Inference through optimal transport

Esteban G. Tabak

New York University

Courant Institute of Mathematical Sciences

Dynamics and Data Assimilation, Physiology and Bioinformatics: Mathematics at the Interface of Theory and Clinical Application

Banff, 2022.

The problem

Data:

x^i: state

z^i: factors

a^i: action

g^i: reference group

Goal:

Estimate $\rho\left(x|z_*, a_*, g_*\right)$

or simulate $x_*^j \sim \rho\left(:|z_*, a_*, g_*\right)$.

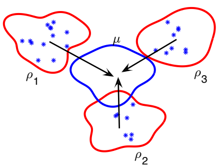
Examples:

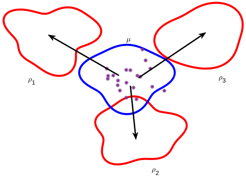
Diagnosis, forecast, treatment effect estimation

Conditional density simulation through the optimal transport barycenter problem

Include the action a among the factors z, forget temporarily the reference group g.

Remove from x the variability attributable to z through a map $y = T(x; z)$ such that $\mu(y) = T\#\rho(x|z)$ is independent of z. Among such maps T, select the minimizer of a cost function $C(T)$ associated with data distortion.





$$x^i \sim \rho\left(: |z^i\right) \to y^i = T\left(x^i; z^i\right) \to x^i_* = T^{-1}\left(y^i; z_*\right) \sim \rho\left(: |z_*\right)$$

An example: hourly temperature in Ithaca, NY

Set 1: Static covariates: time of day, day of year, year.
Set 2: Static + local temperature 24 hours before.
Set 3: Static + temperature at 3 locations 36 hours before.

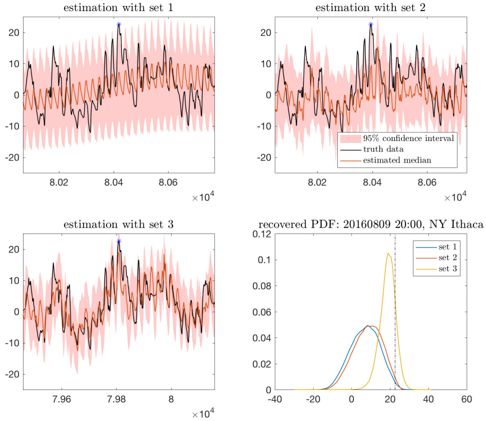
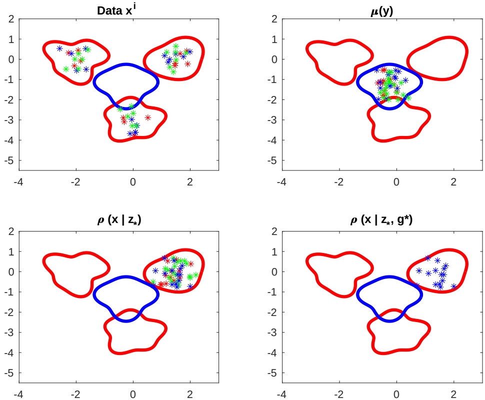


Figure: Observations, estimated median and 95% confidence interval.

Capturing idiosyncratic factors through sub-sampling



Trajectories

$$y_* = T\left(x_*; z_*, a_*\right)).$$

In time, under a prescribed action $a(t)$:

$$x_*(t) = T^{-1}\left(y_*; z(t), a(t)\right) \sim \rho\left(: |z(t), a(t), g_*\right)$$

In action:

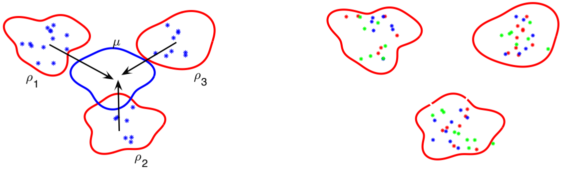
$$x_*(t) = T^{-1}\left(y_*; z_*, a\right) \sim \rho\left(: |z_*, a, g_*\right)$$

Sensitivity, attribution:

$$x_*(t) = T^{-1}\left(y_*; z, a_*\right) \sim \rho\left(: |z, a_*, g_*\right)$$

Factor discovery

Find additional latent factors z_l explaining, jointly with the known factors z_k, as much variability in x as possible:



$$z_l = \arg\min var\left[\mu(y) = T\#\rho(x|z)\right], \quad z = \{z_k, z_l\}$$

$$\text{but} \quad var[\rho] = \min C(T) + var[\mu], \quad \text{so}$$

$$\max_{z_l} \min_{T(x,z)} C(T)$$

To conclude

The barycenter problem provides a natural framework for inference and control, suitable for analysis at various levels of granularity/individualization.

Left out of this talk: how to formulate and solve the data-driven barycenter problem. That's where much of the math fun is!
Some ingredients: weak formulation of the push-forward condition, maps built from continuous flows, minimax problems.

Also left our of this talk: biomedical applications. Work in progress!

Much more to do.

Thanks!