Data-Driven Robust Optimization with Known Marginal Distributions

Rui Gao, Anton J. Kleywegt
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332,
rgao32@gatech.edu, anton@isye.gatech.edu

A key challenge in decision making under uncertainty is seeking robust solutions against high-dimensional data uncertainty. The joint distribution of the stochastic data can hardly ever be obtained exactly, even when estimation on their one-dimensional marginals is rather accurate. To tackle this issue, existing studies hedge a family of distributions with known marginals, but either allow arbitrary dependence structure of these distributions, which tends to be over-conservative, or impose constraints on the deviation — measured by Kullback-Leibler divergence — of the dependence structure from some nominal model, which may lead to pathological worst-case distributions. We propose a distributionally robust approach, which hedges against a family of joint distributions with known marginals and a dependence structure similar to — with similarity measured by Wasserstein distance — that of a nominal joint distribution (e.g., the empirical distribution or the independent product distribution). Tractability of our new formulation is obtained by a novel constructive proof of strong duality, combining ideas from the theory of multi-marginal optimal transport and a new variational argument. Numerical experiments in portfolio selection and nonparametric density estimation demonstrate how the proposed approach outperforms other benchmark approaches.

Key words: distributionally robust optimization; Wasserstein distance; portfolio optimization; copula; density estimation

1. Introduction

Many decision-making problems involves multi-dimensional, dependent random variables. For example, in financial portfolio selection, the total return and the risk of a portfolio depend not only on the return of each individual financial asset, but also on correlations among different risky financial assets. As another example, in large-scale distributed inventory systems design, the designer needs to understand how the dependence of random demands among subsystems affects the overall system performance. Although the joint distributions of such random variables are important, they can hardly ever be obtained accurately in practice.

In contrast, the one-dimensional marginal distribution of each single random variable is often much easier to obtain or estimate, as illustrated by the following three data availability regimes.

(I) *Only joint data are available.* Using a relatively small amount of data, the one-dimensional marginal distribution of each random variable can often be estimated accurately. Moreover, tools for statistical inference regarding one-dimensional distributions are in general more comprehensive than that regarding high-dimensional distributions.

(II) *Only marginal data are available.* This occurs often in a large complex distributed system, in which data sharing among components is limited due to high communication cost or system configuration. For this reason, the joint distribution is often assumed to be an independent product distribution. For example, in many facility location problems, demands at different locations are assumed to be independent.

(III) *Besides joint data, extra marginal data are available for one or more random variables.* When the data streams of different random variables are collected with different frequencies, the decision maker may have more data on the marginal distributions than on the joint distribution. Consider the example in Hall and Neumeyer (2006), in which the decision maker wants to measure the dependence between the lengths of delay of two nonstop flights A and B from Los Angeles to Sydney. One flight operates daily, while the other operates on Mondays, Wednesdays, and Saturdays. Thus, we have joint data on the lengths of delay of the two flights on the days of week when they both operate, and on the remaining days we have additional data on the length of delay of the flight that operates daily.

Based on the discussion above, the central question we want to answer in this paper is:

How to find robust solutions when the joint distribution of random variables are not known exactly but estimates of marginal distributions are relatively accurate?

We next describe two existing approaches to tackle this question and point out their potential issues. In Section 1.1, we present the classical approach using Copula theory and distributions with given marginals, and in Section 1.2, we present a KL-divergence-based distributionally robust approach.

1.1. Copula Theory and Distributions with Known Marginals

Copula theory (Nelsen 2013, Joe 2014) provides a unified way to model the multivariate dependence that is applicable to all the three data availability regimes above. It plays an increasingly important role in many areas, including finance, high-dimensional statistics, and machine learning. A *copula* is a multivariate distribution with all univariate marginals being uniformly distributed on $[0,1]^K$. The seminal Sklar's theorem (Sklar 1959) states that, for every multivariate joint distribution function $F^{\boldsymbol{\mu}}$ with marginal distributions $\{F_k\}_{k=1}^K$, there exists a probability distribution function $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ on $[0,1]^K$, such that

$$F^{\boldsymbol{\mu}}(\xi_1,\ldots,\xi_k) \;=\; \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(F_1(\xi_1),\ldots,F_k(\xi_k)), \quad \forall\, \xi \in \Xi. \tag{1}$$

Such $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ is unique if the marginals are continuous. Conversely, any copula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and marginal distributions $\{F_k\}_k$ together define a K-dimensional joint distribution through (1). This result is phenomenal since it suggests that the analysis of the dependence structure of a multivariate joint distribution can be separated from knowledge of the marginal distributions. For a detailed illustration on constructing copula, we refer to Section 2.1.

Using copula theory, the uncertainty of the joint distribution all boils down to uncertainty of the copula, provided that the marginal distributions are known. A classical approach to address the central question under data availability regime (II) is formulating a minimax problem which hedges against all probability distributions $\mathcal{P}(\Xi)$ on Ξ with the given marginals:

$$\min_{x\in X}\sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}}\left\{\mathbb{E}_{\boldsymbol{\mu}}[\Psi(x,\boldsymbol{\xi})]:\ \boldsymbol{\mu}\text{ has marginals }\{F_k\}_{k=1}^K\text{ and copula }\boldsymbol{\mathcal{C}}\right\}, \tag{2}$$

where x is the decision variable in a feasible set X, and $\Psi: X\times\Xi\to\mathbb{R}$ is the cost function that depends on both the decision x and the random variable $\boldsymbol{\xi}$; $\boldsymbol{\xi}$ has distribution $\boldsymbol{\mu}$ on $\Xi\subset\mathbb{R}^K$, whose value is not known before the decision is made, but its marginal distribution functions $F_1,\ldots,F_k$ are given; and $\mathfrak{C}$ is the set of all copulas on $[0,1]^K$. Such an approach can be traced back at least to Hoeffding (Hoeffding 1940) and Fréchet (Fréchet 1960), who considered the extremes and bounds of (2). Since then, this approach has been extensively studied and applied to many operations management problems (Natarajan et al. 2009, Agrawal et al. 2012, Doan and Natarajan 2012). We refer to Joe (1997) and Benes and Stepán (2012) for a thorough study on this topic.

However, the above worst-case approach (2) does not consider any information at all regarding the joint distribution (such as the joint data in regime (I) and (III)), and thus conceivably, its worst-case distribution often involves fully correlated (i.e., comonotonic or counter-monotonic) components, which may be too extreme for many practical applications. Consider the following example.

Example 1 (Over-conservative worst-case copula). Consider the life insurance model described in Dhaene and Goovaerts (1997), Müller (1997). Each individual risk $\boldsymbol{\xi}_k$ has a two point distribution with $\mathbb{P}(\boldsymbol{\xi}_k=0)=p_k$ and $\mathbb{P}(\boldsymbol{\xi}_k=\alpha_k)=1-p_k$, where α_k represents the value of the k-th claim, p_k denotes the survival probability of the k-th individual, and $p_1\le\cdots\le p_K$. Suppose the function $\Psi_x(\cdot):=\Psi(x,\cdot)$ is supermodular[1], for example, the stop-loss $\max\{0,\sum_{k=1}^K\boldsymbol{\xi}_k-t\}$ of aggregate risks $\sum_{k=1}^K\boldsymbol{\xi}_k$ for some $t>0$. The worst-case copula of (2) is comonotonic[2], and implies that for the corresponding worst-case distribution $\boldsymbol{\mu}^*$, it holds that

$$\mathbb{P}_{\boldsymbol{\mu}^*}[\boldsymbol{\xi}_{k+1}=0|\boldsymbol{\xi}_k=0]\quad=\quad 1,\quad k=1,\ldots,K-1,$$

which means that the death of an individual implies the deaths of all individuals with smaller survival probabilities. In particular, when $p_1=\cdots=p_K$, $\boldsymbol{\mu}^*$ has only two possible scenarios: either all individuals are alive or they all die. Unless the insurance is for some catastrophe, this worst-case distribution seems to be unrealistic, since the dependence of mortality rates among individuals cannot be so strong.

1.2. KL-Divergence-Based DRO with Known Marginals

To overcome the over-conservativeness of (2) and make a better use of joint data, it is natural to restrict $\boldsymbol{\mathcal{C}}$ to a smaller set. Indeed, using the idea from distributionally robust optimization (DRO), recent research considers balls of copulas that are close to some nominal copula $\boldsymbol{\mathcal{C}}^0$ in the sense of Kullback-Leibler (KL) divergence[3] (Dey et al. 2015, Glasserman and Yang 2016, Lam 2017, Dhara et al. 2017):

$$\min_{x\in X}\sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}}\left\{\mathbb{E}_{\boldsymbol{\mu}}[\Psi(x,\boldsymbol{\xi})]:\ \boldsymbol{\mu}\text{ has marginals }\{F_k\}_{k=1}^K\text{ and copula }\boldsymbol{\mathcal{C}},\ KL(\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{C}}^0)\leq\rho\right\},\tag{3}$$

possibly with some additional constraints. However, in a data-driven setting, this approach has limitations as shown by the following example.

Example 2 (KL divergence ball is not suitable for data-driven problem). Consider the nominal distribution is given by $N=30$ i.i.d. observations from a Gaussian distribution. Suppose that we use this empirical distribution as the nominal distribution, then the KL divergence ball $\{\boldsymbol{\mu}:KL(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}},\boldsymbol{\mathcal{C}}^0)\leq\rho\}$ only contains distributions whose support is a subset of the nominal distribution, as indicated by the left image in Fig. 1. However, observe that with probability one, any two data points do not have identical coordinates in either dimension. Hence, if we also consider constraints on the marginals (3), then with probability one, the KL ball is a singleton containing only the empirical distribution itself. To avoid this pathological behavior, one possible remedy is to partition the space into a finite number of bins, such that each bin consists of sufficiently many empirical points. Nevertheless, there is no general guidance on how to make the partition, and it is problematic for high dimensional problems, when the number of data points is less than the dimension of the random variables.

Example 2 demonstrates that formulation (3) is not suitable for high-dimensional data-driven problems. In the next subsection, we propose to use Wasserstein distance instead of KL divergence.

1.3. Our Approach: Wasserstein-Distance-Based DRO with Known Marginals

Motivated by recent progress in distributionally robust stochastic optimization with Wasserstein distance (Esfahani and Kuhn 2015, Gao and Kleywegt 2016, Blanchet and Murthy 2016), we consider all distributions whose associated copula is close to some nominal copula $\boldsymbol{\mathcal{C}}^0$ in Wasserstein distance. More specifically, when the joint data is available (corresponding to data-availability

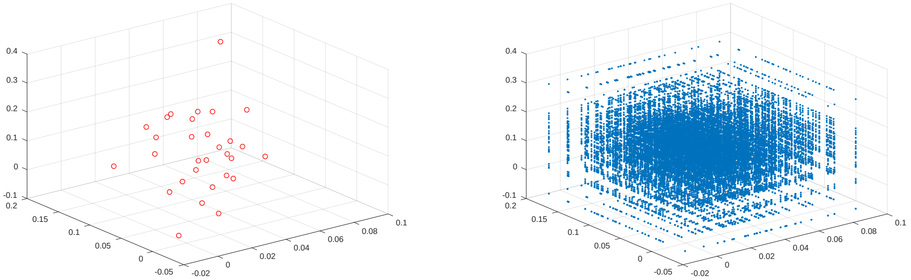

Figure 1 Supports of distributions within a KL divergence ball (3) and a Wasserstein ball (4)

regimes (I) and (III)), we set $\boldsymbol{\mathcal{C}}^0$ to be the empirical copula, and when there is not joint data (corresponding to data-availability regimes (II)), we set $\boldsymbol{\mathcal{C}}^0$ to be the independent copula. Let $W_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}},\boldsymbol{\mathcal{C}}^0)$ denote the p-Wasserstein distance ($p \geq 1$) between $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and $\boldsymbol{\mathcal{C}}^0$. (A more detailed explanation on Wasserstein distance is provided in Section 2.2). Consider the following problem

$$\min_{x\in X} \sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\Psi(x,\boldsymbol{\xi})] : \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}},\ W_p(\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{C}}^0) \leq \rho \right\}, \tag{4}$$

where $\rho > 0$. From the modeling point of view, the advantages of using Wasserstein distance are two-fold.

(i) For copulas of distributions with highly correlated components, Wasserstein distance yields a more intuitive quantitative relationship (Gao and Kleywegt 2017), as illustrated by the following example.

Example 3. Table 1 shows various distances between copulas of Gaussian distributions $\boldsymbol{\mu}_1 = \mathcal{N}(0,[1,0.5;0.5,1])$, $\boldsymbol{\mu}_2 = \mathcal{N}(0,[1,0.99;0.99,1])$, and $\boldsymbol{\mu}_3 = \mathcal{N}(0,[1,0.9999;0.9999,1])$.

Table 1 Distances between copulas of Gaussian distributions

Distances	Fisher-Rao	KL	Burg entropy	Hellinger	Bhattacharya	TV	2-Wasserstein
$\boldsymbol{\mathcal{C}}^{\mu_1}$, $\boldsymbol{\mathcal{C}}^{\mu_2}$	2.77	22.56	1.48	0.69	0.65	2.45	0.15
$\boldsymbol{\mathcal{C}}^{\mu_2}$, $\boldsymbol{\mathcal{C}}^{\mu_3}$	3.26	47.20	1.81	0.75	0.81	4.42	0.03

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 comonotonic distribution. Among the distances above, only Wasserstein metric is consistent with our intuition.

(ii) When the nominal copula is an independent copula (data-availability regime (II)), Wasserstein distance defines a new measure of dependence, and is closely related to Spearman's ranking correlation coefficient. This is established in Section 2.2.

1.4. Our contributions

1. We introduce a Wasserstein-distance-based DRO formulation (4) for decision making under certainty with known marginals, and derive a tractable dual reformulation (Theorem 1). This generalizes the duality results in Kellerer (1984) and Rachev (1985), in which only the marginal constraints are considered, and also generalizes the results in Esfahani and Kuhn (2015) and Gao and Kleywegt (2016), in which only the Wasserstein constraint is considered. Our proof technique combines ideas from a refined constructive approach developed in Gao and Kleywegt (2016), a new variational argument, and the theory of multi-dimensional Monge-Kantorovich optimal transport problem (Rachev and Rüschendorf 1998, Gangbo and Swiech 1998).
2. For a data-driven problem in which the nominal model is the empirical copula, we show that when the objective function Ψ is a piecewise-linear convex function of the random variables, with properly chosen Wasserstein distance, the size of the convex program reformulation of the inner maximization of (4) only linearly depends on the dimension of the random variable, even though the support of the worst-case distribution can contain exponentially many points (Corollary 1). This greatly improves the scalability of our approach.
3. We test the performance of our formulation on two problems. The first is a mean-CVaR portfolio selection problem (Section 4.1), whose parameters are calibrated using real data. The numerical results show superior performance of our approach in high dimension, as opposed to sample average approximation and distributionally robust formulation with only Wasserstein constraints. The second is nonparametric copula density estimation (Section 4.2). Our formulation suggests a novel estimation method. Numerical result on a real dataset illustrates promising results of our approach when the sample size is much less the dimension of the parameters.

2. Copulas and Wasserstein Distance between Copulas

In this section, we describe how to use Wasserstein distance to describe the similarity between dependence structures of distributions. In Section 2.1, we review some results on copula theory, and describe how to construct copula in data-driven problems. In Section 2.2, we introduce the Wasserstein distance between copulas, and investigate its properties.

2.1. Copula and Subcopula in Data-Driven Problems

In the introduction, we have mentioned that the copula is unique for a multivariate continuous distribution. However, in many data-driven problems, the nominal distribution is often finite-supported, which raises the question on the non-uniqueness of copula. To resolve this issue, we consider a slightly general notion called *subcopula*. For ease of exposition, we do not distinguish a probability distribution and its cumulative distribution function as its meaning should be clear from the context. For example, for a distribution $\boldsymbol{\mathcal{C}}$ on $[0,1]^K$, $\boldsymbol{\mathcal{C}}(u)$ is equivalent to $\boldsymbol{\mathcal{C}}([0,u_1]\times\cdots\times[0,u_K])$. Recall that the *support* of a distribution $\boldsymbol{\mu}$ is the complement of the largest open set which has $\boldsymbol{\mu}$-measure zero.

Definition 1 (Subcopula and Copula). A K-dimensional *subcopula* $\boldsymbol{\mathcal{C}}$ is a joint distribution with the following properties:

(i) For all $1 \le k \le K$, the k-th marginal distribution of $\boldsymbol{\mathcal{C}}$, denoted by $\boldsymbol{\mathcal{C}}_k$, has support $\operatorname{supp} \boldsymbol{\mathcal{C}}_k \subset [0,1]$.

(ii) $\boldsymbol{\mathcal{C}}_k(u) = u$ for all $u \in \operatorname{supp} \boldsymbol{\mathcal{C}}_k$.

A K-dimensional subcopula $\boldsymbol{\mathcal{C}}$ is called a K-dimensional *copula* if $\operatorname{supp} \boldsymbol{\mathcal{C}}_k = [0,1]$ for all $1 \le k \le K$.

We next restate Sklar's theorem in terms of subcopula.

Sklar's Theorem. *Let $\boldsymbol{\mu}$ be a K-dimensional distribution on Ξ with marginal distribution functions $F_1,\ldots,F_K$. Then there exists a unique K-subcopula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ such that for all $\xi \in \Xi$,*

$$\boldsymbol{\mu}(\xi_1,\ldots,\xi_K) = \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}\big(F_1(\xi_1),\cdots,F_K(\xi_K)\big),$$

and $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ is a copula if the F_k's are all continuous. Conversely, for any subcopula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and marginal distribution functions $F_1,\ldots,F_K$, the equation above defines a K-dimensional distribution $\boldsymbol{\mu}$ with marginal distributions $F_1,\ldots,F_K$.

Sklar's theorem indicates that the dependence structure of a multivariate distribution is fully characterized by a unique subcopula, which becomes a copula if the marginal distributions are continuous. If we denote the inverse cumulative distribution function of each marginals by F_k^{-1}, then $\boldsymbol{\mathcal{C}}$ can be computed through the formula

$$\boldsymbol{\mathcal{C}}(u_1,\ldots,u_K) = H\big(F_1^{-1}(u_1),\ldots,F_K^{-1}(u_k)\big).$$

We here list some commonly used subcopulas and copulas.

Example 4 (Empirical copula). Let $\frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$ be an empirical distribution, and $\hat{F}_k^{-1}$ be the inverse cumulative empirical distribution of the k-th marginal. The empirical copula (Deheuvels 1979, Tsukahara 2005) is defined by

$$\hat{\boldsymbol{\mathcal{C}}}(u) := \frac{1}{N}\sum_{i=1}^N \prod_{k=1}^K \mathbb{1}\big\{\hat{\xi}_i^k \le \hat{F}_k^{-1}(u_k)\big\}.$$

Thus, empirical copula can be viewed as the empirical distribution of the rank transformed data. Note that empirical copula is a subcopula but not a copula, since $\operatorname{supp} \boldsymbol{\mathcal{C}}_k \subset \{\frac{i}{N} : 1 \le i \le N\}$.

Example 5 (Independent, comonotonic, and counter-monotonic copulas).

- If $\boldsymbol{\xi}$ has mutually independent components, then it has copula $\boldsymbol{\mathcal{C}}(u) = \prod_{k=1}^{K} u_k$.
- If $\boldsymbol{\xi}$ has comonotonic components, i.e., $\boldsymbol{\xi} = (F_1^{-1}(U), \ldots, F_K^{-1}(U))$ for some distribution functions $\{F_k\}_{k=1}^K$ and a uniformly distributed random variable U on $[0,1]$, then $\boldsymbol{\mathcal{C}}(u) = \min_{1 \le k \le K} u_k$.
- If $K = 2$ and $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ are counter-monotonic, i.e., $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (F_1^{-1}(U), F_2^{-1}(1-U))$ for some distribution functions F_1, F_2 and a uniformly distributed random variable U on $[0,1]$, then $\boldsymbol{\mathcal{C}}(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$.

We next illustrate on how to construct a subcopula using the dataset described in data-availability regime (III) in the introduction.

Example 6 (Construction of an Empirical Copula). The joint data of number of minutes delay for two flights on the days of week that they both operate are:

$$(30,4), (-1,0), (-5,7), (12,13), (10,0), (-5,20), (0,15), (32,58), (15,85), (30,45),$$
$$(26,30), (6,23), (40,55), (3,40), (0,-8), (11,12), (7,13), (-5,9), (-11,6), (-10,-20).$$

The additional marginal data of number of minutes delay for the more frequent flight are:

$$20, 4, 5, 48, -30, -10, -22, -3, 80, -23, 0, 26, 10, 90, 90, 24, 30, 45, 17, 35, -10, -1, 30, 5, 18, 0, 40, 16, 6.$$

We denote the joint data by $\{(\hat{\xi}_1^i, \hat{\xi}_2^i)\}_{i=1}^N$, and the extra marginal data by $\{\hat{\xi}_1^{N+j}\}_{j=1}^M$. The empirical copula is constructed in two steps. In the first step, we use all the marginal data information, i.e., $\{\hat{\xi}_1^i\}_{i=1}^{N+M}$ and $\{\hat{\xi}_2^i\}_{i=1}^N$ to estimate the marginal distributions $F_1(\xi_1)$ and $F_2(\xi_2)$. For example, we can simply use empirical cumulative distribution function, or a linear interpolation of the empirical cumulative distribution function. Using the estimated marginal distribution functions, the original joint data set is converted to $(\hat{u}_1^i, \hat{u}_2^i) = (F_1(\hat{\xi}_1^i), F_2(\hat{\xi}_2^i))$, $i = 1, \ldots, N$. Then in the second setup, we estimate the copula density function $c(u_1, u_2)$ using the converted joint dataset $\{(\hat{u}_1^i, \hat{u}_2^i)\}_{i=1}^N$. The scatter plots of the empirical distribution and empirical copula are shown in Figure 2.

2.2. Wasserstein Distance between (Sub)Copulas

Let d be a metric on $[0,1]^K$. In the case of empirical copula, d can be viewed as the distance between two relative rankings. The Wasserstein distance between two subcopulas $\boldsymbol{\mathcal{C}}$, $\boldsymbol{\mathcal{C}}^0$ is defined as follows.

Definition 2 (Wasserstein distance). Let $p \in [1, \infty)$. The p-Wasserstein distance $W_p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0)$ between $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0 \in \mathcal{P}([0,1]^K)$ (under metric d) is defined by

$$W_p^p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0) := \min_{\boldsymbol{\gamma} \in \mathcal{P}([0,1]^{2K})} \left\{ \int_{[0,1]^{2K}} \mathsf{d}^p(u,v) \boldsymbol{\gamma}(du, dv) : \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0 \right\}. \tag{5}$$

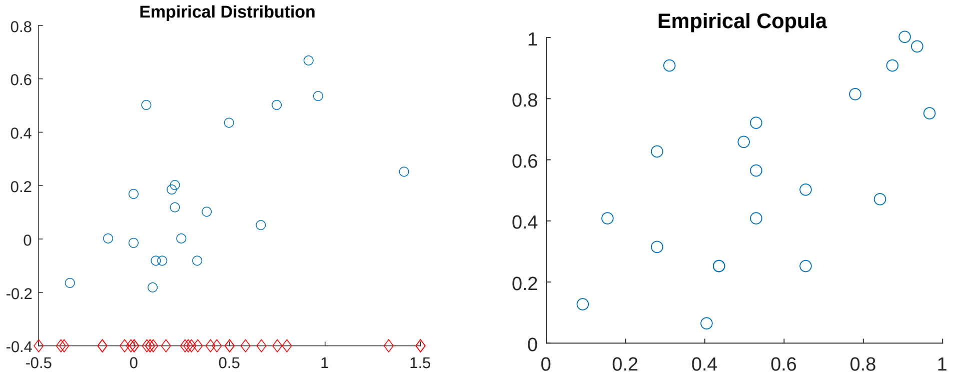

Figure 2 Scatter plots of empirical joint and marginal distributions and empirical copula

Thus, Wasserstein distance between $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0$ is the minimum cost (in terms of d^p) of redistributing mass from $\boldsymbol{\mathcal{C}}$ to $\boldsymbol{\mathcal{C}}^0$. Wasserstein distance is a natural way of comparing two distributions when one is obtained from the other by perturbations.

The expression (5) is written in terms of the integration on $[0,1]^K$. With changing of variables, it can be equivalently represented using integration on the data space Ξ. Let $\boldsymbol{\mu}, \boldsymbol{\nu}$ be two distributions with the same marginals $\{F_k\}_k$, and denote their copulas by $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and $\boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}}$. We define

$$\mathsf{d}_F(\xi,\zeta) := \liminf_{d(\xi^m,\xi), d(\zeta^m,\zeta) \overset{m\to\infty}{\longrightarrow} 0} \mathsf{d}\big((F_1(\xi_1^m),\dots,F_K(\xi_K^m)),\,(F_1(\zeta_1^m),\dots,F_K(\zeta_K^m))\big).$$

It follows that d_F is lower semi-continuous, and d_F is a premetric (Aldrovandi and Pereira 1995), i.e., $\mathsf{d}_F \geq 0$ and $\mathsf{d}_F(\xi,\xi) = 0$. With these definitions, $W_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}})$ can be equivalently represented as

$$W_p^p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}}) = \min_{\gamma \in \mathcal{P}(\Xi\times\Xi)} \left\{ \int_{\Xi^2} \mathsf{d}_F^p(\xi,\zeta)\gamma(d\xi,d\zeta) : \gamma \text{ has marginals } \boldsymbol{\mu},\boldsymbol{\nu} \right\}.$$

Now let us consider the case when the nominal copula $\boldsymbol{\mathcal{C}}^0$ is the independent subcopula Π, which corresponds to the data-availability regime (II) described in the introduction. In this case, the Wasserstein distance $W_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \Pi)$ measures the deviation of $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ away from an independent distribution, and thus can be viewed as a measure of dependence of random variables with joint distribution $\boldsymbol{\mu}$. In particular, when $K = 2$ and $\Pi(u) = u_1u_2$, with a special choice of d, $W_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \Pi)$ reduces to Schweizer and Wolffs L^1-based measure of dependence (Schweizer and Wolff 1981), defined as $\int_0^1\int_0^1 |\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(u_1,u_2) - u_1u_2| du_1 du_2$.

Proposition 1. *Suppose*

$$\mathsf{d}\big((u_1,u_2),(v_1,v_2)\big) = \begin{cases} |u_1 - v_1|, & if\ u_2 = v_2, \\ +\infty, & o.w. \end{cases}, \text{ or } \begin{cases} |u_2 - v_2|, & if\ u_1 = v_1, \\ +\infty, & o.w. \end{cases}$$

Let $K = 2$. Then for any distribution $\boldsymbol{\mu}$ with copula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$, it holds that

$$W_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \Pi) = \int_0^1\int_0^1 |\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(u_1,u_2) - u_1u_2| du_1 du_2.$$

We remark that Schweizer and Wolffs' measure of dependence is closely related to Spearman's rank correlation coefficient, which can be written as $\int_0^1 \int_0^1 \big(\boldsymbol{\mathcal{C}}(u_1,u_2) - u_1 u_2\big) du_1 du_2$.

If we set d to be the ℓ_1-norm, then $W_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^0)$ defines a new measure of dependence which satisfies Rényi's axioms on measure of dependence (Rényi 1959, Schweizer and Wolff 1981).

Proposition 2. *Suppose*

$$\mathsf{d}(u,v) = ||u - v||_1, \quad u, v \in [0,1]^2.$$

Let $(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *be two random variables with continuous distribution* $\boldsymbol{\mu} \in \mathcal{P}([0,1]^K)$*, define*

$$\omega(\boldsymbol{\xi}, \boldsymbol{\zeta}) \;:=\; 12 \cdot W_1(\boldsymbol{\mu}, \Pi).$$

Then $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *defines a measure of dependence that satisfies Rényi's axioms:*

(i) $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta}) = \omega(\boldsymbol{\zeta}, \boldsymbol{\xi})$.

(ii) $0 \le \omega(\boldsymbol{\xi}, \boldsymbol{\zeta}) \le 1$.

(iii) $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta}) = 0$ *if and only if* $\boldsymbol{\xi}$ *and* $\boldsymbol{\zeta}$ *are independent.*

(iv) $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta}) = 1$ *if and only if each of* $\boldsymbol{\xi}$ *is a.s. a strictly monotone function of the other.*

(v) If f *and* g *are strictly monotone a.s. on* $\operatorname{Ran} \boldsymbol{\xi}$ *and* $\operatorname{Ran} \boldsymbol{\zeta}$ *respectively, then* $\omega(f(\boldsymbol{\xi}), g(\boldsymbol{\zeta})) = \omega(\boldsymbol{\xi}, \boldsymbol{\zeta})$.

(vi) If the joint distribution of $\boldsymbol{\xi}$ *and* $\boldsymbol{\zeta}$ *is bivariate normal with correlation coefficient* ρ*, then* $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *is a strictly increasing function of* $|\rho|$.

(vii) If $(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *and* $(\boldsymbol{\xi}^m, \boldsymbol{\zeta}^m)$*,* $m = 1, 2, \cdots$*, are pairs of random variables with joint distribution* $\boldsymbol{\mu}$ *and* $\boldsymbol{\mu}^m$ *respectively, and if the sequence* $\boldsymbol{\mu}^m$ *converges weakly to* $\boldsymbol{\mu}$*, then* $\lim_{m\to\infty} \omega(\boldsymbol{\xi}, \boldsymbol{\mu}) = \omega(\boldsymbol{\xi}, \boldsymbol{\mu})$.

3. Dual reformulation

In this section, we derive a dual reformulation for the inner maximization of problem (4). For ease of notation, we suppress variable x of Ψ. Set

$$v_P \;:=\; \sup_{\boldsymbol{\mathcal{C}} \in \mathfrak{C}} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\Psi(x, \boldsymbol{\xi})] : \; \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}},\; W_p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0) \le \rho \right\}. \tag{6}$$

We assume Ψ is upper semicontinuous on Ξ, and satisfies the growth condition $\sup_{\xi \in \Xi} \frac{\Psi(\xi)}{\mathsf{d}_F^p(\xi, \zeta_0)} < \infty$ for some $\zeta_0 \in \Xi$. Our main result is the following strong duality theorem.

Theorem 1 **(Strong duality)**. *Let* $\boldsymbol{\nu}$ *be a distribution with marginals* $\{F_k\}_{k=1}^K$ *and copula* $\boldsymbol{\mathcal{C}}^0$*. Let* Ξ_k *be the projection of* Ξ *onto the* k*-th marginal component. Then problem* (6) *has a strong dual problem*

$$v_D \;:=\; \inf_{\substack{\lambda \ge 0 \\ f_k \in B(\Xi_k)}} \left\{ \lambda \rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) + \int_{\Xi} \sup_{\xi \in \Xi} \Big[\Psi(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big] \boldsymbol{\nu}(d\zeta) \right\}.$$

Before diving into the proof, we outline the proof idea as follows. To start with, it is straightforward to establish the weak duality using Lagrangian and properties of marginal distribution (Lemma 1). However, the difficulties in proving strong duality lie in the non-compactness of the data space Ξ, and the semi-infinite marginal and Wasserstein constraints. To obtain the strong duality, we first assume certain compactness and continuity assumptions. Under such assumptions, we show the existence of a dual minimizer using convexification trick (see, e.g., Rachev (1985), Gangbo and Swiech (1998)) in the theory of multi-marginal optimal transport (Lemma 2). Next, we derive the first-order optimality condition at the dual minimizer, which helps to construct a primal optimal solution (Lemma 3). Finally using some limiting argument, we relax the continuity and the compactness assumption and thus complete the proof of Theorem 1. We only provide the proof of Lemma 3 here, and proofs of other lemmas and measurability of the integrand involved in the dual program are presented in the Technical Appendix.

Lemma 1 **(Weak duality)**. $v_P \leq v_D$.

Lemma 2 **(Existence of dual minimizer)**. *Assume that Ξ is compact and Ψ and d_F are Lipschitz continuous on Ξ. Then there exists a dual minimizer.*

Lemma 3 **(Strong duality under compactness and continuity assumption)**. *Assume that Ξ is compact and Ψ and d_F are Lipschitz continuous on Ξ. Then $v_P = v_D$.*

Proof of Lemma 3. We start with establishing the first-order optimality condition of the dual problem. We perform a variational analysis on the dual objective function at $(\lambda^*, \{f_k^*\}_k)$. For each $1 \leq k \leq K$, let $\{g_{km}\}_{m=1}^{\infty}$ be a Schauder basis of $B(\Xi_k)$. For any $n \in \mathbb{Z}_+$, we define a function

$$\Phi_n(\lambda, \epsilon, \zeta) := \sup_{\xi \in \Xi} \Big\{ \Psi(\xi) - \sum_k f_k^*(\xi_k) - \sum_k \sum_{m=1}^{n} \epsilon_{km} g_{km}(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\}. \tag{7}$$

By Lemma 4 in Appendix, Φ is random lower semi-continuous. Moreover, for all $\zeta \in \Xi$, $\Phi(\cdot,\cdot,\cdot,\zeta)$ is a convex function on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{nK}$. We further define

$$\begin{aligned} h_n(\lambda, \epsilon) = \lambda \rho^p &+ \sum_{k=1}^{K} \int_{\Xi_k} f_k(t) F_k(dt) + \sum_{k=1}^{K} \sum_{m=1}^{n} \int_{\Xi_k} \epsilon_{km} g_{km}(t) F_k(dt) \\ &+ \int_{\Xi} \sup_{\xi \in \Xi} \Big[\Psi(\xi) - \sum_{k=1}^{K} f_k^*(\xi_k) - \sum_{k=1}^{K} \sum_{m=1}^{n} \epsilon_{km} g_{km}(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big] \boldsymbol{\nu}(d\zeta). \end{aligned}$$

Then by generalized Moreau-Rockafellar theorem (see, e.g., Theorem 7.47 in Shapiro et al. (2009)), for any $(\lambda, \epsilon) \in \operatorname{dom} h_n$ it holds that

$$\partial h_n(\lambda, \epsilon) = \Big(\rho, \Big[\int_{\Xi_k} g_{km}(t) F_k(dt) \Big]_{\substack{1 \leq k \leq K \\ 1 \leq m \leq n}} \Big)^{\top} - \int_{\Xi} \partial_{\lambda,\epsilon} \Phi_n(\lambda, \epsilon, \zeta) \boldsymbol{\nu}(d\zeta) + \mathcal{N}(\lambda, \epsilon),$$

where $\mathcal{N}(\lambda,\epsilon)$ stands for the normal cone at (λ,ϵ) to the feasible region $\mathbb{R}_+\times\mathbb{R}_+\times\mathbb{R}^{nK}$. Furthermore, it follows from Theorem 2.4.18 in Zalinescu (2002) that

$$\partial_{\lambda,\epsilon}\Phi(\lambda,\epsilon,\zeta)=\operatorname{conv}\Bigg\{\bigg(\mathsf{d}^p_F(F(\xi(\zeta)),F(\zeta)),[g_{km}(\xi_k(\zeta))]_{\substack{1\le k\le K\\1\le m\le n}}\bigg)^\top: \\ \xi(\zeta)\in\arg\max_{\xi\in\Xi}\Big[\Psi(\xi)-\sum_{k=1}^K f_k^*(\xi_k)-\lambda\mathsf{d}^p_F(\xi,\zeta)\Big]\Bigg\}.$$

Set

$$T(\zeta):=\arg\max_{\xi\in\Xi}\Big[\Psi(\xi)-\sum_{k=1}^K f_k^*(\xi_k)-\lambda^*\mathsf{d}^p_F(\xi,\zeta)\Big].$$

The first-order optimality condition $0\in\partial h_n(\lambda^*,0)$ implies that there exists $0\le r^*\le\rho$ with $\lambda^*(\rho-r^*)=0$, such that

$$\bigg(r^*,\Big[\int_{\Xi_k}g_{km}(t)F_k(dt)\Big]_{\substack{1\le k\le K\\1\le m\le n}}\bigg)\in\int_{\Xi}\operatorname{conv}\Big\{\Big(\mathsf{d}^p(\xi(\zeta),\zeta),[g_{km}(\xi_k(\zeta))]_{\substack{1\le k\le K\\1\le m\le n}}\Big):\xi(\zeta)\in T(\zeta)\Big\}\boldsymbol{\nu}(d\zeta). \tag{8}$$

We construct a primal optimal solution. (8) suggests that there is a measurable selection $z(\zeta)$ of $\operatorname{conv}\{(\mathsf{d}^p(\xi(\zeta),\zeta),[g_{km}(\xi_k(\zeta))]_{k,m}),\ \xi(\zeta)\in T(\zeta)\}$, such that $\int_\Xi z(\zeta)\boldsymbol{\nu}(d\zeta)=(r^*,[\int_{\Xi_k}g_{km}(t)F_k(dt)]_{k,m})$. Each $z(\zeta)$ can be represented as

$$z(\zeta)=\mathbb{E}_{\boldsymbol{\gamma}^n_\zeta}\big[\big(\mathsf{d}^p(\xi(\zeta),\zeta),[g_{km}(\xi_k(\zeta))]_{1\le k\le K,1\le m\le n}\big)\big],$$

for some finite probability distribution $\boldsymbol{\gamma}^n_\zeta\in\mathcal{P}(T(\zeta))$, and the measurability of $z(\zeta)$ implies the measurability of $\boldsymbol{\gamma}^n_\zeta$ (as a function of ζ). Thus, there exists a probability kernel $\{\boldsymbol{\gamma}^n_\zeta\}_{\zeta\in\Xi}$ such that each $\boldsymbol{\gamma}^n_\zeta$ is a probability distribution on $T(\zeta)$ and satisfies

$$\begin{aligned}r^*&\le\rho,\\ \lambda^*(\rho-r^*)&=0,\\ \int_{\Xi^2}g_{km}(\xi_k(\zeta))\boldsymbol{\gamma}^n_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)&=\int_{\Xi_k}g_{km}(t)F_k(dt),\ \forall 1\le k\le K,1\le m\le n.\end{aligned}\tag{9}$$

Now define a probability measure $\boldsymbol{\mu}^n$ by

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

Then

$$\int_\Xi g_{km}(\xi_k)\boldsymbol{\mu}^n(d\xi)=\int_{\Xi_k}g_{km}(t)F_k(dt),\quad\forall 1\le k\le K,1\le m\le n,$$

due to (9). Since the collection of probability measures $\{\boldsymbol{\mu}^n\}_n$ is tight, by Prokhorov's theorem, there is a convergent subsequence, whose limit is denoted by $\boldsymbol{\mu}^*$. It follows that

$$\int_\Xi g_{km}(\xi_k)\boldsymbol{\mu}^*(d\xi)=\int_{\Xi_k}g_{km}(t)F_k(dt),\quad\forall 1\le k\le K,m\ge 1,$$

that is, $\boldsymbol{\mu}^*$ has marginals $\{F_k\}_k$. Hence $\int_\Xi f_k(\xi_k)\boldsymbol{\mu}^*(d\xi) = \int_{\Xi_k} f_k(t)F_k(dt)$ for all $f_k \in B(\Xi_k)$. In addition, due to (9), we have that $\boldsymbol{\mu}^*$ is primal feasible, and

$$\begin{aligned}
&\int_\Xi \Psi(\xi)\boldsymbol{\mu}^*(d\xi)\\
=&\int_\Xi \big[\Psi(\xi) - \sum_{k=1}^K f_k^*(\xi_k) - \lambda^* \mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\mu}^*(d\xi)\\
&\qquad\qquad\qquad + \int_\Xi \big[\sum_{k=1}^K f_k^*(\xi_k) + \lambda^* \mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\mu}^*(d\xi)\\
=&\int_\Xi \sup_{\xi\in\Xi}\Big[\Psi(\xi) - \sum_{k=1}^K f_k^*(\xi_k) - \lambda^* \mathsf{d}_F^p(\xi,\zeta)\Big]\boldsymbol{\nu}(d\zeta) + \lambda^*\rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k^*(t)F_k(dt)\\
\geq& v_D.
\end{aligned}$$

□

3.1. Data-driven Problem and Size Reduction

Corollary 1. *Suppose* $\Psi(\xi) = \max_{1\leq m\leq M} a^{m\top}\xi + b^m$ *for some* $a^m \in \mathbb{R}^K$ *and* $b^m \in \mathbb{R}$*, and* $\boldsymbol{\nu} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{\delta}_{\hat{\xi}^i}$. *Let* $\Xi_k := \{\hat{\xi}_k^i : \ i = 1,\ldots,N\}$. *Then the dual problem of* (6) *is given by*

$$\begin{aligned}
\inf_{\substack{\lambda\geq 0, f_k^i\in\mathbb{R}\\ y_i\in\mathbb{R}}} \Big\{ &\lambda\rho^p + \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^N f_k^i + \frac{1}{N}\sum_{i=1}^N y^i :\\
&y^i \geq a^{m\top}(\hat{\xi}_1^{j_1},\ldots,\hat{\xi}_K^{j_K}) + b^m - \sum_{k=1}^K f_k^{j_k} - \lambda \mathsf{d}_F^p\big((\hat{\xi}_1^{j_1},\ldots,\hat{\xi}_K^{j_K}),\hat{\xi}^i\big),\\
&\forall 1\leq i\leq N,\ \forall 1\leq j_k\leq N,\ \forall 1\leq k\leq K\Big\}.
\end{aligned}\tag{10}$$

If, in addition, there exists $\{\mathsf{d}_{F,k}\}_k$ *such that*

$$\mathsf{d}_F^p\big((\hat{\xi}_1^{j_1},\ldots,\hat{\xi}_K^{j_K}),\hat{\xi}^i\big) = \sum_{k=1}^K \mathsf{d}_{F,k}(\hat{\xi}_k^{j_k},\hat{\xi}_k^i), \quad \forall i, j_k,\ \forall k,$$

then the above program is equivalent to

$$\begin{aligned}
\inf_{\substack{\lambda\geq 0, f_k^i\in\mathbb{R}\\ y^i, z_k^{im}\in\mathbb{R}}} \Big\{ &\lambda\rho^p + \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^N f_k^i + \frac{1}{N}\sum_{i=1}^N y^i :\ y^i \geq b^m + \sum_{k=1}^K z_k^{im},\ \forall i, m,\\
&z_k^{im} \geq a_k^{m\top}\xi_k^j - f_k^j - \lambda \mathsf{d}_{F,k}(\xi_k^j,\hat{\xi}_k^i),\ \forall i, m, j, k\Big\}.
\end{aligned}\tag{11}$$

Proof of Corollary 1. Formulation (10) follows directly from Theorem 1. Formulation (11) follows from the fact that for any additively separable function $g(\hat{\xi}_1^{j_1},\ldots,\hat{\xi}_K^{j_K}) = \sum_{k=1}^K g_k(\hat{\xi}_k^{j_k})$,

$$\max_{j_1,\ldots,j_K} g(\hat{\xi}_1^{j_1},\ldots,\hat{\xi}_K^{j_K}) = \sum_{k=1}^K \max_j g_k(\hat{\xi}_k^j).$$

□

We remark that (11) indicates that when the metric d_F^p is additively separable, by introducing auxiliary variables z_k^{im}, the original problem admits a reformulation with $MN(K+1)$ constraints, linearly growing in dimension K.

4. Applications

In this section, we discuss two applications.

4.1. Mean-CVaR portfolio selection

We consider a distributionally robust portfolio optimization problem

$$\min_{x\in X}\max_{\boldsymbol{\mu}\in\mathfrak{M}} \quad \mathbb{E}_{\boldsymbol{\mu}}[-x^\top\boldsymbol{\xi}] + c\cdot \mathrm{CVaR}_{\boldsymbol{\mu}}^{\alpha}[-x^\top\boldsymbol{\xi}], \tag{12}$$

where $c>0$, $X := \left\{x\in\mathbb{R}_+^K : \ \sum_{k=1}^K x_k = 1\right\}$ encodes the vectors of weights of K assets without short-selling, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_K)^\top$ is the vector of excessive returns over the risk-free rate, and CVaR is the conditional value-at-risk Rockafellar and Uryasev (2000) under distribution $\boldsymbol{\mu}$. We use the Fama-French three-factor model Fama and French (1993) to model the asset return. The Fama-French three-factor model assumes that the excess return of the k-th asset follows the following three-factor model:

$$\boldsymbol{\xi}_k = b_{k1}\boldsymbol{f}_1 + b_{k2}\boldsymbol{f}_2 + b_{k3}\boldsymbol{f}_3 + \boldsymbol{\epsilon}_k, \quad k=1,\ldots,K,$$

where, the factor $\boldsymbol{f}_1$ are respectively the excess return of the proxy of the market portfolio, which equals the value-weighted return on all NYSE, AMEX and NASDAQ stocks minus the one-month Treasury bill rate; factors $\boldsymbol{f}_2,\boldsymbol{f}_3$ are related to the market capitalizations and and book-to market ratios, more specifically, $\boldsymbol{f}_2$ equals the average return on three small portfolios minus the average return on three big portfolios, and $\boldsymbol{f}_3$ equals the average return on two value portfolios minus the average return on two growth portfolios; b_{k1},b_{k2},b_{k3} are the factor loadings of the k-th stock; and ϵ_k is the idiosyncratic noise independent of the three factors, and independent across the stocks.

The parameters are estimated using the three-year daily data of 30 Industry Portfolios from May 1, 2002 to Aug 29, 2005 (Frech 2017). We borrow the calibration results from Fan et al. (2008) (see Table 2), where the factor loadings (b_{k1},b_{k2},b_{k3}), $k=1,\ldots,K$ are i.i.d. drawn from $Normal(\mu_b,\Sigma_b)$, and once generated, they are fixed as constants throughout simulations. The N-period returns of the three factors $(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3)$ are generated from $Normal(\mu_f,\Sigma_f)$, and the noises are generated from $Gamma(3.3586, 0.1876)$ conditioned on the noise level of at least 0.1950.

Note that the objective function of (12) can be equivalently written as

$$\min_{x\in X,\tau\in\mathbb{R}}\sup_{\boldsymbol{\mu}\in\mathfrak{M}}\left\{\mathbb{E}_{\boldsymbol{\mu}}\left[\max_{1\le m\le M} a_m x^\top\boldsymbol{\xi} + b_m\right]\right\},$$

Table 2 Parameters in the three-factor model

μ_b		Σ_b		μ_f		Σ_f	
0.78282	0.029145	0.023873	0.010184	0.023558	1.2507	-0.034999	-0.20419
0.51803	0.0232873	0.053951	-0.006967	0.012989	-0.034999	0.31564	-0.0022526
0.41003	0.010184	-0.006967	0.086856	0.020714	-0.20419	-0.0022526	0.19303

where $M = 2$, $a_1 = -1$, $a_2 = -1 - c/\alpha$, $b_1 = c$ and $b_2 = c(1 - 1/\alpha)$. We choose $\boldsymbol{\mathcal{C}}_0$ to be the empirical copula in defining $\mathfrak{M}$. In all numerical experiments, we set $\alpha = 0.2$, $c = 10$, $\mathsf{d}(u,v) = ||u - v||_1$. We fix $N = 50$, and vary $K = 10, 50, 100$, corresponding to three regimes $N > K$, $N = K$, and $N < K$. We run the simulation with 200 repetitions. The Wasserstein radius ρ is chosen using hold-out cross validation. More specifically, in each repetition, we generate N-period returns, and the N samples are randomly partitioned into a training dataset with 70% data and a validation set with 30% data. We solve problem (12) using the training dataset for different choices of ρ, and choose the one that has the best out-of-sample performance using validation dataset. Then we resolve problem (12) using the all N samples, and the out-of-sample performance of the optimal solution is evaluated using an independent testing dataset with 10^6 samples.

We compare our approach with two other approaches, sample average approximation (SAA) method, and DRSO with W_1-Wasserstein ball considered in Esfahani and Kuhn (2015), in which there is no constraints on the marginal distributions and the ball is centered at the empirical distribution instead of the copula. Note that our numerical setting is similar to the one in Esfahani and Kuhn (2015), expect that we generate random asset returns based on the three-factor model whose parameters are calibrated using real data. The box plot of the results is shown in Figure 3.

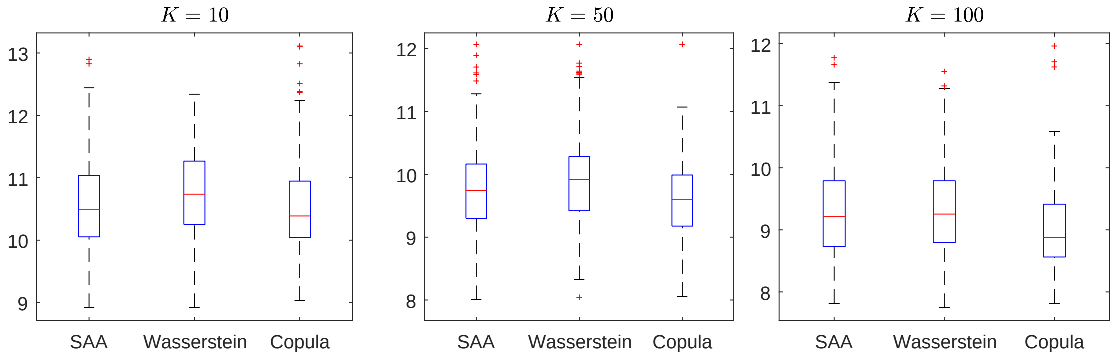

Figure 3 Out-of-sample performances of three approaches

We observe that the DRSO with Wasserstein ball does not have a superior performance over SAA method, and is actually even worse in relatively low dimensional setting when $K \leq N$. Possible explanation of this is that variations of the uncertain asset returns are not that big, so SAA already has a relatively good performance especially in low-dimensional setting, whereas DRSO with Wasserstein ball only provide a conservative solution. Nevertheless, our proposed Copula approach seems to perform better when the dimensional K becomes larger. Note that in our experiments, samples of size $N = 50$ already provide a rather accurate estimate of the one-dimensional marginal distribution. By constraining the marginal distributions and building a ball around the empirical copula, our approach obtain a more robust (comparing to SAA) yet less conservative solution (comparing to Wasserstein ball), and this effect becomes more apparent in high dimensions.

4.2. Nonparametric density estimation with extra marginal data

We focus on the copula density estimation in the second step above, and we are interested in nonparametric estimation. The following setup is based on Qu and Yin (2012). The domain $[0,1]^2$ is partitioned into $M \times M$ rectangle cells with equal size. For each cell $(u_1^{k_1}, u_2^{k_2})$, $k_1, k_2 = 1, \dots, M$, denote by $\boldsymbol{\mathcal{C}}^0_{k_1,k_2}$ the empirical relative frequency of observations $\{(\hat{u}_1^i, \hat{u}_2^i)\}_{i=1}^N$ falling in this cell, and define x_{k_1,k_2} to be the probability mass of this cell that we are going to estimate. Then the maximum likelihood estimation is given by

$$\min_{x \in X} \mathbb{E}_{\boldsymbol{\mathcal{C}}^0}[-\log(x(\boldsymbol{u}))], \tag{13}$$

where

$$X := \left\{ x \in \mathbb{R}_+^{M \times M} : \sum_{k_1} x_{k_1,k_2} = \sum_{k_2} x_{k_1,k_2} = \frac{1}{M} \right\}.$$

In Qu and Yin (2012), it is proposed to consider a total variation penalized likelihood

$$\min_{x \in X} \mathbb{E}_{\boldsymbol{\mathcal{C}}^0}[-\log(x(\boldsymbol{u}))] + \lambda \sum_{k_1,k_2=1}^{M} \sqrt{(x_{k_1+1,k_2} - x_{k_1,k_2})^2 + (x_{k_1,k_2+1} - x_{k_1,k_2})^2}.$$

Here we propose another approach based on our distributionally robust framework. Consider

$$\min_{x \in X} \max_{\boldsymbol{\mathcal{C}} \in \mathfrak{M}} \mathbb{E}_{\boldsymbol{\mathcal{C}}}\big[-\log(x(\boldsymbol{u}))\big], \tag{14}$$

where $\mathfrak{M}$ is a ball of subcopulas centered at $\boldsymbol{\mathcal{C}}_0$. Using our duality result, the problem above can be reformulated as a convex programming

$$\min_{\substack{x \in X, \lambda \geq 0 \\ f_1^{k_1}, f_2^{k_2}, y}} \left\{ \lambda\rho + \frac{1}{M} \sum_{k_1=1}^{M} f_1^{k_1} + \frac{1}{M} \sum_{k_2=1}^{M} f_2^{k_2} + \frac{1}{N} \sum_{i=1}^{N} y_i + \sum_{k_1,k_2=1}^{M} x_{k_1,k_2}^2 : \right.$$
$$\left. y_i \geq -\log(x_{k_1,k_2}) - f_1^{k_1} - f_2^{k_2} - \lambda \cdot ||(u_1^{k_1}, u_2^{k_2}) - (\hat{u}_1^i, \hat{u}_2^i)||_1, \ \forall i, k_1, k_2 \right\}.$$

In our experiment, we use a dataset in Example 6. We compare our approach with total variation penalized likelihood estimation proposed in Qu and Yin (2012), which is, to the best of our knowledge, the only method that fores the marginal constraints on the copula (Many other

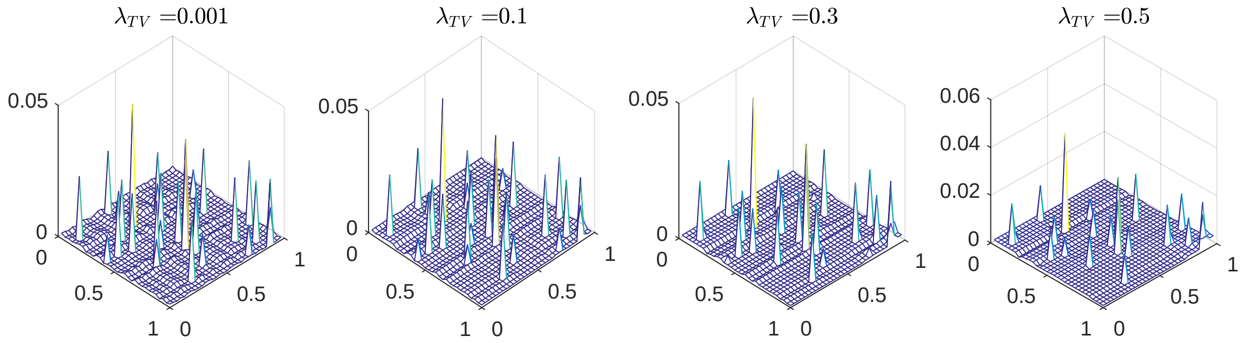

Figure 4 Copula density estimator using TV penalized maximum likelihood

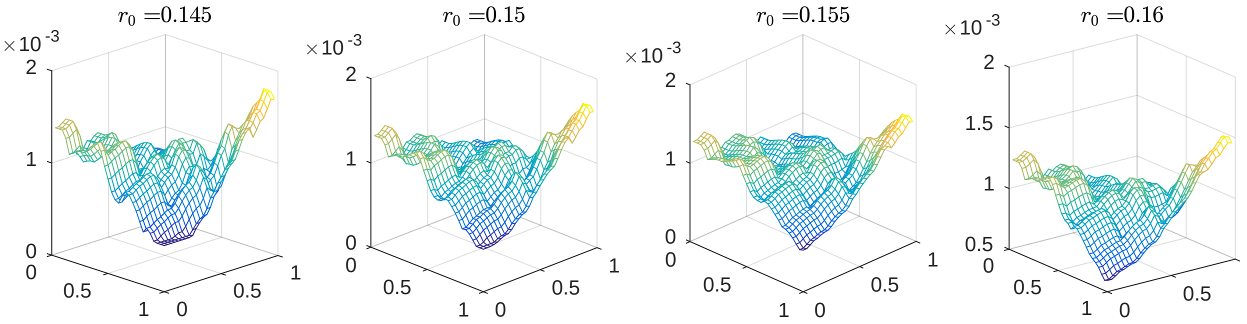

Figure 5 Copula density estimator using Wasserstein-based distributionally robust method

kernel/wavelets-based approach actually do not provide an estimator that satisfies the marginal requirement for a copula). In our experiment, we set $M = 32$. Since the real dataset is very small, we here only provide a qualitative comparison for the copula density estimators.

Figure 4 and 5 show the estimators yielding from the two approaches with different tuning parameters. It is obvious that they differ a lot. In particular, the density estimator using total variation penalized likelihood estimation proposed in Qu and Yin (2012) has disconnected support, which seems unrealistic. In contrast, our density estimator is smoother and seems to be more reasonable using only a small dataset.

5. Concluding remarks

In this paper, we proposed a distributionally robust framework for decision-making under uncertainty when the marginal distributions are fixed. We chose Wasserstein distance to measure the closeness between the considered dependence structure and some nominal model. We used several illustrative examples to show its advantages over previous work on divergence-based approach. Our computational examples on portfolio selection and density estimation show that, for high-dimensional data-driven problems, namely, problems in which the sample size is much less than the number of unknown parameters, our approach outperforms the conventional approaches.

Appendix. Technical Proofs.

Proof of Proposition 1. Given a copula $\boldsymbol{\mathcal{C}}$, set $\boldsymbol{\mathcal{C}}_u$ to be the marginal distribution of $\boldsymbol{v}$ given $\boldsymbol{u}=u$. Let $\mathcal{U}$ be the uniform distribution on $[0,1]$. Under the above condition on d, we have that

$$\omega_{1,\mathsf{d}}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)=\int_0^1 W_1(\boldsymbol{\mathcal{C}}_u,\mathcal{U})du.$$

Then the result follows from the formula for one-dimensional Wasserstein distance Vallender (1974).

□

Proof of Proposition 2. Observe that when choosing ℓ_1-norm, the optimal transportation defining $W_1(\boldsymbol{\mathcal{C}}^M,\Pi)$ can be chosen such that each point is transported only vertically. Then the computational of Wasserstein distance is reduced to the case in Proposition 1, and thus the result follows.

□

Proof of Lemma 1. Observe that for any random vector $(\boldsymbol{\xi},\boldsymbol{\zeta})$ with joint distribution $\boldsymbol{\gamma}\in\mathcal{P}(\Xi\times\Xi)$ and marginals $\boldsymbol{\mu},\boldsymbol{\nu}\in\mathcal{P}(\Xi)$, it holds that

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

where $\boldsymbol{\gamma}_\zeta$ represents the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\zeta}=\zeta$. Also note that $\boldsymbol{\mu}$ has marginal F_k if and only if $\int_\Xi f_k(\xi_k)\boldsymbol{\mu}(d\xi)=\int_{\Xi_k} f_k(t)F_k(dt)$ for all $f_k\in B(\Xi_k)$. With the observations above, using Lagrangian weak duality, we have that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathfrak{M}}\left\{\int_\Xi \Psi(\xi)\boldsymbol{\mu}(d\xi)\right\}\\
&=\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\ \inf_{\substack{\lambda\geq 0\\ f_k\in B(\Xi_k)}}\left\{\int_{\Xi^2}\Psi(\xi)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)+\lambda\rho^p-\lambda\int_{\Xi^2}\mathsf{d}_F^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right.\\
&\qquad\qquad\left.+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)-\int_{\Xi^2}\sum_k f_k(\xi_k)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}\\
&\leq\inf_{\substack{\lambda\geq 0\\ f_k\in B(\Xi_k)}}\left\{\lambda\rho^p+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)\right.\\
&\qquad\qquad\left.+\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\int_{\Xi^2}\left[\Psi(\xi)-\sum_k f_k(\xi_k)-\lambda\mathsf{d}_F^p(\xi,\zeta)\right]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}\\
&\leq\inf_{\substack{\lambda\geq 0\\ f_k\in B(\Xi_k)}}\left\{\lambda\rho^p+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)\right.\\
&\qquad\qquad\left.+\int_\Xi\sup_{\xi\in\Xi}\left[\Psi(\xi)-\sum_k f_k(\xi_k)-\lambda\mathsf{d}_F^p(\xi,\zeta)\right]\boldsymbol{\nu}(d\zeta)\right\}.
\end{aligned}$$

□

Proof of Lemma 2. We claim that there exists $M > 0$ such that

$$v_D = \inf_{\substack{0 \le \lambda \le M \\ f_k \in B(\Xi_k)}} \left\{ \lambda \rho^p + \sum_{k=1}^{K} \int_{\Xi_k} f_k(t) F_k(dt) + \int_{\Xi} \sup_{\xi \in \Xi} \Big[\Psi(\xi) - \sum_{k=1}^{K} f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big] \boldsymbol{\nu}(d\zeta) \right\}. \tag{15}$$

Indeed, according to the assumption on Ψ, there exists $M > 0$ such that $\Psi(\xi) \le M$ for all $\xi \in \Xi$ and thus by choosing $\alpha = \lambda = 0$ and $f_k \equiv 0$, we obtain that $v_D \le M$. On the other hand, fixing $f_k \equiv 0$, the dual objective tends to infinity as $\lambda \to \infty$. Hence the claim holds.

For any feasible solution $(\lambda, \{f_k\}_k)$ of (15) such that the dual objective is finite, we are going to define a modification $(\lambda, \{\bar{f}_k\}_k)$ which yields a dual objective value no worse than $(\lambda, \{f_k\}_k)$, but also has a nicer continuity property. The technique used here is the convexification trick. Setting

$$\Phi(\lambda, \zeta) := \sup_{\xi \in \Xi} \Big\{ \Psi(\xi) - \sum_{k} f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\},$$

we define

$$\bar{f}_1(\xi_1) := \sup_{\substack{\zeta \in \Xi \\ 0 \le u_k \le K, k \ge 2}} \Big\{ \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{k \ge 2} f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\},$$

and inductively define $\bar{f}_k$ by

$$\bar{f}_k(\xi_k) := \sup_{\substack{\zeta \in \Xi \\ 0 \le \xi_j \le K, j \ne k}} \Big\{ \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j < k} \bar{f}_j(\xi_j) - \sum_{j > k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\}, \quad \forall 2 \le k \le K.$$

The definition of Φ implies that

$$f_1(\xi_1) \ge \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j \ge 2} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta), \quad \forall \xi, \zeta \in \Xi,$$

hence $f_1 \ge \bar{f}_1$. Similarly, the definition of $\bar{f}_{k-1}$ implies that

$$f_k(\xi_k) \ge \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j < k} \bar{f}_j(\xi_j) - \sum_{j > k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta), \quad \forall \xi, \zeta \in \Xi,$$

hence $f_k \ge \bar{f}_k$ for $k \ge 2$. Hence, for all $1 \le k \le K$ it holds that

$$\begin{aligned} \bar{f}_k(\xi_k) &= \sup_{\substack{\zeta \in \Xi \\ 0 \le \xi_j \le K, j \ne k}} \Big\{ \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j < k} \bar{f}_j(\xi_j) - \sum_{j > k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\} \\ &\le \sup_{\substack{\zeta \in \Xi \\ 0 \le \xi_j \le K, j \ne k}} \Big\{ \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j \ne k} \bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\}. \end{aligned}$$

But the definition of $\bar{f}_K$ gives that

$$\bar{f}_k(\xi_k) \ge \sup_{\substack{\zeta \in \Xi \\ 0 \le \xi_j \le K, j \ne k}} \Big\{ \Psi(\xi) - \Phi(\lambda, \zeta) - \sum_{j \ne k} \bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\}, \quad \forall k.$$

Combining the previous two inequalities yields

$$\bar{f}_k(\xi_k) = \sup_{\substack{\zeta\in\Xi \\ 0\le\xi_j\le K, j\neq k}} \Big\{ \Psi(\xi) - \Phi(\lambda,\zeta) - \sum_{j\neq k} \bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big\}, \quad \forall k,$$

which also implies that

$$\sup_{\xi\in\Xi} \Big\{ \Psi(\xi) - \sum_{k=1}^K \bar{f}_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big\} = \Phi(\lambda,\zeta) = \sup_{\xi\in\Xi} \Big\{ \Psi(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big\},$$

Together with $\bar{f}_k \le f_k$, we conclude that $(\lambda, \{\bar{f}_k\}_k)$ yields a dual objective no greater than $(\lambda, \{f_k\}_k)$.

Moreover, since the dual objective remains unchanged if $\{\bar{f}_k\}_k$ is modified into $\{\bar{f}_k + a_k\}_k$, where $a_k \in \mathbb{R}$, so we may assume that $\min_{\Xi_k} \bar{f}_k = 0$. It then follows that $\{f_k\}_k$ are also upper bounded. In addition, the Lipschitz continuity of $\Psi - \lambda \mathsf{d}_F^p$ implies $\bar{f}_k$ are also Lipschitz continuous and the Lipschitz constant only depends on that of $\Psi - \lambda \mathsf{d}_F^p$.

Now let $(\lambda^{(m)}, \{f_k^{(m)}\}_k)_m$ be a minimizing sequence of (15). Using the convexification trick as above, we obtain a sequence $(\lambda^{(m)}, \{\bar{f}_k^{(m)}\}_k)_m$. Then the analysis above implies that $(\bar{f}_k^{(m)})_m$ are uniformly bounded and equi-continuous. Hence by Bolzano-Weierstrass theorem and Arzela-Ascoli theorem, there exists a convergent subsequence. Denote its limit by $(\lambda^*, \{f_k^*\}_{k=1}^K)$. Then by dominate convergence $(\lambda^*, \{f_k^*\}_{k=1}^K)$ is a dual minimizer. □

Proof of Theorem 1. Let us first relax the continuity assumption made in Step 2. We will relax the compactness assumption in the last step. Note that any upper semi-continuous function satisfying the growth rate condition can be written as the infimum of a non-increasing sequence of Lipschitz continuous functions, for example, by Moreau-Yosida approximation Ambrosio et al. (2008). Thus we can approximate Ψ by a non-increasing sequence of Lipschitz continuous functions Ψ_n and approximate d_F by a non-decreasing sequence of Lipschitz continuous functions d_n. Let us define

$$\begin{aligned}
v_P^n &:= \sup_{\boldsymbol{\mu}\in\mathfrak{M}} \int_\Xi \Psi_n d\boldsymbol{\mu}, \quad v_P^0 := \sup_{\boldsymbol{\mu}\in\mathfrak{M}} \int_\Xi \Psi d\boldsymbol{\mu}, \\
v_D^n &:= \inf_{\substack{\lambda\ge 0 \\ f_k\in B(\Xi_k)}} \Big\{ \lambda\rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) \\
&\qquad + \int_\Xi \sup_{\xi\in\Xi} \Big[\Psi_n(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_n^q(\xi,\zeta) \Big] \boldsymbol{\nu}(d\zeta) \Big\}, \\
v_D^0 &:= \inf_{\substack{\lambda\ge 0 \\ f_k\in B(\Xi_k)}} \Big\{ \lambda\rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) \\
&\qquad + \int_\Xi \sup_{\xi\in\Xi} \Big[\Psi(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big] \boldsymbol{\nu}(d\zeta) \Big\}.
\end{aligned}$$

Since $\Psi_n \geq \Psi$ and $\mathsf{d}_n \leq \mathsf{d}_F$, we have $v_D^0 \leq v_D^n$. From previous steps we know $v_D^n = v_P^n$. In view of $v_D^0 \geq v_P^0$, it remains to show $\lim_{n\to\infty} v_D^n \leq v_D^0$. From Step 4, we know that there exists $\boldsymbol{\mu}_n^*$ such that $\int_\Xi \Psi_n d\boldsymbol{\mu}_n^* = v_D^n$. Observe that $\mathfrak{M}$ is tight, then by Prokhorov's theorem, it is relatively compact with respect to the weak topology, and thus $\{\boldsymbol{\mu}^n\}_n$ admits a convergent subsequence, whose limit is denoted by $\boldsymbol{\mu}_0^*$. Then $\int_\Xi \Psi_n d\boldsymbol{\mu}_n^* \leq \int_\Xi \Psi_m d\boldsymbol{\mu}_n^*$ for all $n \geq m$ implies that

$$\lim_{n\to\infty} \int_\Xi \Psi_n d\boldsymbol{\mu}_n^* \leq \liminf_{n\to\infty} \int_\Xi \Psi_m d\boldsymbol{\mu}_n^* \leq \int_\Xi \Psi_m d\boldsymbol{\mu}_0^*.$$

Let $m \to \infty$, by monotone convergence $\lim_{n\to\infty} \int_\Xi \Psi_n d\boldsymbol{\mu}_n^* \leq \int_\Xi \Psi d\boldsymbol{\mu}_0^* \leq v_P^0$, which concludes the proof.

We next consider the setting where Ξ is not compact. For any $\epsilon > 0$, let $\Xi^\epsilon \subset \Xi$ be a compact set such that $\boldsymbol{\nu}(\Xi \setminus \Xi^\epsilon) \leq \epsilon$. Set

$$\boldsymbol{\nu}^\epsilon := \frac{\mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}}{\boldsymbol{\nu}(\Xi^\epsilon)},$$

and let F_k^ϵ be the marginal distribution of $\boldsymbol{\nu}^\epsilon$, and Ξ_k^ϵ be its support. Then the previous steps imply that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathcal{P}(\Xi^\epsilon)} \left\{ \int_{\Xi^\epsilon} \Psi d\boldsymbol{\mu} : \ W_p(\boldsymbol{\mu}, \boldsymbol{\nu}^\epsilon) \leq \rho, \ \pi_\#^k \boldsymbol{\mu} = \boldsymbol{\nu}_k^\epsilon, \forall k \right\} \\
= &\inf_{\substack{\lambda\geq 0 \\ f_k \in B(\Xi_k^\epsilon)}} \left\{ \lambda\rho^p + \sum_k \int_{\Xi_k} f_k F_k^\epsilon \right. \\
&\qquad\qquad \left. + \int_\Xi \sup_{\xi\in\Xi^\epsilon} \left[\Psi(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\right] \boldsymbol{\nu}(d\zeta) \right\} \\
=: &v^\epsilon.
\end{aligned}$$

Observe that for any feasible solution $(\lambda, \{f_k\}_k)$ of the dual problem above, the growth condition on Ψ implies that there exists sufficiently large M such that $(\lambda, \{f_k + M\mathbb{1}_{\Xi\setminus\Xi^\epsilon}\}_k)$ is a feasible solution to the original dual problem with the same objective value. Therefore, if we denote by v_P and v_D the optimal value of the original primal and dual problem respectively, then $v^\epsilon \geq v_D \geq v_P$. Let $\boldsymbol{\nu}^\epsilon$ be an optimal primal solution of the primal problem above. Define

$$\tilde{\boldsymbol{\mu}}^\epsilon := \boldsymbol{\nu}(\Xi^\epsilon)\boldsymbol{\mu}^\epsilon + \mathbb{1}_{\Xi\setminus\Xi^\epsilon}\boldsymbol{\nu}.$$

Then it holds that

$$\pi_\#^k \tilde{\boldsymbol{\mu}}^\epsilon = \boldsymbol{\nu}(\Xi^\epsilon)\pi_\#^k \boldsymbol{\mu}^\epsilon + \mathbb{1}_{\Xi\setminus\Xi^\epsilon}\pi_\#^k \boldsymbol{\nu} = \mathbb{1}_{\Xi^\epsilon} F_k + \mathbb{1}_{\Xi\setminus\Xi^\epsilon} F_k = F_k.$$

Moreover, we have that

$$W_p(\tilde{\boldsymbol{\mu}}^\epsilon, \boldsymbol{\nu}) \leq W_p(\boldsymbol{\nu}(\Xi^\epsilon)\boldsymbol{\mu}^\epsilon, \mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}) = \boldsymbol{\nu}(\Xi^\epsilon) W_p(\boldsymbol{\mu}^\epsilon, \mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}^\epsilon) \leq \boldsymbol{\nu}(\Xi^\epsilon) x \leq x,$$

Hence $\tilde{\boldsymbol{\mu}}^\epsilon$ is feasible to the original primal problem. In addition,

$$\int_\Xi \Psi d\tilde{\boldsymbol{\mu}}^\epsilon = \boldsymbol{\nu}(\Xi^\epsilon)\int_{\Xi^\epsilon}\Psi d\boldsymbol{\mu}^\epsilon + \int_{\Xi\setminus\Xi^\epsilon}\Psi d\boldsymbol{\nu} = \quad \boldsymbol{\nu}(\Xi^\epsilon)v^\epsilon + \int_{\Xi\setminus\Xi^\epsilon}\Psi d\boldsymbol{\nu}.$$

Letting $\epsilon \to 0$, we obtain that $v_P \geq v_D$. Therefore, up to a subsequence, $\tilde{\boldsymbol{\mu}}^\epsilon$ converges to $\tilde{\boldsymbol{\mu}}$, and the analysis above shows that $\tilde{\boldsymbol{\mu}}$ is primal optimal and $v_P = v_D$. □

We finally prove the measurability of the integrand involved in the dual problem. 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 Kallenberg (2006)). A function $f : \mathbb{R}^m \times \Xi \to \bar{\mathbb{R}}$ is called a *normal integrand*, if the associated epigraphical multifunction $\zeta \mapsto \text{epi } f(\cdot, \zeta)$ is closed valued and measurable.

Lemma 4. *Let $f_k \in B(\Xi_k)$. The function $\Phi : \mathbb{R} \times \Xi \to \mathbb{R}$ defined by*

$$\Phi(\lambda, \zeta) := \sup_{\xi\in\Xi}\left[\Psi(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta)\right]$$

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

Proof of Lemma 4. Define a function $g : \Xi \times \mathbb{R} \times \mathbb{R} \times \Xi \to \bar{\mathbb{R}}$ by

$$g(\xi, \lambda, \zeta) = \Psi(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta).$$

Then for every $\zeta \in \Xi$, $-g(\cdot,\cdot,\cdot,\zeta)$ is lower semi-continuous, thus g is $\mathscr{B}(\Xi) \otimes \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$-measurable. Hence by joint measurability criterion (see, e.g., Corollary 14.34 in Rockafellar and Wets (2009)), g is a normal integrand, thereby the function Φ is also a normal integrand (Theorem 7.38 in Shapiro et al. (2009)). □

Endnotes

1. A function is supermodular, if

$$\Psi_x(\xi_1,\ldots,\xi_k,\ldots,\xi_{k'},\ldots,\xi_K) + \Psi_x(\xi_1,\ldots,\xi_k+\epsilon,\ldots,\xi_{k'}+\delta,\ldots,\xi_K)$$
$$\geq \Psi_x(\xi_1,\ldots,\xi_k+\epsilon,\ldots,\xi_{k'},\ldots,\xi_K) + \Psi_x(\xi_1,\ldots,\xi_k,\ldots,\xi_{k'}+\delta,\ldots,\xi_K)$$

for all $\xi \in \Xi$, $1 \leq k < k' \leq K$ and $\epsilon, \delta > 0$

2. A distribution is comonotonic if its cumulative distribution function satisfies

$$F^{\boldsymbol{\mu}^*}(\xi_1,\ldots,\xi_K) \quad = \quad \min_{1\leq k\leq K} F_k^{\boldsymbol{\mu}^*}(\xi_k), \quad \forall\, \xi.$$

3. Some authors consider KL ball centered at some nominal distribution instead of nominal copula. Nevertheless, it can be easily shown that the Kullback-Leibler divergence between two distributions equals the Kullback-Leibler divergence between their associated copulas (cf. Sec 10.4 in Schmid et al. (2010).

References

Agrawal S, Ding Y, Saberi A, Ye Y (2012) Price of correlations in stochastic optimization. *Operations Research* 60(1):150–162.

Aldrovandi R, Pereira JG (1995) *An introduction to geometrical physics* (World scientific).

Ambrosio L, Gigli N, Savaré G (2008) *Gradient flows: in metric spaces and in the space of probability measures* (Springer Science & Business Media).

Benes V, Stepán J (2012) *Distributions with given marginals and moment problems* (Springer Science & Business Media).

Blanchet J, Murthy K (2016) Quantifying distributional model risk via optimal transport .

Deheuvels P (1979) La fonction de dépendance empirique et ses propriétés. un test non paramétrique dindépendance. *Acad. Roy. Belg. Bull. Cl. Sci.(5)* 65(6):274–292.

Dey S, Juneja S, Murthy KR (2015) Incorporating views on marginal distributions in the calibration of risk models. *Operations Research Letters* 43(1):46–51.

Dhaene J, Goovaerts MJ (1997) On the dependency of risks in the individual life model. *Insurance: Mathematics and Economics* 19(3):243–253.

Dhara A, Das B, Natarajan K (2017) Worst-case expected shortfall with univariate and bivariate marginals. *arXiv preprint arXiv:1701.04167* .

Doan XV, Natarajan K (2012) On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems. *Operations research* 60(1):138–149.

Esfahani PM, Kuhn D (2015) Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116* .

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33(1):3–56.

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

Frech KR (2017) 30 industry portfolios. URL `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`, online; accessed April 2017.

Fréchet M (1960) Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut international de statistique* 10–32.

Gangbo W, Swiech A (1998) Optimal maps for the multidimensional monge-kantorovich problem. *Communications on pure and applied mathematics* 51(1):23–45.

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Gao R, Kleywegt AJ (2017) Distributionally robust stochastic optimization with dependence structure. *arXiv preprint arXiv:1701.04200* .

Glasserman P, Yang L (2016) Bounding wrong-way risk in cva calculation. *Mathematical Finance* .

Hall P, Neumeyer N (2006) Estimating a bivariate density when there are extra data on one or both components. *Biometrika* 439–450.

Hoeffding W (1940) *Massstabinvariante korrelationstheorie* (Teubner), (Translated in: The CoHected Works of Wassily Hoeffding, N. I. Fisher and P. K. Sen (eds.), Springer Verlag, New York 1994).

Joe H (1997) *Multivariate models and multivariate dependence concepts* (CRC Press).

Joe H (2014) *Dependence modeling with copulas* (CRC Press).

Kallenberg O (2006) *Foundations of modern probability* (Springer Science & Business Media).

Kellerer HG (1984) Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 67(4):399–432.

Lam H (2017) Sensitivity to serial dependency of input processes: a robust approach. *Management Science* .

Müller A (1997) Stop-loss order for portfolios of dependent risks. *Insurance: Mathematics and Economics* 21(3):219–223.

Natarajan K, Song M, Teo CP (2009) Persistency model and its applications in choice modeling. *Management Science* 55(3):453–469.

Nelsen RB (2013) *An introduction to copulas*, volume 139 (Springer Science & Business Media).

Qu L, Yin W (2012) Copula density estimation by total variation penalized likelihood with linear equality constraints. *Computational Statistics & Data Analysis* 56(2):384–398.

Rachev ST (1985) The monge-kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications* 29(4):647–676.

Rachev ST, Rüschendorf L (1998) *Mass Transportation Problems: Volume I: Theory*, volume 1 (Springer Science & Business Media).

Rényi A (1959) On measures of dependence. *Acta mathematica hungarica* 10(3-4):441–451.

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *Journal of risk* 2:21–42.

Rockafellar RT, Wets RJB (2009) *Variational analysis*, volume 317 (Springer Science & Business Media).

Schmid F, Schmidt R, Blumentritt T, Gaißer S, Ruppert M (2010) Copula-based measures of multivariate association. *Copula theory and its applications*, 209–236 (Springer).

Schweizer B, Wolff EF (1981) On nonparametric measures of dependence for random variables. *The annals of statistics* 879–885.

Shapiro A, Dentcheva D, Ruszczynski A (2009) Lectures on stochastic programming, volume 9 of mps/siam series on optimization. *Philadelphia, PA: SIAM. Modeling and theory* .

Sklar M (1959) *Fonctions de répartition à n dimensions et leurs marges* (Université Paris 8).

Tsukahara H (2005) Semiparametric estimation in copula models. *Canadian Journal of Statistics* 33(3):357–375.

Vallender S (1974) Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18(4):784–786.

Zalinescu C (2002) *Convex analysis in general vector spaces* (World Scientific).