CASE STUDY: Minimization of Kantorovich-Rubinstein Distance between Two Distributions of Atoms in Euclidean Space (linear, sqrt_quadratic)

Background

This case study presents numerical algorithm for approximation of discrete distribution in *k*-dimension space by some other discrete distribution with a smaller number of atoms. The approximation is done with the Kantorovich-Rubinstein distance between distributions. The case study is described in Kuzmenko and Uryasev [1].

Kantorovich–Rubinstein distance was defined in 1958 [2]. This distance is closely related with the general continuous transportation problem on a compact metric space, which was formulated in Kantorovich's paper in 1942 [3]. Kantorovich established a relation to Monge's problem of excavations and embankments (transportation problem in Euclidean space) [4].

Positions and probabilities of atoms of the approximating distribution are variables of the optimization problem. The algorithm is based on a consecutive solution of a sequence of optimization problems reducing the distance between distributions. The initial positions of atoms in the approximating distribution are chosen by solving *k*-means clustering problem. One iteration of the approximation algorithm solves two optimization problems. The first problem changes position of atoms of the approximating distribution by finding a minimal sum of distances between atoms. The second problem finds a nearest atom of the approximating (variable) distribution for every atom of the fixed (target) distribution (i.e., this problem links atoms of two distributions). The iteration process stops when the sum of distances achieves a minimum value (i.e., the distance stops changing).

This case study is implemented in MATLAB environment. The MATLAB code reads initial data, generates an initial approximating solution, organizes cycles, prepares and modifies data for Optimization Problems 1 and 2. Problem 1 is solved with Portfolio Safeguard (PSG) called from MATLAB. One instance of Problems 1 is exported to text files and it is demonstrated how it can be solved in the Run-Files environment.

PSG Lp-norm function can be used to calculate distance between atoms in a multidimensional case. In this case study we have used Euclidian distance, which was calculated with the PSG function Sqrt_qudratic.

References

1. Kuzmenko V. and S. Uryasev. *Kantorovich-Rubinstein distance minimization: application to location problems.* In Large Scale Optimization Applied to Supply Chain & Smart Manufacturing: Theory & Real Applications. Springer Optimization and Its Applications. 2019.
2. 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.
3. Kantorovich, L.V. *On the translocation of masses*, Dokl. Akad. Nauk SSSR, Vol. 37, No. 7-8, pp. 227-229, 1942.
4. Kantorovich, L.V. *On a problem of Monge*, Uspekhi Mat. Nauk, Vol. 3, No. 2, pp. 225-226, 1948.

Notations

m = number of atoms in the target distribution;

n = number of atoms in the approximating distribution, $n < m$;

$Y = \{\vec{y}_1, \dots, \vec{y}_m\}$ = positions of atoms in the target distribution, $\vec{y}_j \in R^k, j = 1, \dots m$;

$\vec{q} = \{q_1, \dots, q_m\}$ = vector of probabilities of atoms in the target distribution;

$X = \{\vec{x}_1, \dots, \vec{x}_n\}$ = positions of atoms in the approximating (variable) distribution, $\vec{x}_i \in R^k, i = 1, \dots n$;

$\vec{p} = (p_1, \dots, p_n)$ = vector of probabilities of atoms in the approximating (variable) distribution;

$dist(\vec{x}_i, \vec{y}_j) = \sqrt[p]{\sum_{l=1}^{k} |x_{il} - y_{jl}|^p}$, $p > 0$ = l_p–norm = the distance between position $\vec{x}_i$ and $\vec{y}_j$. The case $p = 2$ corresponds to the Euclidian norm, which is used in this case study.

The positions and probabilities of atoms of the approximating distribution with the smallest Kantorovich-Rubenstein distance to the target distribution is found by solving the following optimization problem,

$$\min_{\vec{x}_i, w_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j)\, w_{ij} \tag{1}$$

subject to

$$\sum_{i=1}^{n} w_{ij} = q_j, \;\; j = 1, \dots m, \tag{2}$$

$$w_{ij} \geq 0, \;\; j = 1, \dots m, \;\; i = 1, \dots n. \tag{3}$$

Suppose that $\vec{x}_i^*, w_{ij}^*, \;\; i = 1, \dots n, \;\; j = 1, \dots m$, is an optimal solution of problem (1)-(3). The optimal probabilities of the approximating distribution are equal to $p_i^* = \sum_{j=1}^{m} w_{ij}^*$, $i = 1, \dots n$.

The algorithm solves in a cycle the following pair of optimization problems.

Optimization Problem 1 (with fixed w_{ij})

$$\min_{\vec{x}_i} \sum_{i=1}^{n} \sum_{j=1}^{m} dist(\vec{x}_i, \vec{y}_j) w_{ij},$$

Optimization Problem 2

$$\min_{w_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} w_{ij}$$

subject to

$$\sum_{i=1}^{n} w_{ij} = q_j, \;\; j = 1, \ldots m,$$
$$w_{ij} \geq 0, \;\; j = 1, \ldots m, \;\; i = 1, \ldots n.$$

where $c_{ij} = dist(\vec{x}_i, \vec{y}_j)$ for the fixed $\vec{x}_i$.