Statistical inference with optimal transport

Vladimir Spokoiny

WIAS, April 12, 2018

The team and possible cooperation

Stochastic: Vladimir Spokoiny, Andrey Sobolevskiy, Alexey Kroshnin, Alexandra Suvorikova

Optimization: Alexander Gasnikov, Pavel Dvurechensky

Cooperation:

Massachusetts Institute of Technology

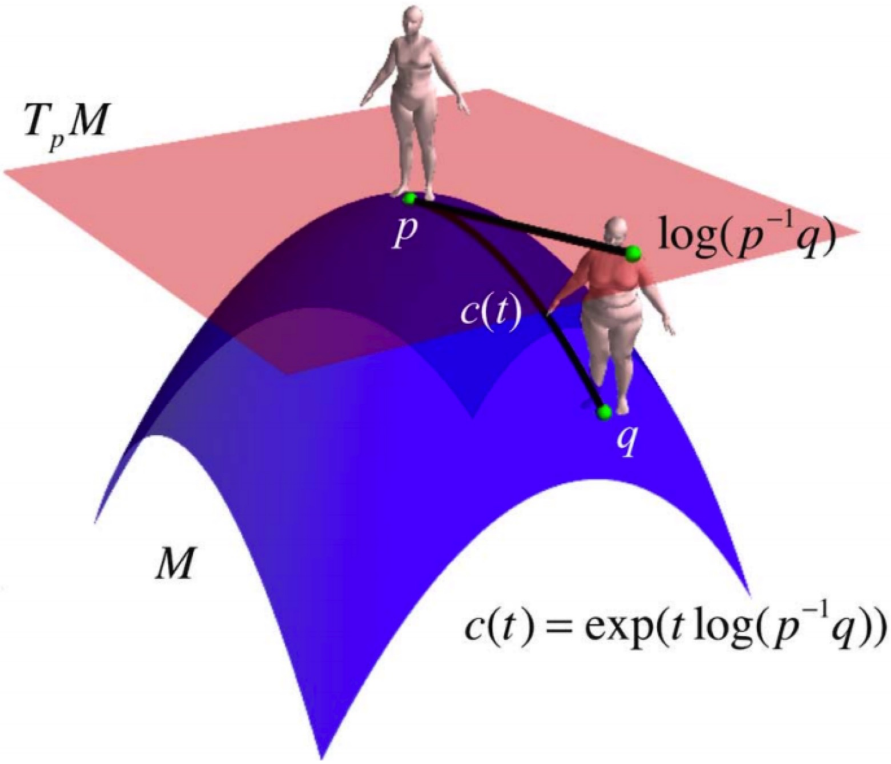
Université Paul Sabatier TOULOUSE III

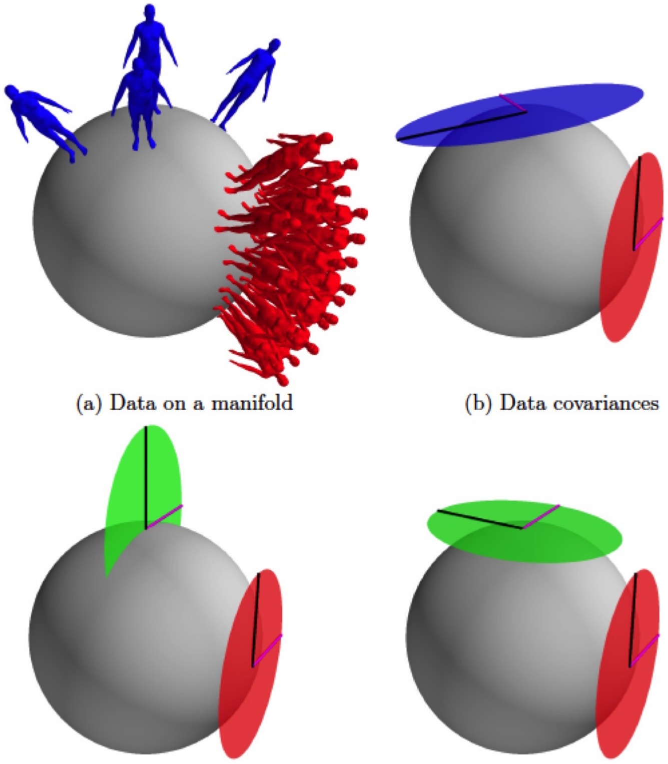
Weierstrass Institute for Applied Analysis and Stochastics

Analysis of data on manifolds

Modern data lives in manifolds:

- underlying geometry
- non-linearity of the space





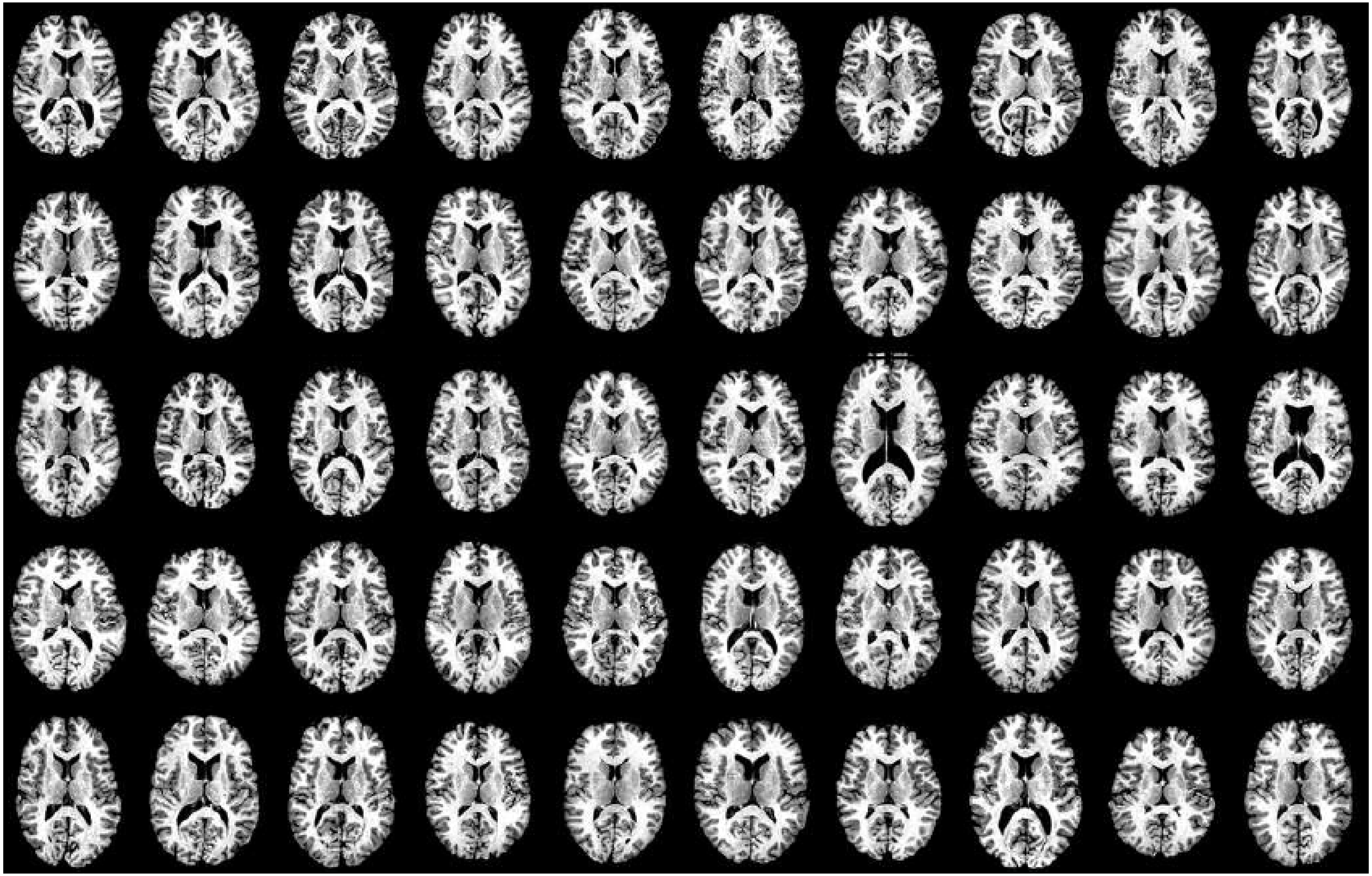
(a) Data on a manifold (b) Data covariances

(c) Linear translation (d) Covariance Transport

Source: ps.is.tuebingen.mpg.de

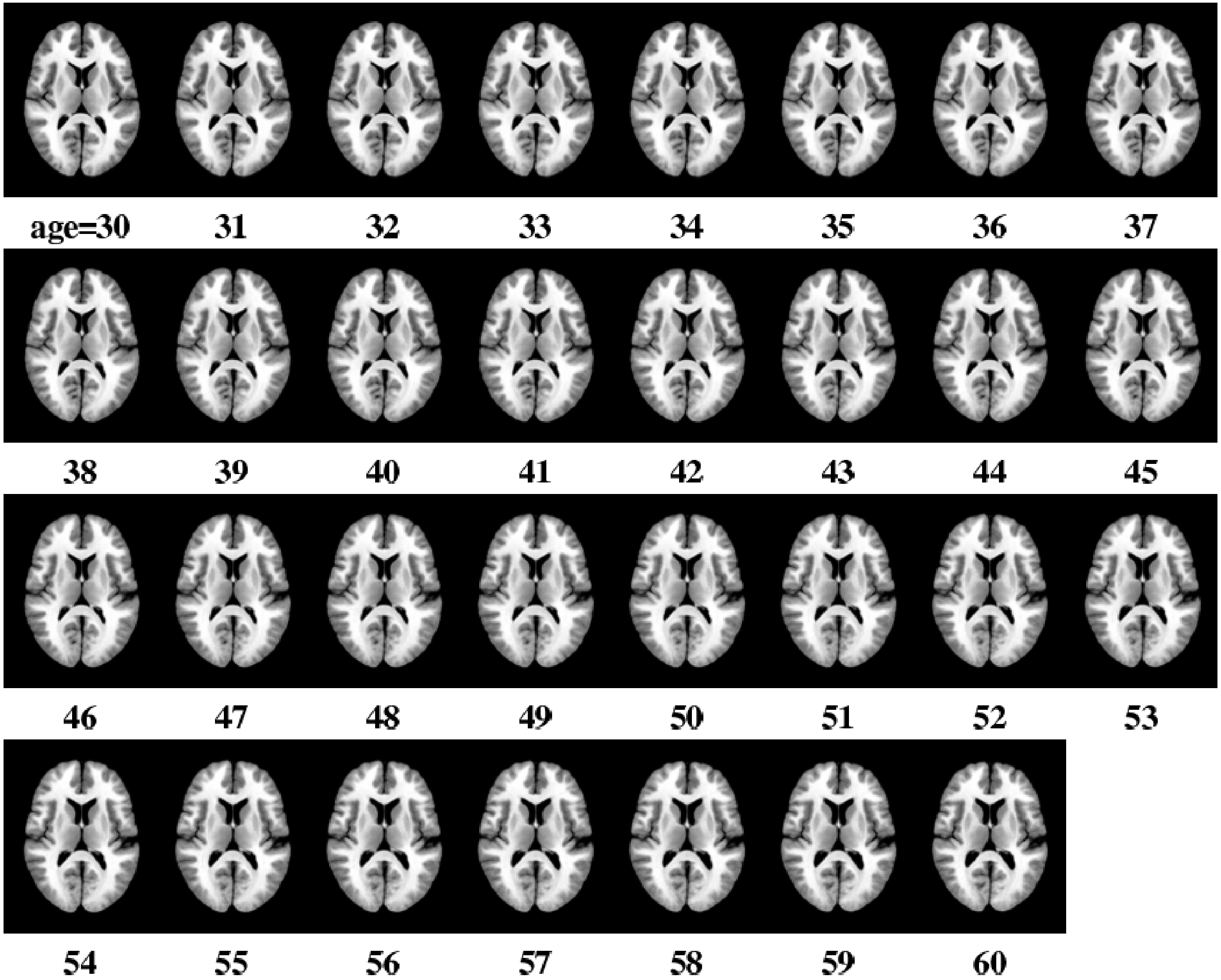
I. Motivation

Example 1: pattern extraction by averaging



Healthy brains, m + f, age of 21 to 72

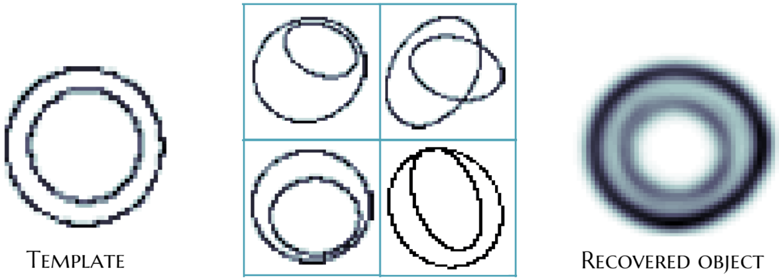
Example 1: pattern extraction by averaging



Typical ageing pattern, f

Example 1: pattern extraction by averaging

Recovery of the template object



Given $(\mathcal{X}, d)$ and $I\!P$ on $\mathcal{X}$ and iid $Y_1, ..., Y_n$, $Y_i \backsim I\!P$:
Template:

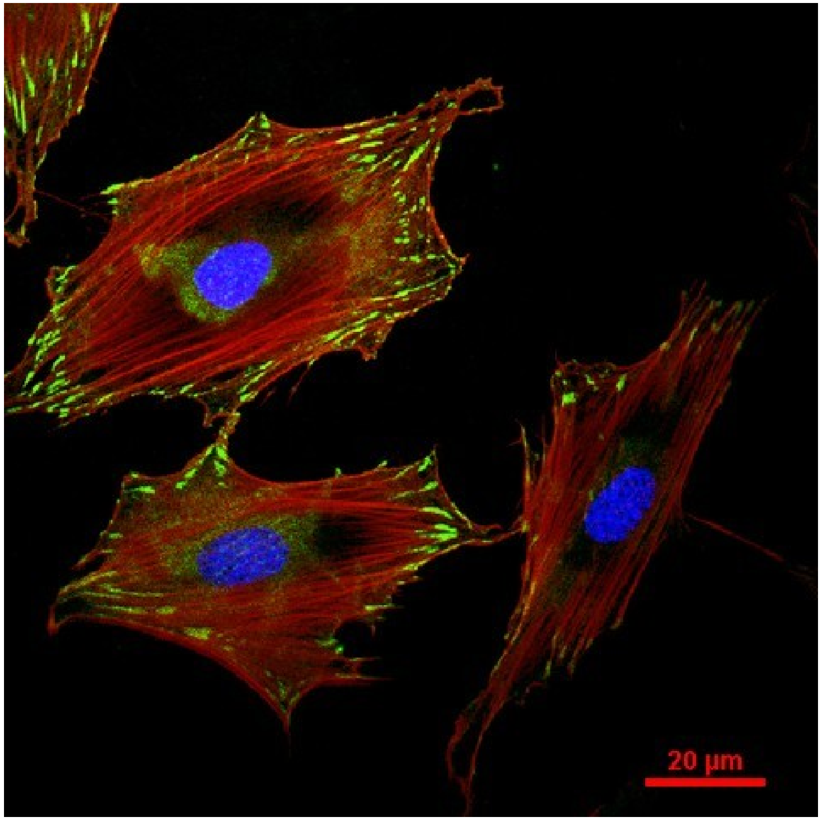
$$X^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{X \in \mathcal{X}} \int_{\mathcal{X}} d^2(X, Y) I\!P(dY)$$

Recovered object:

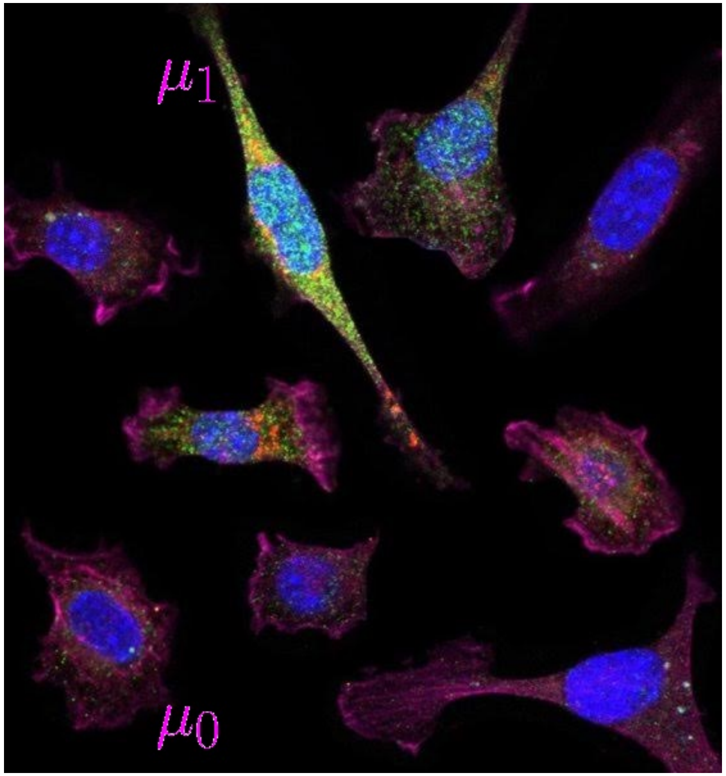
$$X_n \stackrel{\text{def}}{=} \operatorname*{argmin}_{X \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} d^2(X, Y_i).$$

- how to chose d
- confidence sets around X_n

Example 2: stem cell differentiation



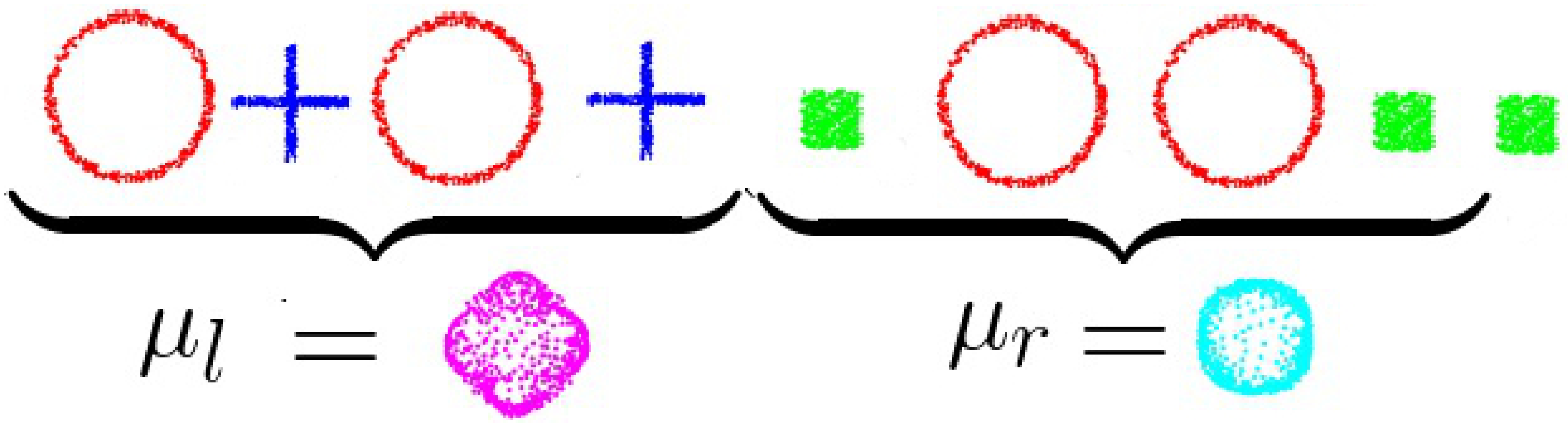
Mesenchymal stem cells, μ_0



Chondrogenesis, $t \in [0, 1]$

Detect time t when a cell specifies its "type".

Example 2: stem cell differentiation



$$\begin{cases} H_0: \text{ data is homogeneous} \iff t \text{ is not a change point} \\ H_1: \text{ data is non-homogeneous} \iff t \text{ is a change point} \end{cases}$$

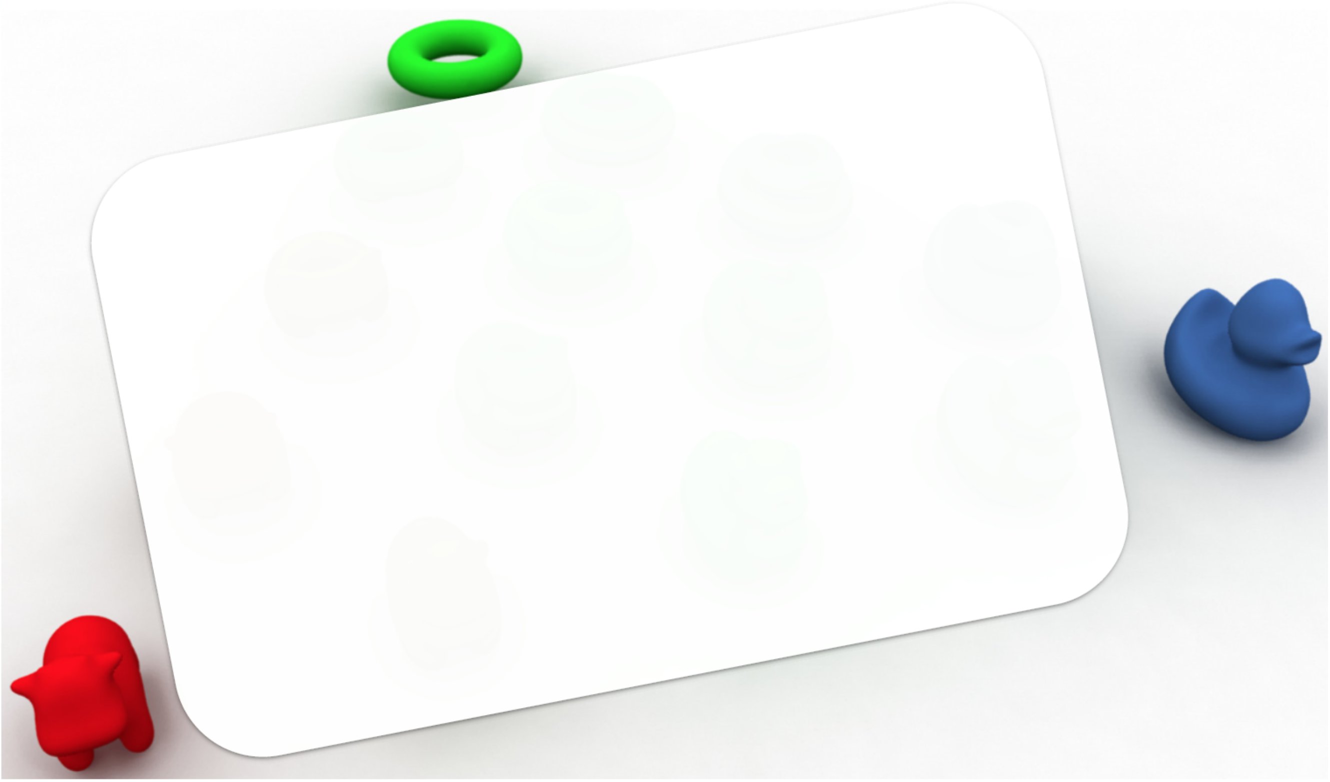
1. compute some cumulative statistics, e.g. means $\mu_l(t)$ and $\mu_r(t)$
2. compare them, e.g. $\text{dist}\big(\mu_l(t), \mu_r(t)\big)$
3. if $\text{dist}\big(\mu_l(t), \mu_r(t)\big) \geq \mathfrak{z}_\alpha(t)$, then H_0 is rejected

Goal: non-asymptotic data driven rejection level $\mathfrak{z}_\alpha$

Common features

All above mentioned examples have a common problem: data sets possess inner geometry

$\mathcal{Q}$: what is a good way to define a distance between such objects?



Common features

All above mentioned examples have a common problem: data sets possess inner geometry

$\mathcal{A}$: 2-Wasserstein distance might be a solution

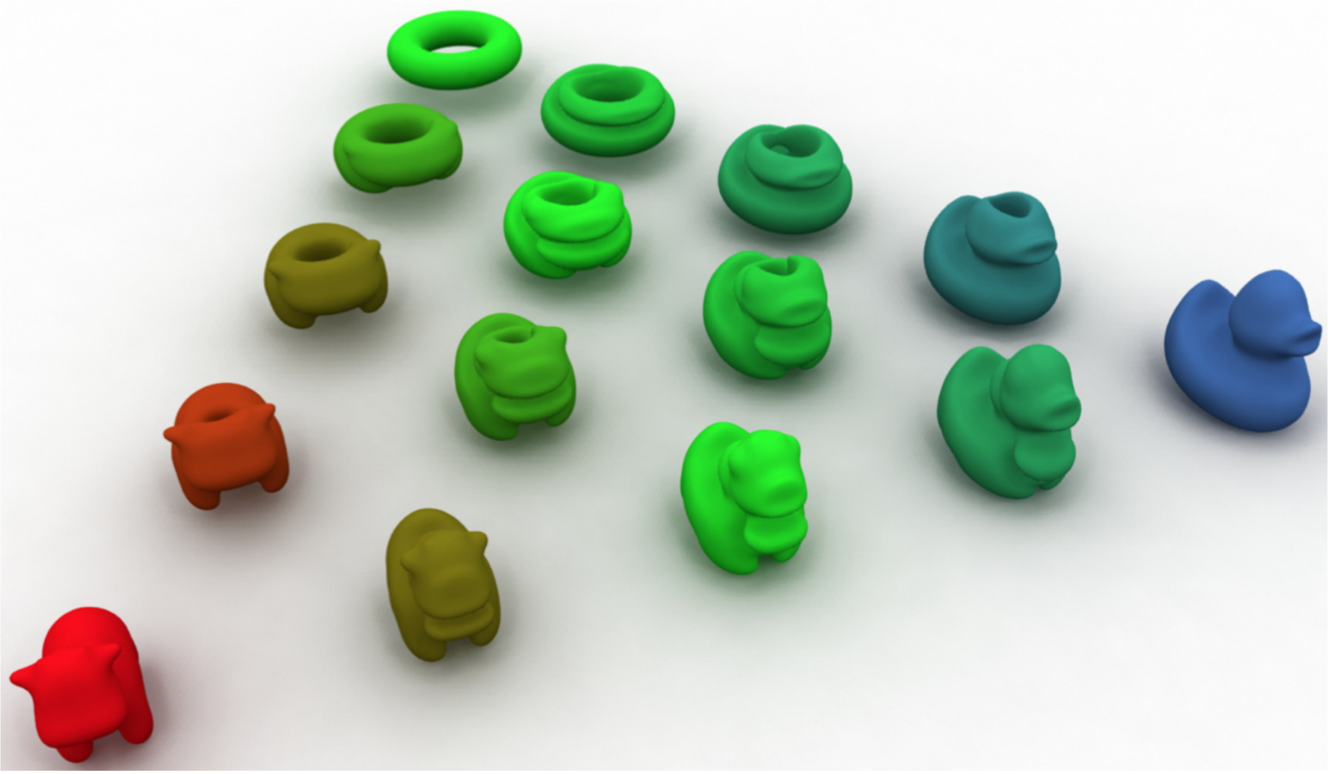


Image source: Marco Cuturi's OT tutorial

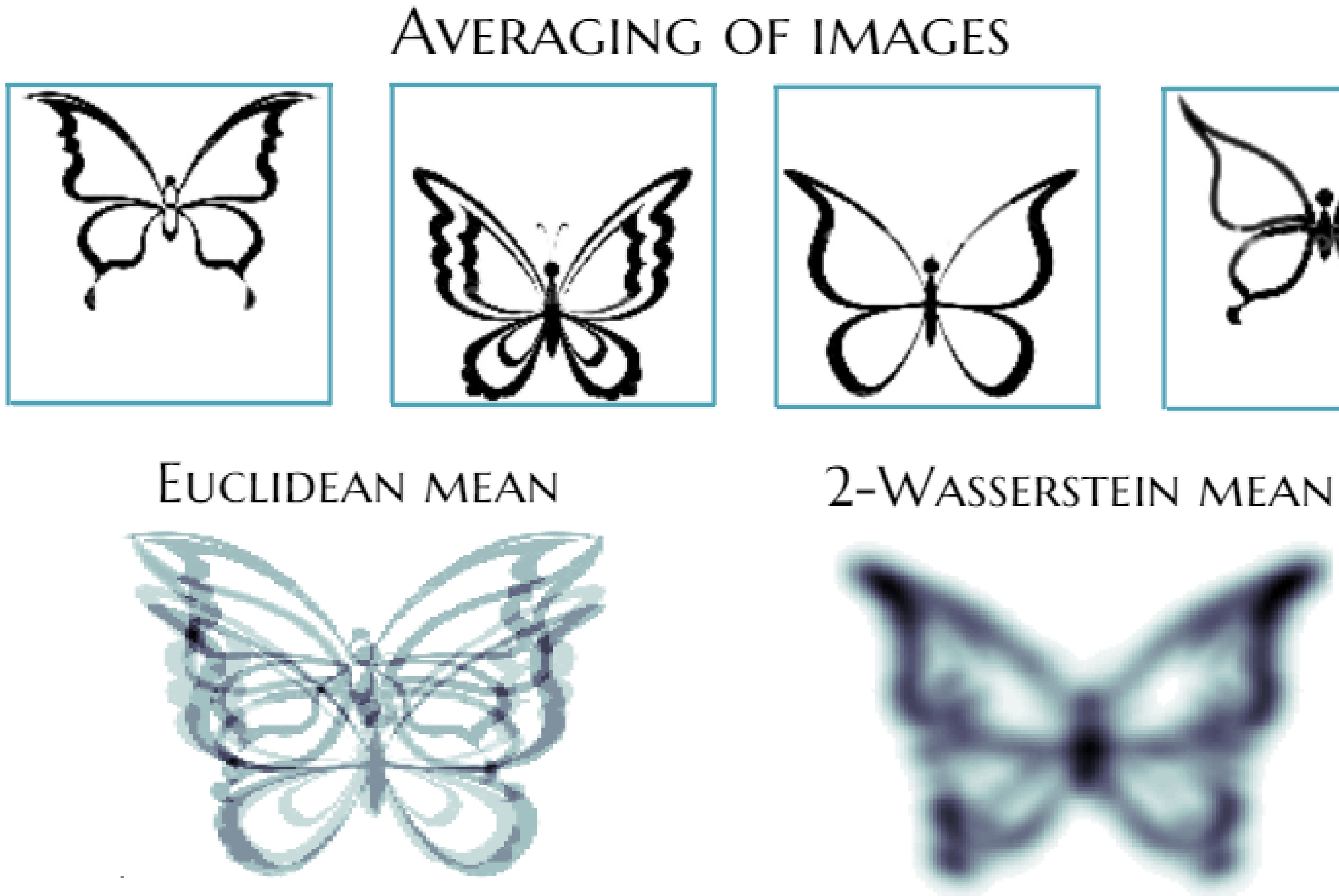
II. Introduction to OT

What is Optimal Transport?

The natural geometry for probability measures

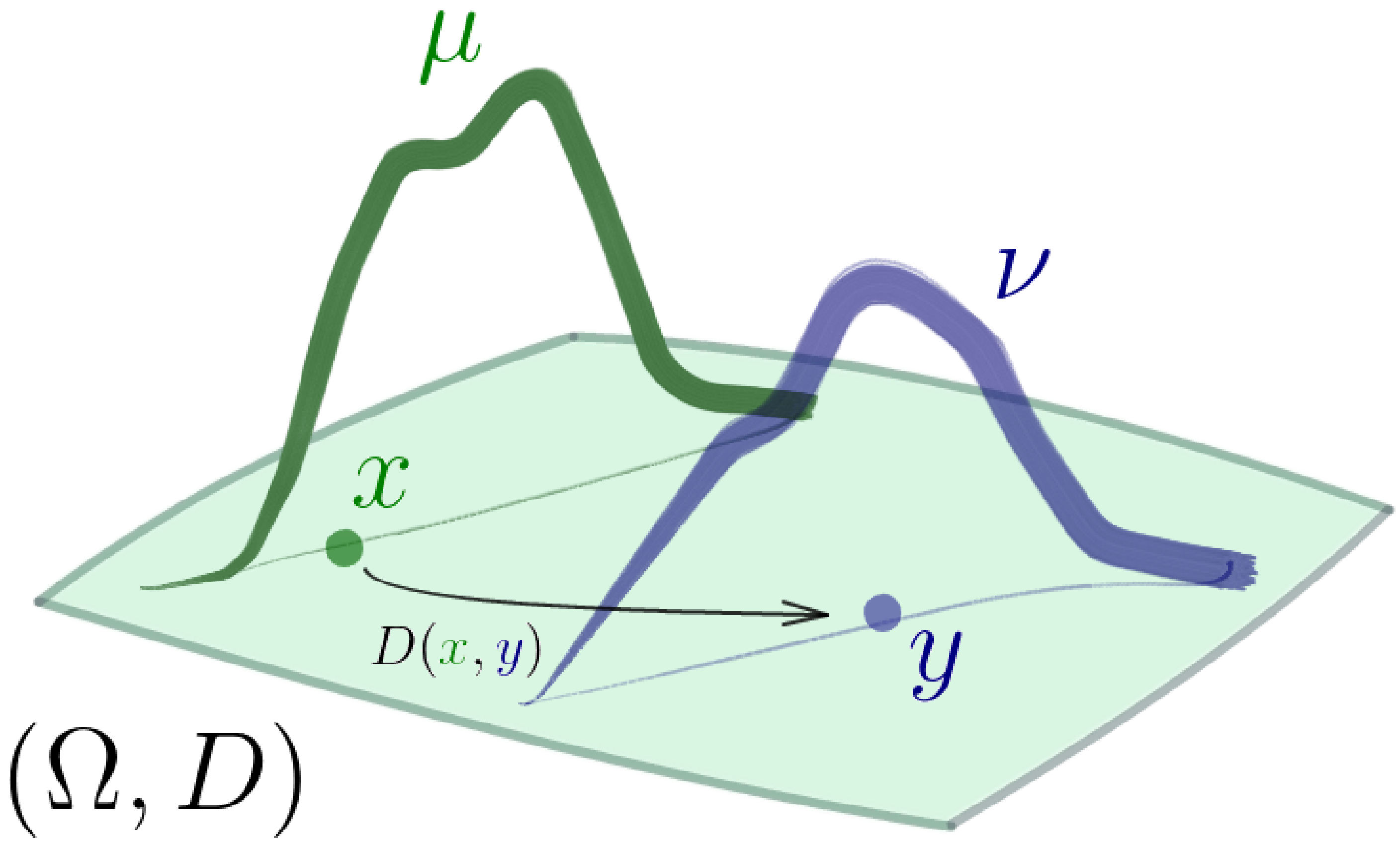
Image source: Marco Cuturi’s OT tutorial

The natural geometry for probability measures



Wasserstein distance and optimal transport

- W-*distance* — minimum amount of work that is necessary to convert μ into ν
- $D(x, y)$ — cost of transportation of unit mass from x to y



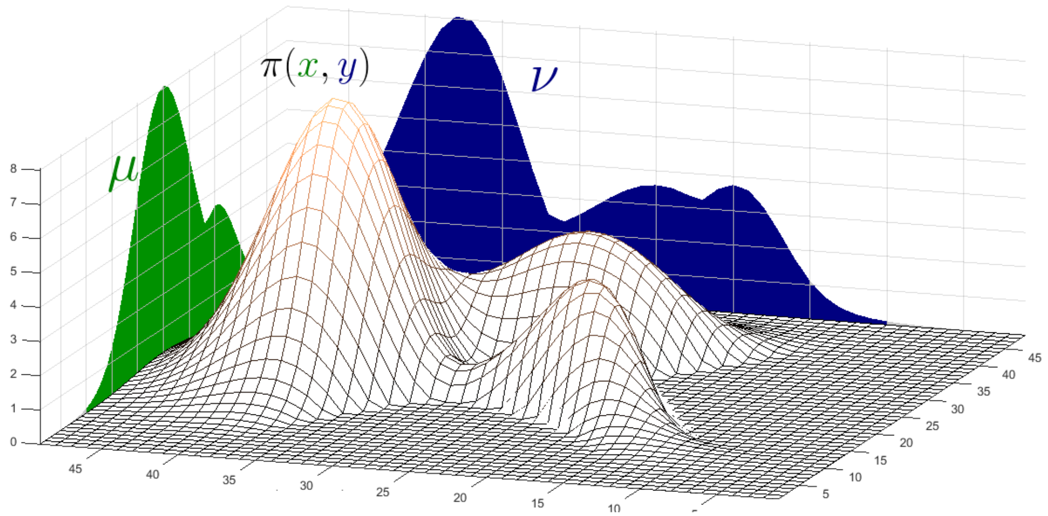
Space of images $(\mathcal{P}(\mathbb{R}^d), W_2)$

$\mathcal{P}_2(\mathbb{R}^d) = \{$Probability measures ν on $\mathbb{R}^d\}$,
metrizied by 2 -W distance:

$$W_2^2(\mu, \nu) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\pi(x, y)$$

$\Pi(\mu, \nu)$ – set of all prob. measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν

$$\begin{cases} \int_{\Omega \times \Omega} \pi(x, y) dy = \mu(x) \\ \int_{\Omega \times \Omega} \pi(x, y) dx = \nu(y) \end{cases}$$

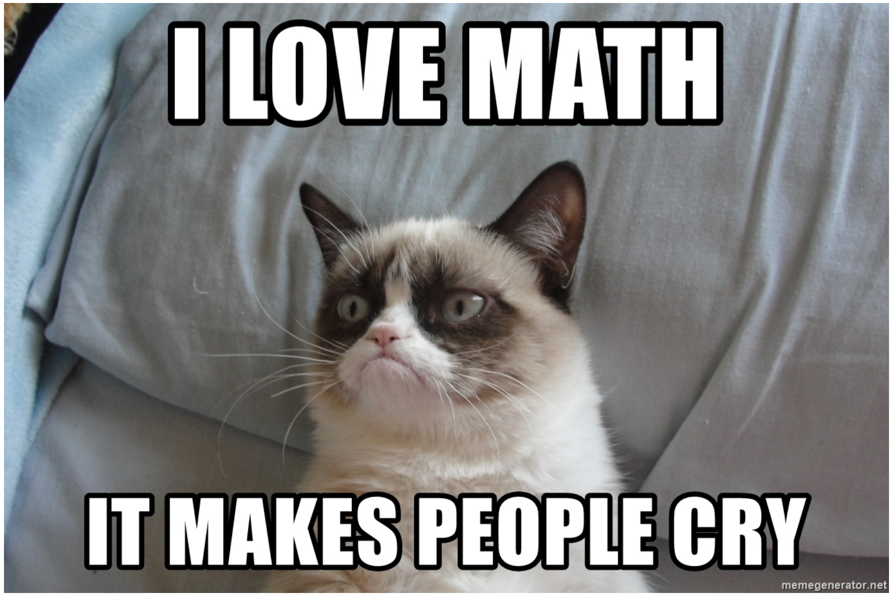


Recommended literature

- Villani, C. (2008). Optimal transport: old and new (Vol. 338). Springer Science & Business Media.
- Santambrogio, F. (2015). Optimal transport for applied mathematicians. Birkaeuser, NY, 99-102.
- Ambrosio, L. (2003). Lecture notes on optimal transport problems. In Mathematical aspects of evolving interfaces (pp. 1-52). Springer Berlin Heidelberg.

and many more...

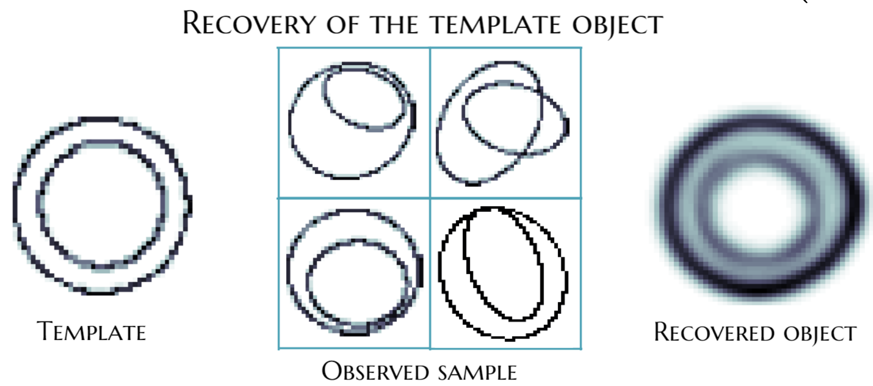
III. Some obtained results



III.a Non-asymptotic confidence sets

Non-asymptotic confidence sets, problem statement

$$\mathcal{P}_2(I\!R^d), W_2 \text{ – metric space and } I\!P \in I\!P\big(\mathcal{P}_2(I\!R^d)\big)$$



Wasserstein *population* barycenter

$$\mu^* \subseteq \operatorname*{argmin}_{\mu \in \mathcal{P}_2(I\!R^d)} \int_{\mathcal{P}_2(I\!R^d)} W_2^2(\mu, \nu) I\!P(d\nu)$$

$\nu_1, ..., \nu_n$ – observed random iid sample, $\nu_i \backsim I\!P$

Wasserstein *empirical* barycenter

$$\mu_n \subseteq \operatorname*{argmin}_{\mu \in \mathcal{P}_2(I\!R^d)} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i).$$

Confidence set around μ_n

Real world

$$T_n \stackrel{\text{def}}{=} \sqrt{n} W_2(\mu^*, \mu_n)$$

we do not know $\mathfrak{z}(\alpha)$:

$$\mathfrak{z}(\alpha) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathfrak{z}>0} \bigl\{ I\!\!P\bigl(T_n \geq \mathfrak{z}\bigr) \leq \alpha \bigr\}$$

Bootstrap world (mimics T_n)

$$T_n^\flat \stackrel{\text{def}}{=} \sqrt{n} W_2(\mu_n, \mu_n^\flat)$$

we know $\mathfrak{z}^\flat(\alpha)$:

$$\mathfrak{z}^\flat(\alpha) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathfrak{z}>0} \bigl\{ I\!\!P^\flat(T_n^\flat \geq \mathfrak{z}) \leq \alpha \bigr\}$$

Goal: replace $\mathfrak{z}(\alpha)$ with $\mathfrak{z}^\flat(\alpha)$

Result: Under some technical assumptions it holds with h.p. $I\!\!P$, $I\!\!P^\flat$

$$\bigl| I\!\!P\bigl(T_n \geq \mathfrak{z}^\flat(\alpha)\bigr) - \alpha \bigr| \leq \mathtt{C}/\sqrt{n}.$$

Multiplier bootstrap

Real world

Observed sample $\nu_i \overset{\text{iid}}{\backsim} \mathbb{P}$

W -*population* barycenter:

$$\mu^* = \operatorname*{argmin}_{\mu} \int W_2^2(\mu, \nu) \mathbb{P}(d\nu),$$

W -*empirical* barycenter:

$$\mu_n = \operatorname*{argmin}_{\mu} \frac{1}{n} \sum W_2^2(\mu, \nu_i),$$

$$T_n = \sqrt{n} W_2(\mu^*, \mu_n)$$

Bootstrap world, $u_i \backsim$ Po(1)

Training sample $\nu_i' \overset{\text{iid}}{\backsim} \mathbb{P}$

$W^\flat$ -*population* barycenter:

$$\mu_n = \operatorname*{argmin}_{\mu} \frac{1}{n} \sum W_2^2(\mu, \nu_i'),$$

$W^\flat$ -*empirical* barycenter:

$$\mu_n^\flat = \operatorname*{argmin}_{\mu} \frac{1}{n} \sum W_2^2(\mu, \nu_i') u_i,$$

$$T_n^\flat = \sqrt{n} W_2(\mu_n, \mu_n^\flat)$$

III.b Non-parametric 2-sample test

Two-sample testing

T-statistics: William S. Gosset (1908), economical way to monitor the quality of stout for Guinness

A 1904 brand poster (Guinness)

Two-sample test

$$X = (X_1, ..., X_n),\ X_i \overset{iid}{\backsim} \mu_X,$$

$$Y = (Y_1, ..., Y_m),\ Y_j \overset{iid}{\backsim} \mu_Y$$

Goal

$$H_0:\ \mu_X = \mu_Y, \quad H_1:\ \mu_X \neq \mu_Y$$

Underlying idea

Find some transformation T of μ_X , μ_Y :

- measure-preserving,
- yields universal critical level $\mathfrak{z}_{nm}$, which does not depend on μ_X , μ_Y .

Test:

$$\text{dist}\big(T\#\mu_X^n, T\#\mu_Y^m\big) \geq \mathfrak{z}_{nm} \longrightarrow H_1$$

1-D case: T is a ranking or quantile map (OT to $\mathcal{U}[0,1]$)

d-D case: T is a Monge map (optimal transport with $c(x,y) = \|x-y\|^2$)

$$T := \underset{T'_{\#}\mu=\nu}{\operatorname{argmin}} \int_{\text{supp}(\mu)} \big\|T'(x) - x\big\|^2 \mu(dx),$$

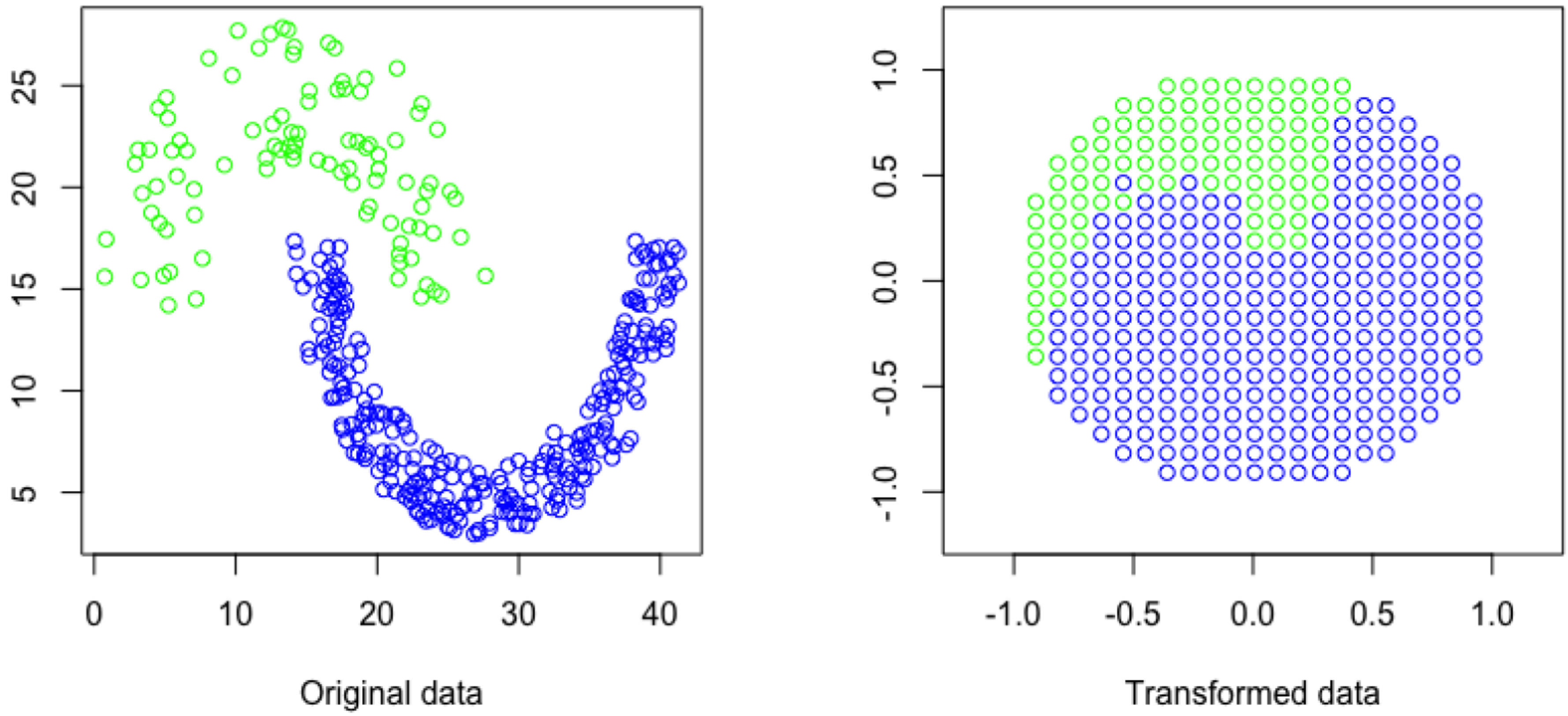
where ν some fixed reference measure and

$$\mu := \alpha\mu_X + (1-\alpha)\mu_Y$$

Intuition: a kind of d-dim quantile map

Observed distribution: $\mu = \alpha\mu_X + (1-\alpha)\mu_Y$

Reference distribution: $\nu = \mathcal{U}[B_0(1)]\,,\; B_0(1) \subset \mathbb{R}^d$



Idea of the method

Observed distribution: $\mu = \alpha\mu_X + (1-\alpha)\mu_Y$

Reference distribution: $\nu = \mathcal{U}[B_0(1)]\,,\ B_0(1) \subset \mathbb{R}^d$

Monge map:

$$T_{\#}\mu = \nu, \quad \nu = \alpha\nu_X + (1-\alpha)\nu_Y,$$

$$\nu_X\big(T(B)\big) = \mu_X(B), \quad \nu_Y\big(T(B)\big) = \mu_Y(B), \quad \forall B \in \mathcal{B}.$$

Under H_0:

$$\nu_X = \nu_Y = \mathcal{U}[B_0(1)]$$

Test:

$$D^T_{nm} \stackrel{\text{def}}{=} W_2(\nu_X^n, \nu_Y^m) \geq \mathfrak{z}_{nm} \ \Rightarrow \ H_1$$

IV. Open problems

What we are currently doing

The list of open problems and related literature: http://strlearn.ru/topics/

Hypothesis testing with Hellinger-Kantorovich distance

Responsible persons: Alexanda Suvorikova, Pavel Dvurechensky, Alexey Kroshnin, Andrey Sobolevskii, Vladimir Spokoiny

Domain adaptation using optimal transportation

Responsible persons: Alexanda Suvorikova, Pavel Dvurechensky, Alexey Kroshnin, Andrey Sobolevskii, Vladimir Spokoiny

Bootstrap for empirical barycenters

Responsible persons: Alexanda Suvorikova, Alexey Kroshnin, Andrey Sobolevskii, Vladimir Spokoiny

Two sample test for high dimensional data using Monge-Kantorovich transform

Responsible persons: Alexanda Suvorikova, Alexey Kroshnin, Andrey Sobolevskii, Vladimir Spokoiny

References:

[SAN15] Santambrogio F. Optimal transport for applied mathematicians. Birkäuser, NY, 2015.

[VIL08] Villani, C. Optimal transport: old and new. Springer Science and Business Media, 2008.

Thank you for your attention!