make the scaling, we use F_k, the marginal distribution of the nominal distribution $\boldsymbol{\nu}$. Consequently, the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \le r_0$ also controls the values, not only the ordinal relationship, of the worst-case distribution. For instance, the box uncertainty set in Example 5 and Theorem 3(ii) confines the region where each data point can be perturbed. The following example shows that the worst-case distribution yielding from (Rank-CRSO) has both similar linear and rank correlation to those of the nominal distribution, whereas (Linear-CRSO) even does not control the linear correlation, since the one-sided moment constraint may be not tight.

Example 7 Consider a concave objective function

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_1} \mathbb{E}_{\boldsymbol{\mu}}[-\boldsymbol{\xi}_1^2 - \boldsymbol{\xi}_1\boldsymbol{\xi}_2 - \boldsymbol{\xi}_2^2],$$

where $\Xi = \mathbb{R}^2$ equipped with ℓ_1-norm, $\boldsymbol{\nu}$ is the empirical distribution i.i.d. drawn from a normal distribution $N\left([1,1], \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}\right)$, and $R_0 = 0.3$. A tractable reformulation based on Theorem 1 is provided in the appendix. The empirical distribution and the worst-case distributions yielding from (Linear-CRSO) are shown in Fig. 2.

For the worst-case distribution, the moment constraint in (Linear-CRSO) is loose, with left-hand side equal to $\begin{bmatrix} 0.216 & -0.056 \\ -0.056 & 0.150 \end{bmatrix}$, less than the sample covariance matrix $\begin{bmatrix} 0.360 & -0.008 \\ -0.008 & 0.235 \end{bmatrix}$. The correlation of the worst-case distribution is -0.491, whereas the sample correlation is -0.029. Thereby the correlation structures of the worst-case distribution differs from that of the nominal distribution a lot. The intuition is as follows. The worst-case distribution of Wasserstein uncertainty set tends to perturb the sample points with large gradient, and since the objective function is concave, the perturbation often leads to a distribution with smaller variance. Hence the second-order moment constraint on the main diagonal elements is much smaller than the right-hand side, which gives more flexibility on the off-diagonal elements. More specifically, let $\Sigma_0 = \begin{bmatrix} \sigma_1^2 & \rho_0\sigma_1\sigma_2 \\ \rho_0\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$, and suppose the variance of the worst-case distribution is $[(1-\epsilon_1)\sigma_1^2, (1-\epsilon_2)\sigma_2^2]$, and let ρ be the correlation of the worst-case distribution. Then the covariance matrix of the worst-case distribution is written as $\begin{bmatrix} (1-\epsilon)\sigma_1^2 & \rho\sqrt{(1-\epsilon_1)(1-\epsilon_2)}\sigma_1\sigma_2 \\ \sqrt{(1-\epsilon_1)(1-\epsilon_2)}\rho\sigma_1\sigma_2 & (1-\epsilon)\sigma_2^2 \end{bmatrix}$. To satisfy the moment constraint, the correlation ρ of the worst-case distribution must satisfy $|\rho - \rho_0| \le \frac{\epsilon}{1-\epsilon}$. When ϵ is not so small, ρ could have a large discrepancy comparing with the nominal correlation ρ_0.

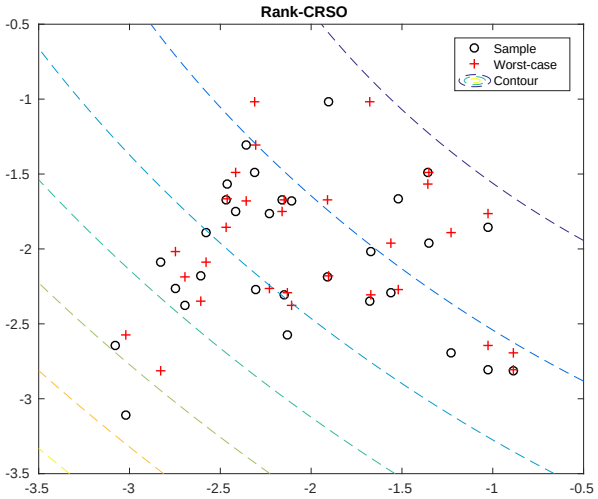

Fig. 3 (Rank-CRSO) controls both rank and linear dependence

Table 1 Distances between bivariate Gaussian copulas

Distances	Fisher-Rao	KL	Burg	Hellinger	Bhattacharya	TV	W_2
$\mathcal{C}^{\boldsymbol{\mu}_1}$, $\mathcal{C}^{\boldsymbol{\mu}_2}$	2.77	22.56	1.48	0.69	0.65	2.45	0.15
$\mathcal{C}^{\boldsymbol{\mu}_2}$, $\mathcal{C}^{\boldsymbol{\mu}_3}$	3.26	47.20	1.81	0.75	0.81	4.42	0.03

In contrast, if we use reformulation (24) of Rank-CRSO with $r_0 = 0$, the worst-case distribution is shown in Fig. 3. It has the same rank correlation as the nominal distribution, and its linear correlation coefficient equals -0.072. □

3.3 Comparison of Wasserstein distance with other distances

In Example 6, we show that for continuous distribution, the constraint $\mathsf{W}_q(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq r_0$ controls the Wasserstein distance between the copula of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Then one may ask consider some other distances and formulate a problem similar to (Rank-CRSO). We point out that comparing to many other commonly used distances, Wasserstein metrics yields a more intuitive quantitative relationship for copulas of distribution with highly correlated components, as shown by the following example.

Example 8 (Distances between Gaussian copulas) Let $R \in [-1, 1]^{K \times K}$ be a correlation matrix. The Gaussian copula with parameter R is given by

$$\mathcal{C}_R(u) = \phi_R(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d)),$$

where ϕ is the cumulative distribution function of a standard Gaussian distribution, and ϕ_R is the cumulative distribution function of a multivariate Gaussian distribution $\mathcal{N}(0, R)$. Now let $\boldsymbol{\mu}_i$, $i = 1, 2, 3$ be three bivariate Gaussian distributions, each of which has standard Gaussian marginal distributions, and the correlation matrices are respectively $R_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $R_2 = \begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$, and $R_3 = \begin{bmatrix} 1 & 0.9999 \\ 0.9999 & 1 \end{bmatrix}$. Their copulas are denoted by $\mathcal{C}^{\boldsymbol{\mu}_i}$ respectively. Various distances between $\mathcal{C}^{\boldsymbol{\mu}_1}$ and $\mathcal{C}^{\boldsymbol{\mu}_2}$ and between $\mathcal{C}^{\boldsymbol{\mu}_2}$ and $\mathcal{C}^{\boldsymbol{\mu}_3}$ are shown in Table 1, in which total variation distance and W_2-Wasserstein metric are computed numerically, while the others use closed-form formulas (cf. [17]).

Intuitively, distance between $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ should be smaller since both $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ are close to a comonotone relationship. Among these distances above, only Wasserstein metric is consistent with our intuition. Therefore, this example renders another justification of the usefulness of our formulation (Rank-CRSO). □

4 Concluding remarks

In this paper, motivated from the drawback of DRSO with moment uncertainty set, we proposed two new formulations, Linear-CRSO and Rank-CRSO, that capture different dependence structure for DRSO

problems, and derive their dual reformulation. In particular, the dual reformulation is tractable for data-driven Linear-CRSO and Rank-CRSO with ∞-Wasserstein distance. An interesting future direction is to find efficient algorithms for Rank-CRSO with q-Wasserstein distance ($q \in [1,\infty)$). Moreover, various examples and numerical results demonstrate the flexibility and usefulness of our new formulations.

A Appendix

A.1 Dual reformulation in Example 7

Let A be a positive semidefinite matrix. Then $\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1}\mathbb{E}_{\boldsymbol{\mu}}[-\boldsymbol{\xi}^\top A\boldsymbol{\xi}]$ admits a strong dual reformulation

$$\begin{aligned}\min_{\substack{x\in X, y_i\in\mathbb{R}\\ \lambda\geq 0,\Lambda\succeq 0}}\quad & \lambda R_0^2+\langle\Lambda,\Sigma_0\rangle+\frac{1}{N}\sum_{i=1}^N y_i\\ s.t.\quad & \begin{bmatrix}\Lambda+A & \frac{1}{2}\zeta-\Lambda m_0\\ (\frac{1}{2}\zeta-\Lambda m_0)^\top & -\zeta^\top\hat{\xi}^i+m_0^\top\Lambda m_0+y_i\end{bmatrix}\succeq 0,\\ & \qquad\qquad \forall\ 1\leq i\leq N, 1\leq j\leq J.\end{aligned}$$

Proof The result is a consequence of Theorem 1. Observe that $||\xi-\hat{\xi}^i||=\sup_{||\zeta||_*\leq 1}\zeta^\top(\xi-\hat{\xi}^i)$, then by convex programming duality, we have that

$$\begin{aligned}&\max_{\xi\in\Xi}\Big\{-\xi^\top A\xi-\lambda||\xi-\hat{\xi}^i||-(\xi-m_0)^\top\Lambda(\xi-m_0)\Big\}\\ =&\max_{\xi\in\Xi}\inf_{||\zeta||_*\leq 1}\Big\{-\xi^\top A\xi-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\}\\ =&\inf_{||\zeta||_*\leq 1}\max_{\xi\in\Xi}\Big\{-\xi^\top A\xi-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\}.\end{aligned}$$

Hence the constraint

$$y_i\geq\max_{\xi\in\Xi}\Big\{-\xi^\top A\xi-\lambda||\xi-\hat{\xi}^i||-(\xi-m_0)^\top\Lambda(\xi-m_0)\Big\}$$

can be written as

$$\exists||\zeta||_*\leq 1,\ s.t.\ y_i\geq\max_{\xi\in\Xi}\Big\{-\xi^\top A\xi-\lambda\zeta^\top(\xi-\hat{\xi}^i)-(\xi-m_0)^\top\Lambda(\xi-m_0)\Big\},$$

which is further equivalent to

$$\exists||\zeta||_*\leq 1,\ s.t.\begin{bmatrix}\Lambda+A & \frac{1}{2}\lambda\zeta-\Lambda m_0\\ (\frac{1}{2}\lambda\zeta-\Lambda m_0)^\top & -\lambda\zeta^\top\hat{\xi}^i+m_0^\top\Lambda m_0+y_i\end{bmatrix}\succeq 0.$$

Replacing $\lambda\zeta$ by ζ we obtain the result. □

References

1. Agrawal, S., Ding, Y., Saberi, A., Ye, Y.: Price of correlations in stochastic optimization. Operations Research **60**(1), 150–162 (2012)
2. Bayraksan, G., Love, D.K.: Data-driven stochastic programming using phi-divergences. Tutorials in Operations Research pp. 1–19 (2015)
3. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science **59**(2), 341–357 (2013)
4. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research **58**(3), 595–612 (2010)
5. Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. Journal of the Royal Statistical Society. Series B (Methodological) pp. 262–268 (1977)
6. Doan, X.V., Li, X., Natarajan, K.: Robustness to dependency in portfolio optimization using overlapping marginals. Operations Research **63**(6), 1468–1488 (2015)
7. Doan, X.V., Natarajan, K.: On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems. Operations research **60**(1), 138–149 (2012)
8. Dupačová, J.: The minimax approach to stochastic programming and an illustrative application. Stochastics: An International Journal of Probability and Stochastic Processes **20**(1), 73–88 (1987)
9. Erdoğan, E., Iyengar, G.: Ambiguous chance-constrained problems and robust optimization. Mathematical Programming **107**(1-2), 37–61 (2006)
10. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. arXiv preprint arXiv:1505.05116 (2015)
11. Galton, F.: Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland **15**, 246–263 (1886)
12. Gao, R., Kleywegt, A.J.: Distributionally robust stochastic optimization with wasserstein distance. arXiv preprint arXiv:1604.02199 (2016)
13. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. Operations research **58**(4-part-1), 902–917 (2010)

14. Jiang, R., Guan, Y.: Data-driven chance constrained stochastic program. Mathematical Programming pp. 1–37 (2015)
15. Kallenberg, O.: Foundations of modern probability. Springer Science & Business Media (2006)
16. Kendall, M.G.: Rank correlation methods. (1948)
17. Marti, G., Andler, S., Nielsen, F., Donnat, P.: Optimal transport vs. fisher-rao distance between copulas for clustering multivariate time series. arXiv preprint arXiv:1604.08634 (2016)
18. Pearson, K.: Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London **58**, 240–242 (1895)
19. Popescu, I.: Robust mean-covariance solutions for stochastic optimization. Operations Research **55**(1), 98–112 (2007)
20. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2009)
21. Scarf, H., Arrow, K., Karlin, S.: A min-max solution of an inventory problem. Studies in the Mathematical Theory of Inventory and Production **10**, 201–209 (1958)
22. Shapiro, A.: On duality theory of conic linear problems. In: Semi-infinite programming, pp. 135–165. Springer (2001)
23. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on stochastic programming, volume 9 of mps/siam series on optimization. Philadelphia, PA: SIAM. Modeling and theory (2009)
24. Sklar, M.: Fonctions de répartition à n dimensions et leurs marges. Université Paris 8 (1959)
25. Spearman, C.: The proof and measurement of association between two things. The American journal of psychology **15**(1), 72–101 (1904)
26. Sun, H., Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research (2015)
27. Wang, Z., Glynn, P.W., Ye, Y.: Likelihood robust optimization for data-driven problems. Computational Management Science **13**(2), 241–261 (2016)
28. Žáčková, J.: On minimax solutions of stochastic linear programming problems. Časopis pro pěstování matematiky **91**(4), 423–430 (1966)
29. Zalinescu, C.: Convex analysis in general vector spaces. World Scientific (2002)
30. Zymler, S., Kuhn, D., Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming **137**(1-2), 167–198 (2013)