Atelier
« Méthodes de calcul pour le transport optimal
18—22 juillet 2016

Workshop
"Computational Optimal Transportation"
July 18—22, 2016

Explanation of variability and removal of confounding factors from data through optimal transport

Giulio Trigila*

trigila@cims.nyu.edu

A methodology based on the theory of optimal transport is developed to attribute variability in data sets to known and unknown factors and to remove such attributable components of the variability from the data. Denoting by x the quantities of interest and by z the explanatory factors, the procedure transforms x into filtered variables y through a z-dependent map, so that the conditional probability distributions $\rho(x|z)$ are pushed forward into a target distribution $\mu(y)$, independent of z. Among all maps and target distributions that achieve this goal, the procedure selects the one that minimally distorts the original data: the barycenter of the $\rho(x \mid z)$.

In the language of optimal transport, the resulting $\mu(y)$ is the weighted barycenter of the $\rho(x \mid z)$.

Connections are found to unsupervised learning –particularly simple instances of the methodology are shown to be equivalent to k-means and principal component analysis— and to fundamental problems in statistics such as conditional density estimation and sampling. An applications is shown to a time-series of ground temperature hourly data across the United States.

*Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA