CASE STUDY: Minimization of Kantorovich-Rubinstein distance between two distributions (kantor, ksm_avg, cardn, linear, pcvar)

Background

Approximation one distribution by some other distribution is a popular topic in academic literate. Various metrics are used to estimate distance between the target (approximated) and approximating distributions.

This case study considers two finite one-dimensional discrete distributions. A target discrete distribution has fixed parameters and an approximating distribution has a given number of atoms and variable positions and probabilities of atoms. The case study demonstrates how to solve the approximation problems by minimizing Kantorovich-Rubinstein and Average Kolmogorov-Smirnov distances between distributions.

By definition Kantorovich-Rubinstein distance is an area between two CDFs. We minimized Kantorovich-Rubinstein distance between the fixed target discrete distribution and approximating distribution with a fixed number of atoms and variable probabilities and locations of the atoms. The optimization problem is multi-extremal and optimization may give a local minimum (instead of global). We have considered two variants of the Problem Statement: without and with initial point (approximating distribution).

In minimizing Average Kolmogorov-Smirnov distance both distributions should have the same positions. The minimization problem contains a cardinality constraint on the number of used position for the approximating distribution with the variable probabilities of atoms. The optimization problem is multi-extremal and optimization may give a local minimum (instead of global).

This case study considers 3 Problems for minimization Average Kolmogorov-Smirnov distance.

The first Problem minimizes Average Kolmogorov-Smirnov distance and finds an approximating distribution with fixed locations and variable probabilities of the atoms. In this case, locations of atoms of the approximating distribution coincide with the location of atoms of the target distribution. To constraint the number of atoms with nonzero probabilities in the approximating distribution, we have imposed a cardinality constraint. We considered two options for the cardinality constraint: 1) “linearize=1”; 2) “mip=1”. These two options initiate different optimization approaches with cardinality constraints.

The second and the third Problem minimize Average Kolmogorov-Smirnov distance and find an approximating distribution with fixed locations and variable probabilities of atoms. The number of atoms of the approximating distribution is 4 times smaller than the number of atoms of the target distribution. To assure that the right tail of the approximating distribution correctly matches the right tail of the target distribution we have imposed two CVaR constraints. We want to minimize distance between distributions and match two CVaRs of approximating and target distributions. We have considered two cases: 1) CVaRs of approximating distribution are larger than CVaRs of target distribution (in this case tails of approximating distribution are heavier than tail of target distribution). This problem is convex and optimality of the solution in guaranteed; 2) CVaRs of approximating distribution are equal to CVaRs of target distribution. This problem may have multiple extremums and solver finds some extremum (may be a local one).

References

- Kantorovich, L.V., and Rubinstein, G.Sh. On a space of totally additive functions, Vestn. Lening. Univ., Vol. 13, No. 7, pp. 52-59, 1958.

Notations

m = number of atoms in the target distribution;

n = number of atoms in the approximating distribution;

$\{y_1, \dots, y_m\}$ = set of positions of atoms in the target distribution;

$\{q_1, \dots, q_m\}$ = set of probabilities of atoms in the approximating distribution;

$\vec{p} = (p_1, \dots, p_n)$ = vector of variable probabilities p_i;

$\vec{x} = (x_1, \dots, x_n)$ = vector of variable positions x_i;

$Z = \{x_1, \dots, x_n\} \cup \{y_1, \dots, y_m\}$ = union of positions of atoms;

z_j = j-th element in the ascending ordered set Z (values z_j depend on variable positions $\vec{x}$),

$F_Y(z_j)$ = CDF value of the target distribution at the point $z_j \in Z$;

$F_X(\vec{p}, \vec{x}, z_j)$ = CDF value of the approximating distribution at the point $z_j \in Z$;

$A_j(\vec{p}, \vec{x}) = |F_Y(z_j) - F_X(\vec{p}, \vec{x}, z_j)|$ = difference between distributions at the point $z_j \in Z$;

$w_j(\vec{x}) = z_{j+1}(\vec{x}) - z_j(\vec{x})$ = weight of difference $A_j(\vec{p}, \vec{x}),\ j = 1, \dots, |Z| - 1$;

$\boldsymbol{kantor}(\vec{x}, \vec{p}) = \sum_{j=1}^{|Z|-1} w_j(\vec{x}) A_j(\vec{p}, \vec{x})$ = Kantorovich-Rubinstein distance between two distributions;

$\boldsymbol{linear}(\vec{p}) = \sum_{i=1}^{n} p_i$ = sum of variable probabilities;

$\boldsymbol{ksm\_avg}(\vec{p}, \vec{x}) = \frac{\sum_{j=1}^{|Z|-1} w_j(\vec{x}) A_j(\vec{p}, \vec{x})}{\sum_{j=1}^{|Z|-1} w_j(\vec{x})}$ = Average Kolmogorov-Smirnov distance between two distributions

(probabilities $\vec{p} = (p_1, \dots, p_n)$ are variables and $\vec{x}$ is a fixed parameter);

$\boldsymbol{cardn}(\vec{p}) = \sum_{i=1}^{n} u(p_i, w)$ = Cardinality function,
where

$$u(y, w) = \begin{cases} 0, & \text{if } -w < y < w \\ 1, & \text{otherwise} \end{cases};$$

$w > 0$ is a threshold value;

$\boldsymbol{pcvar}(\vec{p})$ = CVaR for Discrete Distribution as Function of atom probabilities $\vec{p}$ with fixed atom positions $\vec{x}$.

Optimization Problem 1

minimizing Kantorovich-Rubinstein distance

$$\min_{\vec{x}, \vec{p}} \boldsymbol{kantor}(\vec{x}, \vec{p})$$

subject to

constraints on probabilities

$$\boldsymbol{linear}(\vec{p}) = 1,$$

$$p_i \geq 0, i = 1, \dots, n.$$

***Note*:** variables for Kantor function and constraints on probabilities are generated by PSG automatically, so user should not define these variables and constraints in the PSG Problem Statement.

Optimization Problem 2

minimizing Average Kolmogorov-Smirnov distance

$$\min_{\vec{p}_y} \boldsymbol{ksm\_avg}(\vec{p}_y, \vec{y})$$

subject to

constraint on a number of used atoms

$$\boldsymbol{cardn}(\vec{p}_y) \leq n,$$

constraint on probabilities

$$\boldsymbol{linear}(\vec{p}_y) = 1,$$

$$p_{yi} \geq 0, i = 1, \dots, m.$$

***Note*:** positions of atoms in *Optimization Problem* 2 are the same for the target and approximating distributions.

Optimization Problem 3

minimizing Average Kolmogorov-Smirnov distance

$$\min_{\vec{p}_x} \boldsymbol{ksm\_avg}(\vec{p}_x, \vec{x})$$

subject to

inequality constraints on CVaRs

$$\boldsymbol{pcvar}(0.9, \vec{p}_x) \geq \; CVaR(0.9, \vec{y}, \vec{q}),$$

$$\boldsymbol{pcvar}(0.95, \vec{p}_x) \geq \; CVaR(0.95, \vec{y}, \vec{q}),$$

constraint on probabilities

$$\boldsymbol{linear}(\vec{p}_x) = 1,$$

$$p_{xi} \geq 0, i = 1, \dots, n.$$

***Note*:** positions of atoms for approximating distribution in *Optimization Problem* 3 are a subset of positions of atoms in the target distribution.

Optimization Problem 4

minimizing Average Kolmogorov-Smirnov distance

$$\min_{\vec{p}_x} \boldsymbol{ksm\_avg}(\vec{p}_x, \vec{x})$$

subject to

equality constraints on CVaRs

$$\boldsymbol{pcvar}(0.9, \vec{p}_x) = CVaR(0.9, \vec{y}, \vec{q}),$$

$$\boldsymbol{pcvar}(0.95, \vec{p}_x) = CVaR(0.95, \vec{y}, \vec{q}),$$

constraint on probabilities

$$\boldsymbol{linear}(\vec{p}_x) = 1,$$

$$p_{xi} \geq 0, i = 1, \dots, n.$$

***Note*:** positions of atoms for approximating distribution in *Optimization Problem* 3 are a subset of positions of atoms in the target distribution.