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

Background

Approximation one distribution by some other distribution is a popular topic in academic literate. To estimate a distance between approximated and approximating distributions a lot of metrics between two distributions are considered in probability and risk theory.

This case study considers two finite discrete distributions. Approximated one has fixed parameters and other has given number of atoms and variable positions and probabilities. The case study demonstrates how approximate fixed distribution by variable distribution minimizing Kantorovich-Rubinstein distance.

The same result may be obtained minimizing Average Kolmogorov-Smirnov distance between distributions because there exists an optimal solution in minimizing Kantorovich-Rubinstein distance for which variable positions are a subset of fixed positions. In minimizing Average Kolmogorov-Smirnov distance both distributions should have the same positions and minimization problem should contain a constraint on the number of used position for distribution with variable probability.

Objective in minimization Kantorovich-Rubinstein distance is an area between two CDFs. Objective in minimization Average Kolmogorov-Smirnov distance is an average difference between two CDFs. If width of fixed distribution (the difference between maximal point where fixed CDF is 0 and minimal point where fixed CDF is 1) is 1 then values of objectives in optimal points should be equal. But both problems are multi-extremal and optimization may give local minima instead of global.

This case study provides two Problem statements for minimization Kantorovich-Rubinstein distance and two Problem statements for minimization Average Kolmogorov-Smirnov distance.

For case of Kantorovich-Rubinstein distance one Problem statement contains initial point, other does not. When initial point is present in minimization Kantorovich-Rubinstein distance then optimization begins from this point. If initial point is not present then PSG uses a heuristic to find initial point with small value of Kantorovich-Rubinstein distance.

For case of Average Kolmogorov-Smirnov distance one Problem statement contains “linearize=1” option, other one contains “mip=1” option for constraint on cardinality function. These options provide different approaches for working with cardinality function.

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 a fixed distribution;

n = number of atoms in a variable distribution;

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

$\{q_1, \dots, q_m\}$ = set of probabilities of atoms in fixed 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 atoms positions;

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

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

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

$A_j(\vec{p}, \vec{x}) = \left|F_Y(z_j) - F_X(\vec{p}, \vec{x}, z_j)\right|$ = difference between distributions at 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 ($\vec{x}$ is fixed parameter in this function) ;

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

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

$w > 0$ is threshold value;

Optimization Problem 1

minimizing Kantorovich-Rubinstein distance

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

subject to

constraints on probabilities

$$\boldsymbol{linear}(\vec{p}) = 1,$$
$$p_i \geq 0, i = 1, \dots, n.$$

***Remark*:** variables for Kantor function and constraints on probabilities are generated by PSG automatically, so user should not define these variables and constraints in 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

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

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

***Remark*:** positions of atoms in *Optimization problem* 2 are the same for both distributions.