Data-Driven Distributionally Robust Optimization Applied to Probability Density Estimation and Portfolio Optimization

Diego. F. Fonseca. V. Department of Mathematics, Universidad de los Andes, df.fonseca@uniandes.edu.co

Universidad de los Andes

Introduction

Stochastic programming is a framework for modeling optimization problems that involves uncertainty in the form of a random variable. Usually, the probability distribution of the variable is unknown, hence another approach known as Distributionally Robust Optimization (DRO) has emerged recently. This approach assumes that the true distribution of the random variable involved in the problem belongs to a set of distributions called Ambiguity Set. When this set is defined as a ball with respect to the Wasserstein metric centered at the empiric distribution, the DRO problem can be reformulated as a semi-infinite optimization problem. In this work we show two applications. First, in the context of statistics, we show that the choice of the *bandwidth* parameter in the kernel density estimation (KDE) is a stochastic problem and we formulate its DRO version. Finally, a DRO version of Markowitz Mean-Variance portfolio optimization problem is presented.

Stochastic programming and DRO

A *stochastic program* is an optimization program given by

$$J^* = \min_{x\in\mathbb{X}} \mathbb{E}_{\mathbb{P}}[f(x,\xi)]$$

where $f : \mathbb{X}\times\Xi \to \mathbb{R}$, $\mathbb{X}$ is the feasible region and ξ is random parameter with distribution $\mathbb{P}$ supported in Ξ. The objective is to estimate J^*.

Definition [Wasserstein Metric]

The *Wasserstein distance* $W_p(\mu,\nu)$ between $\mu,\nu \in \mathcal{P}_p(\Xi)$ is defined by

$$W_p^p(\mu,\nu) := \inf_{\Pi\in\mathcal{P}(\Xi\times\Xi)} \left\{ \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi(d\xi,d\zeta) \;\middle|\; \begin{array}{l}\Pi(\cdot\times\Xi)=\mu(\cdot),\\ \Pi(\Xi\times\cdot)=\nu(\cdot)\end{array} \right\}$$

where d is a metric on Ξ and $\mathcal{P}_p(\Xi)$ is the set of probability measures μ such that $\int_\Xi d^p(\xi,\zeta_0)\mu(d\xi) < \infty$ for some $\zeta_0\in\Xi$.

In $\mathcal{P}_p(\Xi)$ the ball of ratio $\varepsilon > 0$ and centered at $\mu \in \mathcal{P}_p(\Xi)$ is $\mathcal{B}_\varepsilon(\mu) = \left\{\nu\in\mathcal{P}_p(\Xi) \;\middle|\; W_p^p(\mu,\nu) \le \varepsilon^p\right\}$.

Let $\widehat{\xi}_1,\dots,\widehat{\xi}_N$ be a sample of $\mathbb{P}$ and $\widehat{\mathbb{P}}_N$ the empirical distribution determined by this sample. Given $\varepsilon > 0$ we consider $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$. The DRO problem that estimates J^* with hight probability is

$$\widehat{J}_N := \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]. \tag{1}$$

Let us study the internal problem of (1). If we omit x, then we obtain:

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(\xi)]. \tag{2}$$

Assumption 1

f satisfies one of following conditions:

1. f is continuous and there exists $C > 0$ and $\xi_0\in\Xi$ such that $|f(\xi)| \le C(1+d^p(\xi,\xi_o))$ for all $\xi\in\Xi$.
2. f is bounded.
3. f is the point-wise maximum of concave functions.

Theorem [Main theorem]

Under Assumption 1 the problem (2) can be formulated as the semi-infinite optimization program

$$\begin{cases} \inf\limits_{\lambda,\,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^N s_i \\ \text{subject to} & \sup\limits_{\xi\in\Xi}\left(f(\xi)-\lambda d^p(\xi,\widehat{\xi}_i)\right) \le s_i \;\forall i\le N. \\ & \lambda \ge 0. \end{cases} \tag{3}$$

Probability Density Estimation

Let ξ be a random variable and f its density function, given a sample $\widehat{\xi}_1,\dots,\widehat{\xi}_N$ we have the Kernel Density Estimator of f given by

$$\widehat{f}_h(x) := \frac{1}{Nh}\sum_{i=1}^N \mathcal{K}\left(\frac{x-\widehat{\xi}_i}{h}\right),$$

where $\mathcal{K}$ is a density function such that $\int x\mathcal{K}(x)dx = 0$ and $\int x^2\mathcal{K}(x)dx = 1$, and $h>0$ is a parameter called *bandwidth*.

Choice of the bandwidth as a DRO

The *bandwidth* $h>0$ is chosen in such a way that minimizes the *mean integrated squared error*:

$$\mathrm{MISE}(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int\left(\widehat{f}_h(x)-f(x)\right)^2 dx\right]. \tag{4}$$

Since f is unknown then $\mathbb{P}$ is also unknown. Note that the h that minimizes the MISE is the solution of a stochastic problem. In fact, calling

$$J(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int\left(\widehat{f}_h(x)\right)^2 dx\right] - 2\mathbb{E}_{\mathbb{P}^N}\left[\int \widehat{f}_h(x)f(x)dx\right], \tag{5}$$

we have $\mathrm{MISE}(h) = J(h) + \int (f(x))^2 dx$. Therefore

$$\operatorname*{argmin}_{h>0} \mathrm{MISE}(h) = \operatorname*{argmin}_{h>0} J(h).$$

But J can be expressed as

$$J(h) = \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\frac{1}{N^2h^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right)\right].$$

Let us denote

$$F(h,\xi,\zeta) := \frac{1}{Nh^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right).$$

We have the stochastic program

$$J^* := \min_{h\ge0} J(h) = \min_{h\ge0}\mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[F(h,\xi,\zeta)\right] \tag{6}$$

where $\mathbb{P}$ is unknown.

We assume that the support of ξ is $\Xi = \mathbb{R}$ and we consider the 2-Wasserstein metric with d as the euclidean distance in $\mathbb{R}^2$. We consider other sample $\widehat{\xi}_1^*,\dots,\widehat{\xi}_N^*$ of ξ, then $(\widehat{\xi}_1,\widehat{\xi}_1^*),\dots,(\widehat{\xi}_N,\widehat{\xi}_N^*)$ is a sample of (ξ,ζ) what allows to define the empirical distribution $\widehat{\mathbb{P}}_N = \sum_{i=1}^N \delta_{(\widehat{\xi}_i,\widehat{\xi}_i^*)}$.

Therefore, the DRO counterpart of (6) is

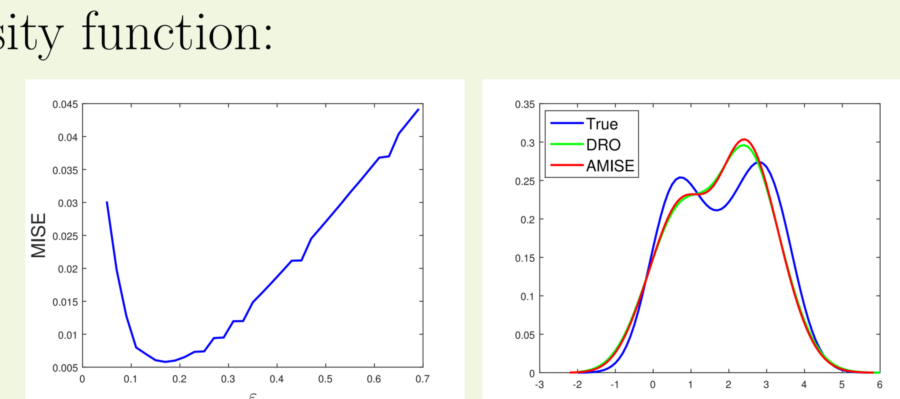
$$\widehat{J}_N = \min_{h\ge0}\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[F(h,\xi,\zeta)\right]$$

By *Main Theorem* it is equivalent to

$$\begin{cases} \inf\limits_{h,\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^N s_i \\ \text{subject to} & \sup\limits_{(\xi,\zeta)\in\mathbb{R}^2}\left(F(h,\xi,\zeta) - \lambda\left\|(\xi,\zeta)-(\widehat{\xi}_i,\widehat{\xi}_i^*)\right\|^2\right) \le s_i \;\forall i=1,\dots,N, \\ & \lambda\ge0, \\ & h\ge0. \end{cases} \tag{7}$$

Results

The following is the result of estimating a two-peak density function:



(a) (b)

Figure 1: (a) ε vs $MISE$. (b) True density and estimations DRO and AMISE.

Portfolio Optimization

We consider a portfolio of m assets whose rate of return are described by the random vector $\xi\in\mathbb{R}^m$ and we denote by $x\in\mathbb{R}^m$ the vector of weights associated with the portfolio.

$$\text{Expected Return} := \mathbb{E}_{\mathbb{P}}\left[\langle x,\xi\rangle\right], \text{Volatility} := \mathrm{Var}_{\mathbb{P}}\left[\langle x,\xi\rangle\right].$$

In the context of Markowitz theory, the investor wants to minimize volatility subject to obtain a expected return bigger than a given value.

$$J := \begin{cases}\min\limits_{x\in\mathbb{R}^m} & \mathrm{Var}_{\mathbb{P}}\left[\langle x,\xi\rangle\right] \\ \text{subject to} & \mathbb{E}_{\mathbb{P}}\left[\langle x,\xi\rangle\right]\ge\mu,\;\; \sum_{i=1}^m x_i = 1.\end{cases} \tag{8}$$

DRO version

Let $\widehat{\xi}_1,\dots,\widehat{\xi}_N$ be a sample of ξ, for $x\in\mathbb{X}$ fixed we define $\zeta^x := \langle x,\xi\rangle$, then $\widehat{\zeta}_1^x,\dots,\widehat{\zeta}_N^x$ defined by $\widehat{\zeta}_i^x := \langle x,\widehat{\xi}_i\rangle$ is a sample of ζ^x. Now, we consider $\widehat{\mathbb{P}}_N^x$ the empirical distribution of ζ^x and $\mathcal{B}_{\varepsilon\|x\|}(\widehat{\mathbb{P}}_N^x)$ the ball of ratio $\varepsilon\|x\|$. We define

$$\mathbb{X} := \left\{x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^m x_i = 1,\;\; \inf_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathbb{E}_{\mathbb{Q}}[\zeta^x]\ge\mu\right\}.$$

The DRO counterpart of (8) is

$$\widehat{J}_N := \min_{x\in\mathbb{X}}\sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathrm{Var}_{\mathbb{Q}}\left[\zeta^x\right]. \tag{9}$$

By *Main Theorem* it is equivalent to

$$\begin{cases}\min\limits_{x\in\mathbb{R}^m} & \left(\sqrt{x^TK_Nx}+\varepsilon\|x\|\right)^2 \\ \text{suject to} & \dfrac{1}{N}\displaystyle\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle - \varepsilon\|x\| \ge \mu,\;\; \sum_{i=1}^m x_i = 1,\end{cases}$$

where K_N is the sample Covariance matrix.

Choice of ε

Prioritizing the risk: ε_{var} the lowest value of $\varepsilon\ge0$ that satisfies $V(\varepsilon) := \mathbb{E}_{\mathbb{P}^N}\left[\widehat{J}_N(\varepsilon) - \mathrm{Var}_{\mathbb{P}}\left[\langle\widehat{x}_N(\varepsilon),\xi\rangle\right]\right]\ge0$.

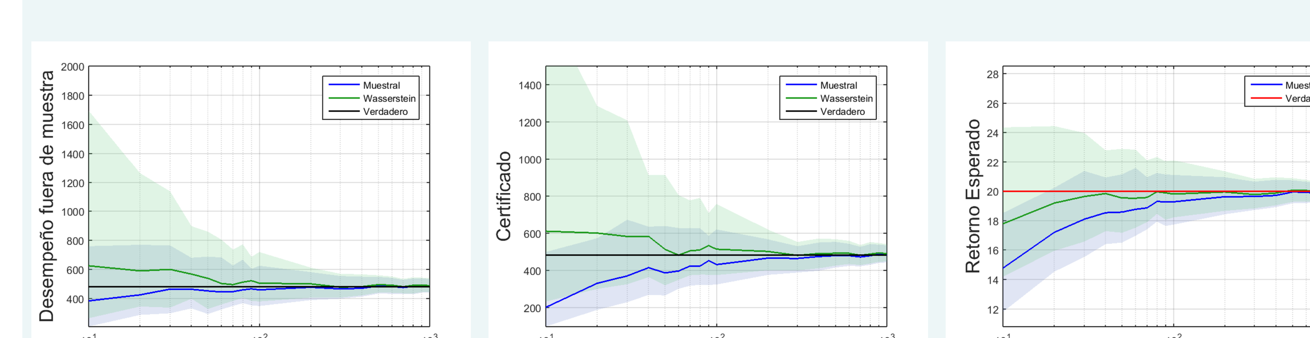
Prioritizing the return: ε_{ret} the lowest value of $\varepsilon\ge0$ that satisfies $R(\varepsilon) := \mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}\left[\langle\widehat{x}_N(\varepsilon),\xi\rangle\right]\right]\ge\mu$.

Results

We consider $\mu = 20$, $m = 4$ and ξ with multinormal distribution with covariance matrix and mean vector

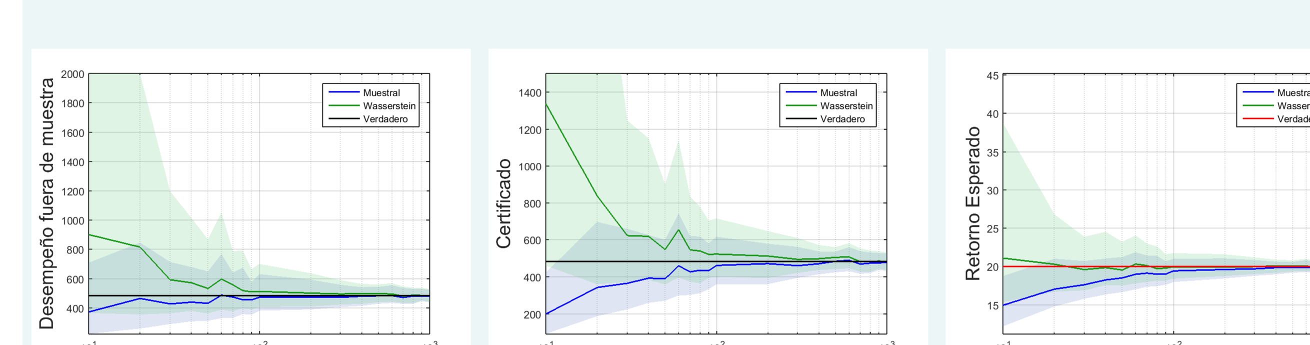
$$C = \begin{bmatrix}185 & 86.5 & 80 & 20\\ 86.5 & 196 & 76 & 13.5\\ 80 & 76 & 411 & -19\\ 20 & 13.5 & -19 & 25\end{bmatrix} \text{ y } \mathbf{m} = (14,12,15,17).$$

(a) Out-of-sample performance $\mathrm{Var}_{\mathbb{P}}[\langle\widehat{x}_N,\xi\rangle]$ (green) and $\mathrm{Var}_{\mathbb{P}}[\langle\widehat{x}_N^{sam},\xi\rangle]$ (blue) where $\widehat{x}_N^{sam}$ is sampled solution. (b) certificate $\widehat{J}_N$ (green) y $\mathrm{Var}_{\widehat{\mathbb{P}}_N}[\langle\widehat{x}_N^{sam},\xi\rangle]$ (blue). (c) Expected return $\mathbb{E}_{\mathbb{P}}[\langle\widehat{x}_N,\xi\rangle]$ (green) y $\mathbb{E}_{\mathbb{P}}[\langle\widehat{x}_N^{sam},\xi\rangle]$ (blue).



(a) (b) (c)

Figure 2: **Prioritizing the risk:** $\varepsilon = \varepsilon_{\text{var}}$



(a) (b) (c)

Figure 3: **Prioritizing the return:** $\varepsilon = \varepsilon_{\text{ret}}$