Distance to the Measure

Geometric inference for measures based on distance functions

The DTM-signature for a geometric comparison of metric-measure spaces from samples

Zhengchao Wan

the Ohio State University

wan.252@osu.edu

Geometric inference problem

Question

Given a noisy point cloud approximation C of a compact set $K \subset \mathbb{R}^d$, how can we recover geometric and topological informations about K, such as its curvature, boundaries, Betti numbers, etc. knowing only the point cloud C?

Inference using distance functions

One idea to retrieve information of a point cloud is to consider the R-**offset** of the point cloud - that is the union of balls of radius R whose center lie in the point cloud.

This offset makes good estimation of the topology, normal cones, and curvature measures of the underlying object, shown in previous literature.

The main tool used is a notion of **distance function**.

Inference using distance functions

For a compact $K \subset \mathbb{R}^d$,

$$\begin{aligned} d_K : \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto \operatorname{dist}(x, K) \end{aligned}$$

1. d_K is 1-Lipschitz.
2. d_K^2 is 1-semiconcave.
3. $\|d_K - d_{K'}\|_\infty \leq d_H(K, K')$.

Unfortunately, offset-based methods do not work well at all in the presence of outliers. For example, the number of connected components will be overestimated if one adds just a single data point far from the original point cloud.

Solution to outliers

Replace the distance function to a set K by a **distance function to a measure**. (Chazal, et al 2010)

Distance to a Measure

Notice
$d_K(x) = \min_{y \in K} \|x - y\| = \min\{r > 0 : B(x, r) \cap K \neq \emptyset\}$.

Given a probability measure μ on $\mathbb{R}^d$, we mimick the formula above:

$$\delta_{\mu,m} : x \in \mathbb{R}^d \mapsto \inf\{r > 0; \mu(\bar{B}(x, r)) > m\},$$

which is 1-Lipschitz but not semi-concave.

Distance to a Measure

Definition

For any measure μ with finite second moment and a positive mass parameter $m_0 > 0$, the distance function to measure (DTM) μ is defined by the formula:

$$d^2_{\mu,m_0} : \mathbb{R}^n \to \mathbb{R}, x \mapsto \frac{1}{m_0} \int_0^{m_0} \delta_{\mu,m}(x)^2 dm.$$

Recall $\delta_{\mu,m}(x) = \inf\{r > 0; \mu(\bar{B}(x,r)) > m\}$.

Example

Let $C = \{p_1, \cdots, p_n\}$ *be a point cloud and* $\mu_C = \frac{1}{n}\sum_i \delta_{p_i}$.
Then function δ_{μ_C, m_0} *with* $m_0 = k/n$ *evaluated at* $x \in \mathbb{R}^d$ *equal to the distance between* x *and its* k*th nearest neighbor in* C.
Given $S \subset C$ *with* $|S| = k$, *define*
$\mathrm{Vor}_C(S) = \{x \in \mathbb{R}^d : \forall p_i \notin S, d(x, p_i) > d(x, S).\}$, *which means its elements take* S *as their* k *first nearest neighbors in* C.

$$\forall x \in \mathrm{Vor}_C(S), d^2_{\mu_C, \frac{k}{n}}(x) = \frac{n}{k} \sum_{p \in S} \|x - p\|^2 .$$

Equivalent formulation

Proposition

1. *DTM is the minimal cost of the following problem:*

$$d_{\mu,m_0}(x) = \min_{\tilde{\mu}} \left\{ W_2\left(\delta_x, \frac{1}{m_0}\tilde{\mu}\right); \tilde{\mu}(\mathbb{R}^d) = m_0, \tilde{\mu} \leq \mu \right\}$$

2. *Denote the set of minimizers as* $\mathcal{R}_{\mu,m_0}(x)$. *Then for each* $\tilde{\mu}_{x,m_0} \in \mathcal{R}_{\mu,m_0}(x)$,
   - $\mathrm{supp}(\tilde{\mu}_{x,m_0}) \subset \bar{B}(x, \delta_{\mu,m_0}(x))$;
   - $\tilde{\mu}_{x,m_0}\big|_{B(x,\delta_{\mu,m_0}(x))} = \mu\big|_{B(x,\delta_{\mu,m_0}(x))}$;
   - $\tilde{\mu}_{x,m_0} \leq \mu$.

3. *For any* $\tilde{\mu}_{x,m_0} \in \mathcal{R}_{\mu,m_0}(x)$,

$$d^2_{\mu,m_0}(x) = \frac{1}{m_0}\int_{h\in\mathbb{R}^d} \|h - x\|^2 \, d\tilde{\mu}_{x,m_0} = W_2^2\left(\delta_x, \frac{1}{m_0}\tilde{\mu}_{x,m_0}\right).$$

Regularity Properties

Proposition

1. d^2_{μ,m_0} *is semiconcave, which means* $\|x\|^2 - d^2_{\mu,m_0}$ *is convex;*
2. d^2_{μ,m_0} *is differentiable at a point* x *iff* $\mathrm{supp}(\mu) \cap \partial B(x, \delta_{\mu,m_0}(x))$ *contains at most 1 point;*
3. d^2_{μ,m_0} *is differentiable almost everywhere in* $\mathbb{R}^d$ *in Lebesgue measure. (directly from item 1)*
4. d_{μ,m_0} *is 1-Lipschitz.*

Stability of DTM

Theorem (DTM stability theorem)

If μ, ν are two probability measures on $\mathbb{R}^d$ and $m_0 > 0$, then

$$\|d_{\mu,m_0} - d_{\nu,m_0}\|_\infty \leq \frac{1}{\sqrt{m_0}} W_2(\mu, \nu).$$

Uniform Convergence of DTM

Lemma

If μ is a compactly-supported measure, then d_S is the uniform limit of d_{μ,m_0} as m_0 converges to 0, where $S = \mathrm{supp}(\mu)$, i.e.,

$$\lim_{m_0 \to 0} \|d_{\mu,m_0} - d_S\|_\infty = 0.$$

Remark

If μ has dimension at most $k > 0$, i.e. $\mu(B(x,\epsilon)) \geq C\epsilon^k, \forall x \in S$ when ϵ is small, then we can control the convergence speed:

$$\|d_{\mu,m_0} - d_S\|_\infty = O(m_0^{1/k}).$$

Reconstruction from noisy data

If μ is a probability measure of dimension at most $k > 0$ with compact support $K \subset \mathbb{R}^d$, and μ' is another probability measure, one has

$$\begin{aligned}\left\| d_K - d_{\mu',m_0} \right\|_\infty &\leq \| d_K - d_{\mu,m_0} \|_\infty + \left\| d_{\mu,m_0} - d_{\mu',m_0} \right\|_\infty \\ &\leq O(m_0^{1/k}) + \frac{1}{\sqrt{m_0}} W_2(\mu,\mu').\end{aligned}$$

Reconstruction from noisy data

Define α-**reach** of K, $\alpha \in (0,1]$ as $r_\alpha(K) = \inf\{d_K(x) > 0 : \|\nabla_x d_K\| \leq \alpha\}$.

Theorem

Suppose μ has dimension at most k with compact support $K \subset \mathbb{R}^d$ such that $r_\alpha(K) > 0$ for some α. For any $0 < \eta < r_\alpha(K)$, $\exists m_1 = m_1(\mu, \alpha, \eta) > 0$ and $C = C(m_1) > 0$ such that: for any $m_0 < m_1$ and μ' satisfying $W_2(\mu, \mu') < C\sqrt{m_0}$, $d^{-1}_{\mu', m_0}([0, \eta])$ is **homotopy equivalent** *to the offset $d_K^{-1}([0, r])$ for $0 < r < r_\alpha(K)$.*

Example

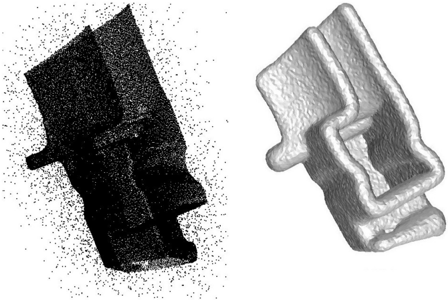


Figure: On the left, a point cloud sampled on a mechanical part to which 10% of outliers have been added- the outliers are uniformly distributed in a box enclosing the original point cloud. On the right, the reconstruction of an isosurface of the distance function d_{μ_C, m_0} to the uniform probability measure on this point cloud.

How to determine that two N-samples are from the same underlying space?

DTM based asymptotic statistical test. (Brecheteau 2017)

DTM-signature

Definition (DTM-signature)

The **DTM-signature** *associated to some mm-space* (X, δ, μ), *denoted* $d_{\mu,m}(\mu)$, *is the distribution of the real valued random variable* $d_{\mu,m}(Y)$ *where* Y *is some random variable of law* μ.

Stability of DTM

Proposition

Given two mm-spaces $(X, \delta_X, \mu), (Y, \delta_Y, \nu)$, we have

$$W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu)) \leq \frac{1}{m} GW_1(X, Y).$$

Proposition

If $(X, \delta_X, \mu), (Y, \delta_Y, \nu)$ are embedded into some metric space (Z, δ), then we can upper bound $W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu))$ by

$$W_1(\mu, \nu) + \min\{\|d_{\mu,m} - d_{\nu,m}\|_{\infty,\mathrm{supp}(\mu)}, \|d_{\mu,m} - d_{\nu,m}\|_{\infty,\mathrm{supp}(\nu)}\},$$

and more generally by $(1 + \frac{1}{m})W_1(\mu, \nu)$.

Non discriminative example

There are non isomorphic $(X, \delta, \mu), (X, \delta, \nu)$ with $d_{\mu,m}(\mu) = d_{\nu,m}(\nu)$.

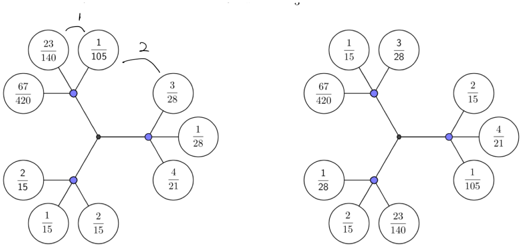


Figure: Each cluster has the same weight $1/3$.

Discriminative results

Proposition

Let $(O, \|\|_2, \mu_O), (O', \|\|_2, \mu_{O'})$ be two mm-spaces, for O, O' two non-empty bounded open subset of $\mathbb{R}^d$ satisfying $O = (\bar{O})^\circ$ and $O = (\bar{O}')^\circ$, $\mu_O, \mu_{O'}$ uniform measures. A lower bound for $W_1(d_{\mu_O,m}(\mu_O), d_{\mu_{O'},m}(\mu_{O'}))$ is given by:

$$C|\mathrm{Leb}_d(O)^{\frac{1}{d}} - \mathrm{Leb}_d(O')^{\frac{1}{d}}|,$$

where C depends on m, ϵ, O, O', d.

Remark

DTM can be discriminative under some conditions.

Statistic test

Given two N-samples from the mm-spaces $(X, \delta, \mu), (Y, \gamma, \nu)$, we want to build a algorithm using these two samples to test the null hypothesis:

$$H_0 \text{ "two mm-spaces } X, Y \text{ are isomorphic"},$$

against its alternative:

$$H_1 \text{ "two mm-spaces } X, Y \text{ are not isomorphic"},$$

The test proposed in the paper is based on the fact that the DTM-signature associated to two isomorphic mm-spaces are equal, which leads to $W_1(d_{\mu,m}(\mu), d_{\nu,m}(\nu)) = 0$.

Idea

Given two N-samples from the mm-spaces $(X,\delta,\mu),(Y,\gamma,\nu)$, choose randomly two n-samples from them respectively, which gives four empirical measures, $\hat{\mu}_n,\hat{\mu}_N,\hat{\nu}_n,\hat{\nu}_N$.

Test statistic: $T_{N,n,m}(\mu,\nu)=\sqrt{n}W_1(d_{\hat{\mu}_N,m}(\hat{\mu}_n),d_{\hat{\nu}_N,m}(\hat{\nu}_n))$.

Denote the law of $T_{N,n,m}(\mu,\nu)$ as $\mathcal{L}_{N,n,m}(\mu,\nu)$.

Lemma

If two mm-spaces are isomorphic, then $\mathcal{L}_{N,n,m}(\mu,\nu) = \mathcal{L}_{N,n,m}(\nu,\nu) = \mathcal{L}_{N,n,m}(\mu,\mu) = \frac{1}{2}\mathcal{L}_{N,n,m}(\mu,\mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu,\nu)$.

Remark

$\frac{1}{2}\mathcal{L}_{N,n,m}(\mu,\mu) + \frac{1}{2}\mathcal{L}_{N,n,m}(\nu,\nu)$ *is the distribution of* $ZT_{N,n,m}(\mu,\mu) + (1-Z)T_{N,n,m}(\nu,\nu)$*, where* Z *is another independent random variable with Bernoulli distribution.*

The α-quantile $q_{\alpha,N,n}$ of $\frac{1}{2}\mathcal{L}_{N,n,m}(\mu,\mu)+\frac{1}{2}\mathcal{L}_{N,n,m}(\nu,\nu)$ will be approximated by the α-quantile $\hat{q}_{\alpha,N,n}$ of $\frac{1}{2}\mathcal{L}^*_{N,n,m}(\hat{\mu}_N,\hat{\mu}_N)+\frac{1}{2}\mathcal{L}^*_{N,n,m}(\hat{\nu}_N,\hat{\nu}_N)$.

Here $\mathcal{L}^*_{N,n,m}(\hat{\mu}_N,\hat{\mu}_N)$ stands for the distribution of $T_{N,n,m}(\hat{\mu}_N,\hat{\mu}_N)=\sqrt{n}W_1(d_{\hat{\mu}_N,m}(\mu^*_n),d_{\hat{\mu}_N,m}({\mu'^*}_n))$ conditionally to $\hat{\mu}_N$, where μ^*_n and ${\mu'^*}_n$ are two independent n-samples of law $\hat{\mu}_N$.

We deal with the **test**:

$$\phi_N=1_{T_{N,n,m}(\mu,\nu)\geq\hat{q}_{\alpha,N,n}}.$$

Bootstrap method

Algorithm 1: Test Procedure

Input : P and Q N-samples from μ (respectively ν), N, n, m, α, N_{MC} even ;
`# Compute` T `the test statistic`
Take P' a random subset of P of size n ;
Take Q' a random subset of Q of size n ;
$T \leftarrow \sqrt{n} W_1(\mathrm{d}_{\mathbb{1}_P,m}(\mathbb{1}_{P'}), \mathrm{d}_{\mathbb{1}_Q,m}(\mathbb{1}_{Q'}))$;
`# Compute` *boot* `a` N_{MC}`-sample from the bootstrap law`
$dtmP \leftarrow (\mathrm{d}_{\mathbb{1}_P,m}(x))_{x\in P}$;
$dtmQ \leftarrow (\mathrm{d}_{\mathbb{1}_Q,m}(x))_{x\in Q}$;
Let *boot* be empty ;
for j **in** $1\,..\lfloor N_{MC}/2 \rfloor$:
  Let $dtmP_1$ and $dtmP_2$ be two independent n-samples from $\mathbb{1}_{dtmP}$;
  Let $dtmQ_1$ and $dtmQ_2$ be two independent n-samples from $\mathbb{1}_{dtmQ}$;
  Add $\sqrt{n} W_1(\mathbb{1}_{dtmP_1}, \mathbb{1}_{dtmP_2})$ and $\sqrt{n} W_1(\mathbb{1}_{dtmQ_1}, \mathbb{1}_{dtmQ_2})$ to *boot* ;
`# Compute` *qalph*`, the` α`-quantile of` *boot*
Let *qalph* be the $\lfloor N_{MC} - N_{MC} \times \alpha \rfloor$th smallest element of *boot* ;
Output : $(T \geq qalph)$

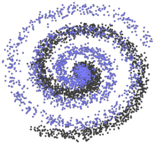
Asymptotic level α

For properly chosen n depending on N, for example, $N = cn^{\rho}$, with $\rho > \frac{\max\{d,2\}}{2}$, test is of asymptotic level α, i.e.

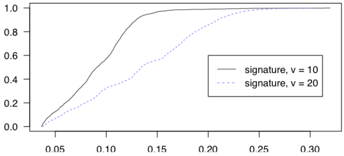
$$\limsup_{N\to\infty} \mathbb{P}_{(\mu,\nu)\in H_0}(\phi_N = 1) \leq \alpha.$$

Numerical illustrations

μ_v: distribution of $(R\sin(vR) + 0.03M, R\cos(vR) + 0.03M')$ with R, M, M' independent variables; M and M' from the standard normal distribution and R uniform on $(0, 1)$.
Sample $N = 2000$ points from two measure, choose $\alpha = 0.05, m = 0.05, n = 20, N_{MC} = 1000$.



Numerical illustrations



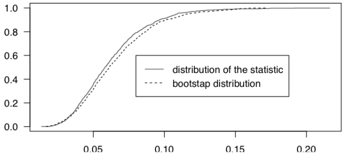


Figure: Left: DTM-signature estimates. Right: Bootstrap validity, $v = 10$.

v	15	20	30	40	100
type I error **DTM**	0.050	0.049	0.051	0.044	0.051
power **DTM**	0.525	0.884	0.987	0.977	0.985
power **KS**	0.768	0.402	0.465	0.414	0.422

Figure: Type 1 error and power approximations by repeating 1000 times.

Thank you!