Geometry Understanding in Higher Dimensions

CHAIRE D'INFORMATIQUE ET SCIENCES NUMÉRIQUES

Collège de France - June 2017

Statistics and Topological Data Analysis

Bertrand MICHEL

Ecole Centrale de Nantes - Lab. de mathématiques Jean Leray

Laboratoire de Mathémariques Jean Leray

UMR 6629 - Nantes

Centrale Nantes

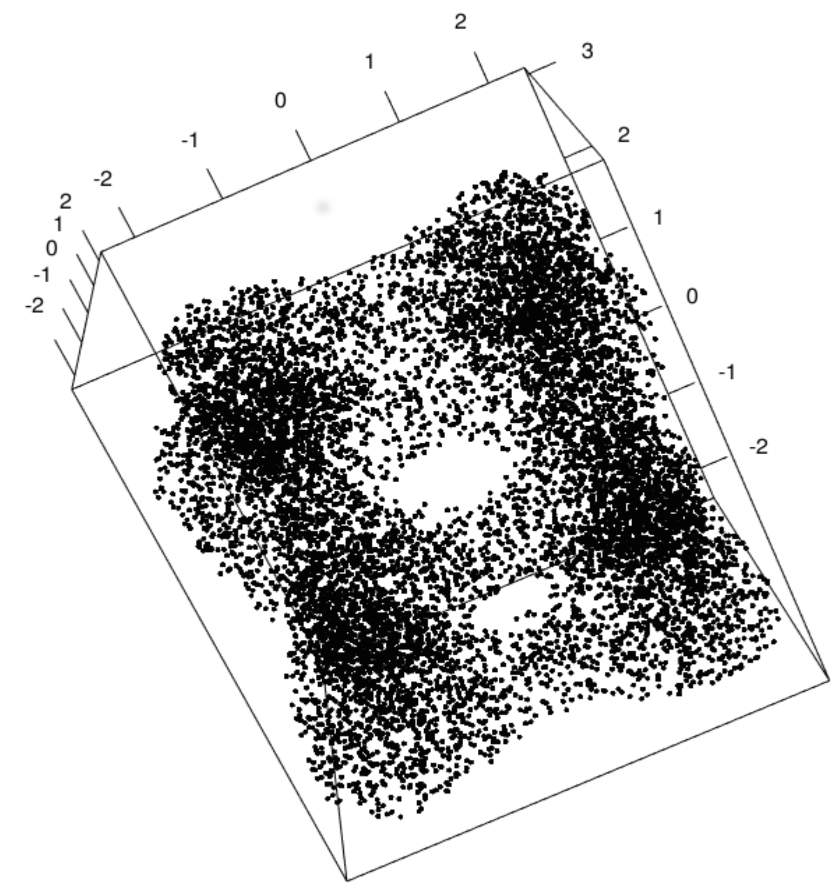
informatics mathematics

Introduction : Topological Data Analysis and Statistics

Topological Data Analysis and Topological Inference

- The aim of TDA is to infer relevant qualitative and quantitative **topological structures** (clusters, holes ...) directly from the data.

    - data : typically point cloud $\mathbb{X}_n$

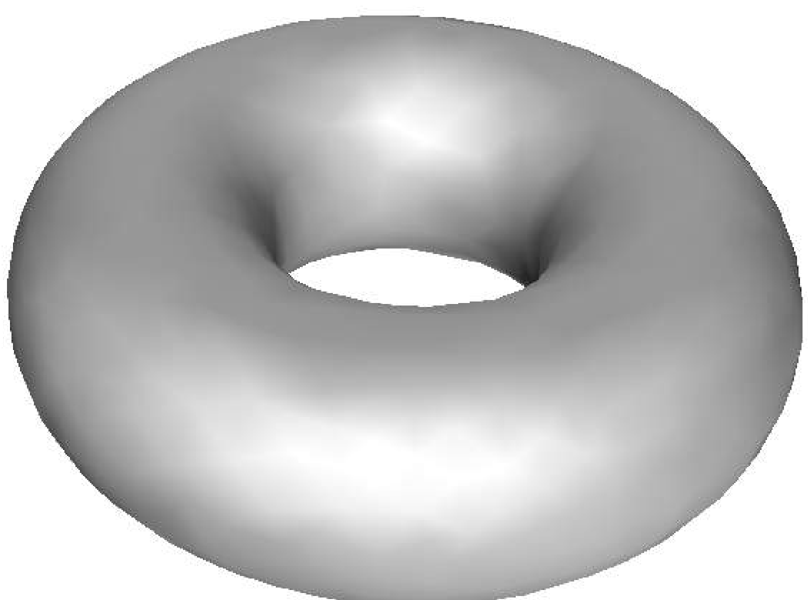


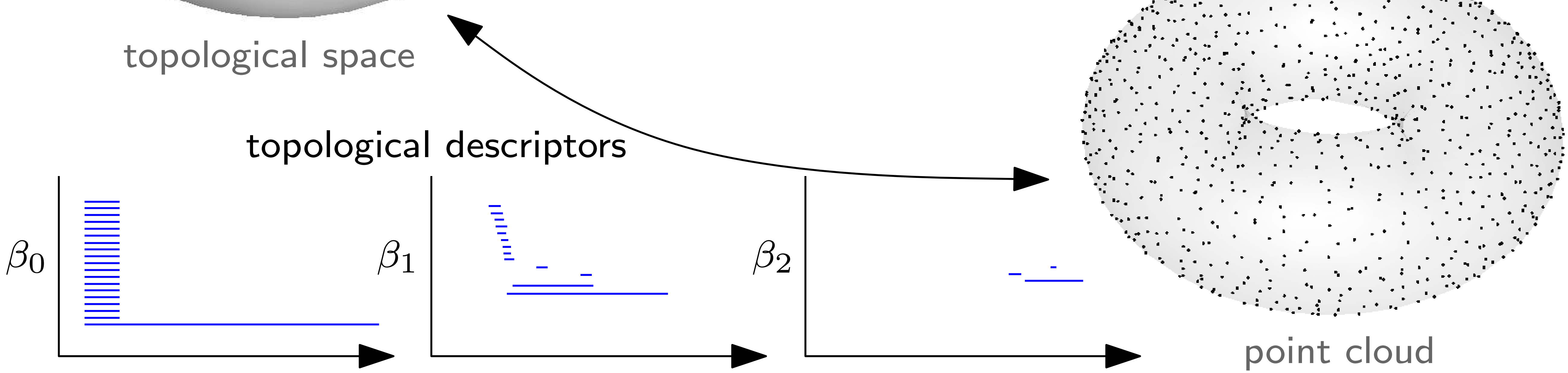
- Two popular methods in TDA : **Mapper algorithm** [Singh et al., 2007] and **persistent homology** [Edelsbrunner et al., 2002].

Topological Data Analysis (TDA)

Why is topology interesting for data analysis?

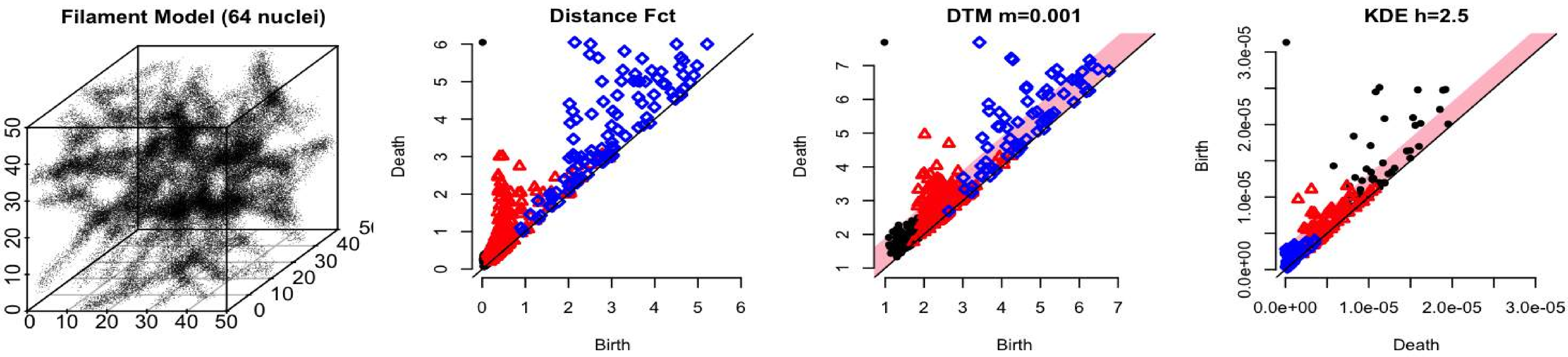
- multiscale
- compact
- invariant under coordinate changes
- stable with respect to (small) perturbations
- informative





Topological Data Analysis (TDA)

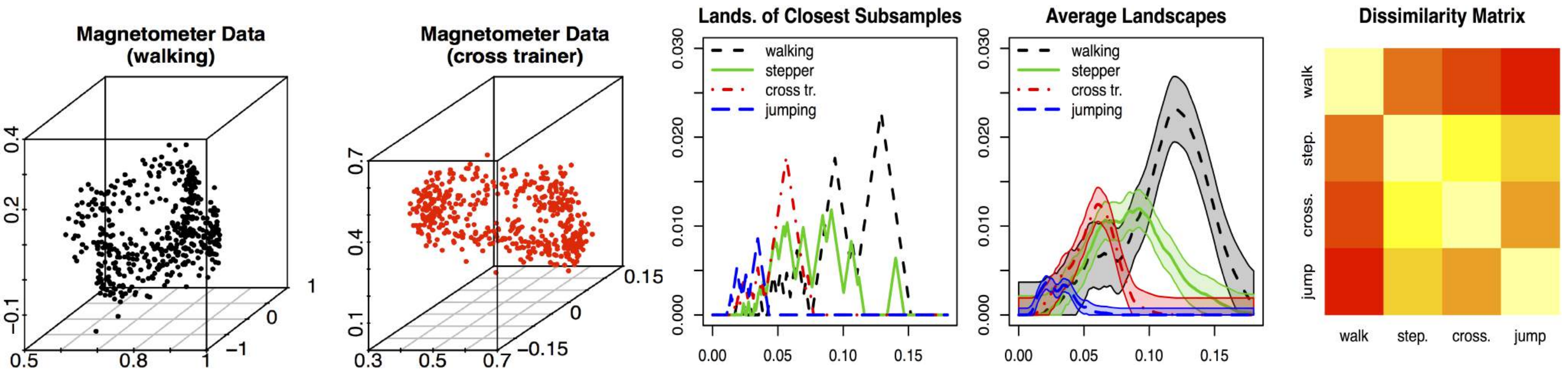
- For **exploratory analysis**, visualization



Topological Data Analysis (TDA)

- For **exploratory analysis**, visualization
- For **feature extraction** and statistical learning

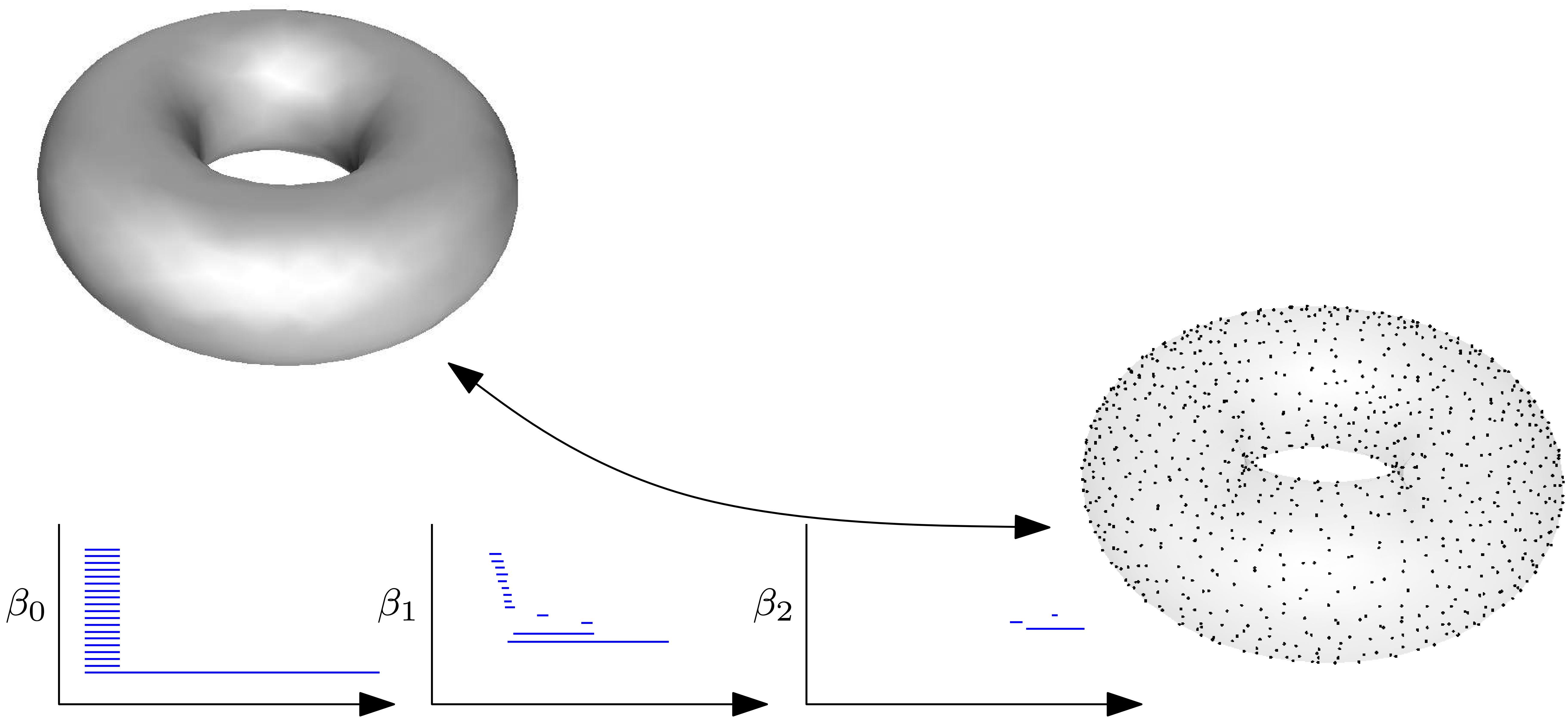
raw data	→	• Topological descriptors • Geometric descriptors • Other signatures	→	• supervised methods • unsupervised methods



Statistics, Learning and TDA

A **statistical approach to TDA** means that :

- we consider data as generated from an unknown distribution
- the inferred topological features by TDA methods are seen as estimators of topological quantities describing an underlying object.



Statistics, Learning and TDA

Directions of research (non-exhaustive list):

- Consistency / convergence of TDA methods: [Chazal15 JMLR], [Bobrowski 17 Bernouilli]
- Confidence regions for TDA [Fasy 14 AoS] [Chazal 15 JOCG]
- Central tendency for persistent homology [Turner 14 DCG] [Fasy15 Nips]
- Robust methods for TDA [Chazal 17, EJS Chazal 17 JMLR]
- Representations of persistence in Euclidean spaces [Bubenik15 JMLR] [Adams15]
- Develop kernels for topological descriptors [Reininghaus 15 IEEE] [Carriere 17 ICML]
- Statistical analysis of Mapper [Carriere 17]
- ...

Homology
and
Persistent homology

Topological Stability and Regularity

Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

Topological Stability and Regularity

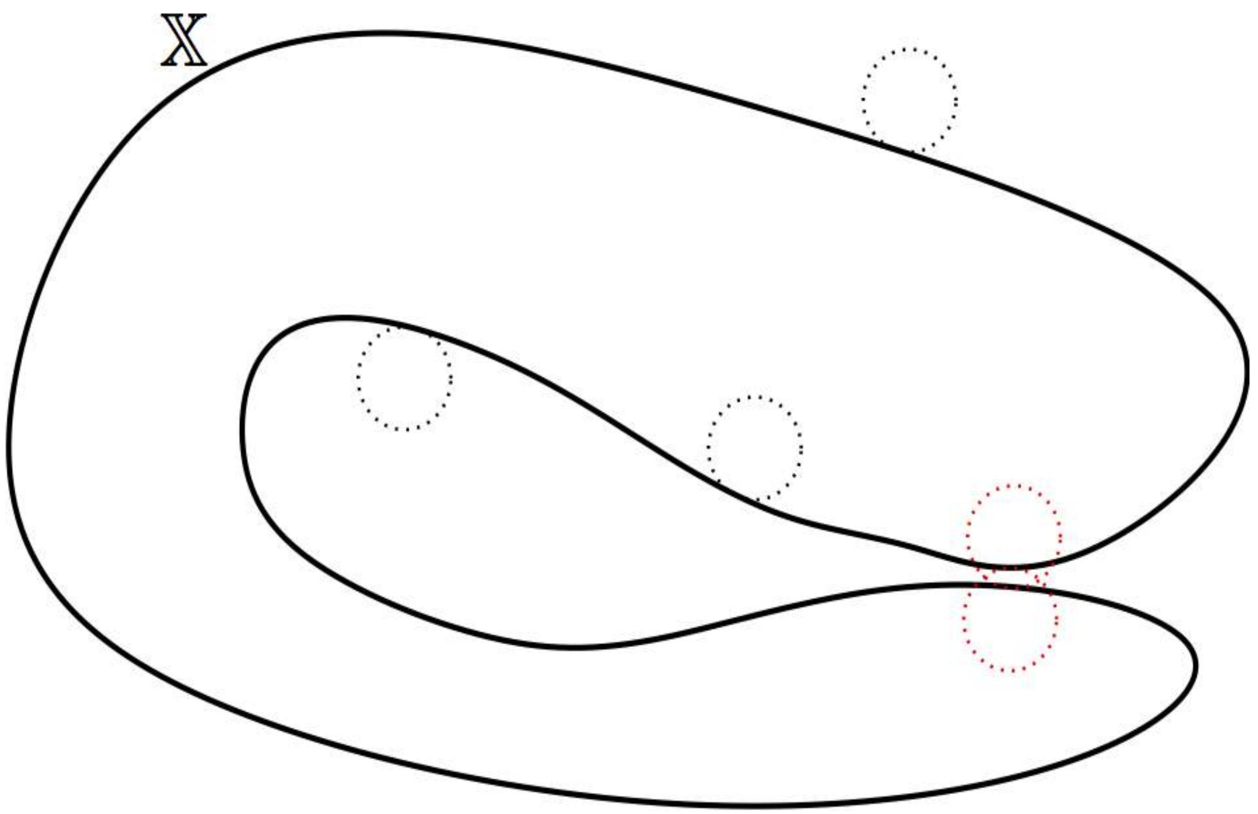
Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

- The *local feature size* is a local notion of regularity :
  For $x \in \mathbb{X}$, $\mathrm{lfs}_{\mathbb{X}}(x) := d\left(x, \mathcal{M}(\mathbb{X}^c)\right)$.

- The global version of the local feature size is the *reach* [Federer, 1959] :

$$\kappa(\mathbb{X}) = \inf_{x \in \mathbb{X}^c} \mathrm{lfs}_{\mathbb{X}}(x).$$

  The reach is small if either $\mathbb{X}$ is not smooth or if $\mathbb{X}$ is close to being self-intersecting.



- Weak feature size and its extensions [Chazal and Lieutier, 2007] (by considering the critical values of $d_{\mathbb{X}}$).

Topological Stability and Regularity

Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

$$\mathrm{d_H}(\mathbb{X}, \mathbb{Y}) = \inf \{\alpha \geq 0 \mid \mathbb{X} \subset \mathbb{Y}^\alpha \text{ and } \mathbb{Y} \subset \mathbb{X}^\alpha\}$$

Example :

Theorem [Chazal and Lieutier, 2007]: Let $\mathbb{X}$ and $\mathbb{Y}$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $\mathrm{d_H}(\mathbb{X}, \mathbb{Y}) < \varepsilon$, $\mathrm{wfs}(\mathbb{X}) > 2\varepsilon$ and $\mathrm{wfs}(\mathbb{Y}) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $\mathbb{X}^\alpha$ and $\mathbb{Y}^\beta$ are homotopy equivalent.

Topological Stability and Regularity

Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

$$\mathrm{d}_{\mathsf{H}}(\mathbb{X}, \mathbb{Y}) = \inf \{\alpha \geq 0 \mid \mathbb{X} \subset \mathbb{Y}^\alpha \text{ and } \mathbb{Y} \subset \mathbb{X}^\alpha\}$$

Example :

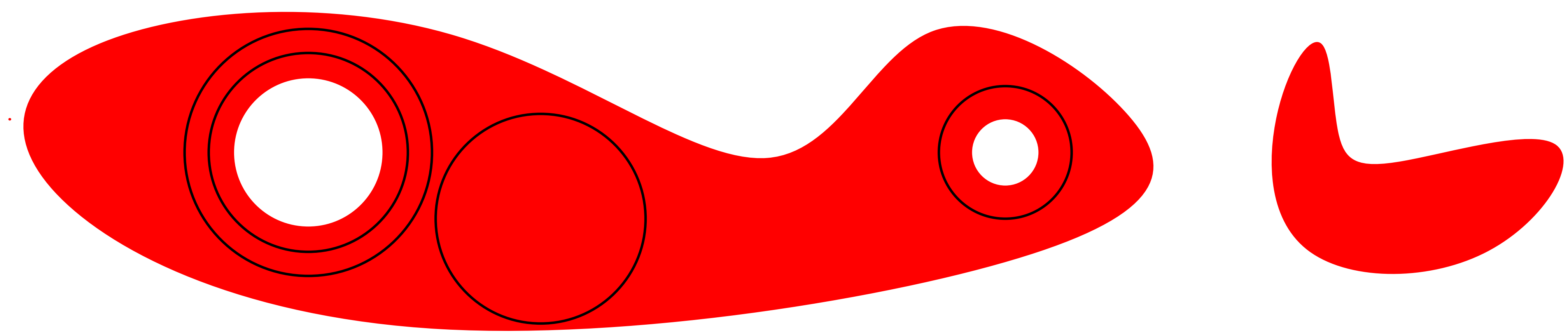
Theorem [Chazal and Lieutier, 2007]: Let $\mathbb{X}$ and $\mathbb{Y}$ be two compact sets in $\mathbb{R}^d$ and let $\varepsilon > 0$ be such that $\mathrm{d}_{\mathsf{H}}(\mathbb{X}, \mathbb{Y}) < \varepsilon$, $\mathrm{wfs}(\mathbb{X}) > 2\varepsilon$ and $\mathrm{wfs}(\mathbb{Y}) > 2\varepsilon$. Then for any $0 < \alpha < 2\varepsilon$, $\mathbb{X}^\alpha$ and $\mathbb{Y}^\beta$ are homotopy equivalent.

Sampling conditions in Hausdorff metric.

Statistical analysis of homotopy inference can be deduced from support estimation of a distribution under the Hausdorff metric.

Homology inference

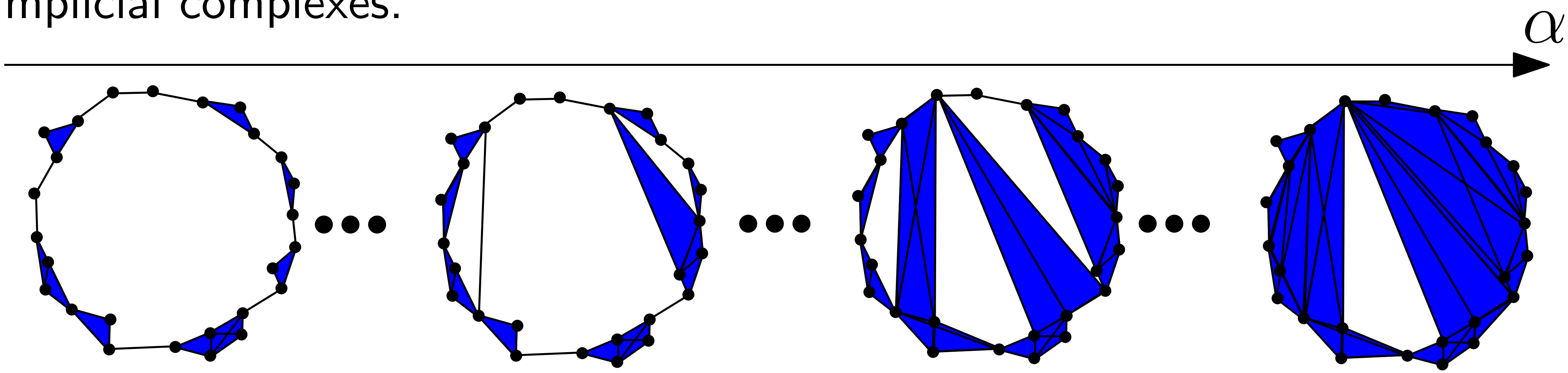
- **Homotopy** is not easy to compute in practice.
- **Singular homology** provides a algebraic description of "holes" in a geometric shape (connected components, loops, etc ...)
- **Betti number** β_k is the rank of the k-th homology group.
- **Computational Topology** : Betti numbers can be computed on simplicial complexes.



Homology inference [Niyogi et al., 2008 and 2011] [Balakrishnan et al., 2012] : The Betti number (actually the homotopy type) of Riemannian manifolds with positive reach can be recovered with high probability from offsets of a sample on (or close to) the manifold.

Persistent homology

Starting from a point cloud $\mathbb{X}_n$, let $\mathrm{Filt} = (\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$ be a fitration of nested simplicial complexes.

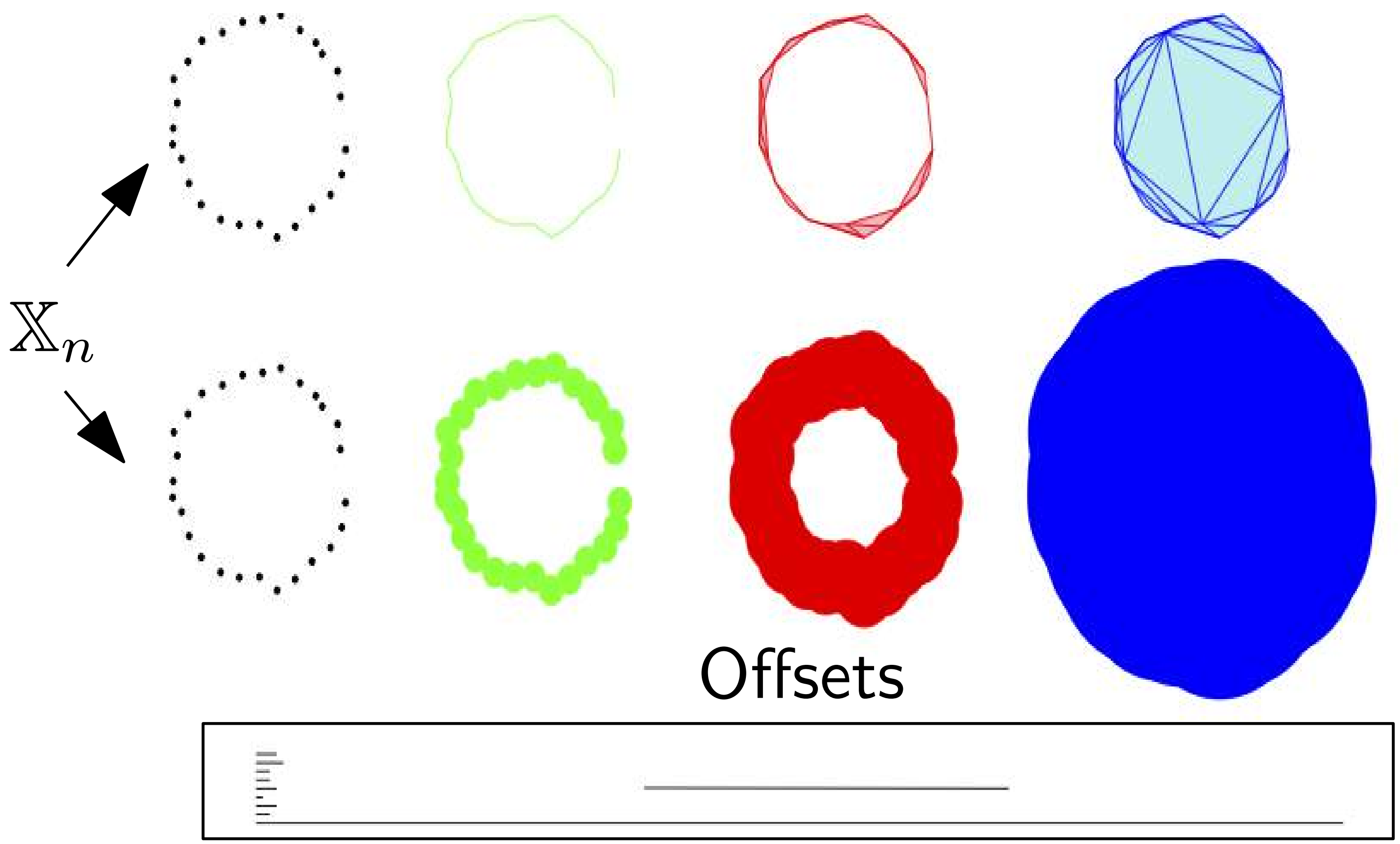


Persistent homology: identification of “persistent” topological features along the filtration.

- multiscale information ;
- more stable and more robust ;

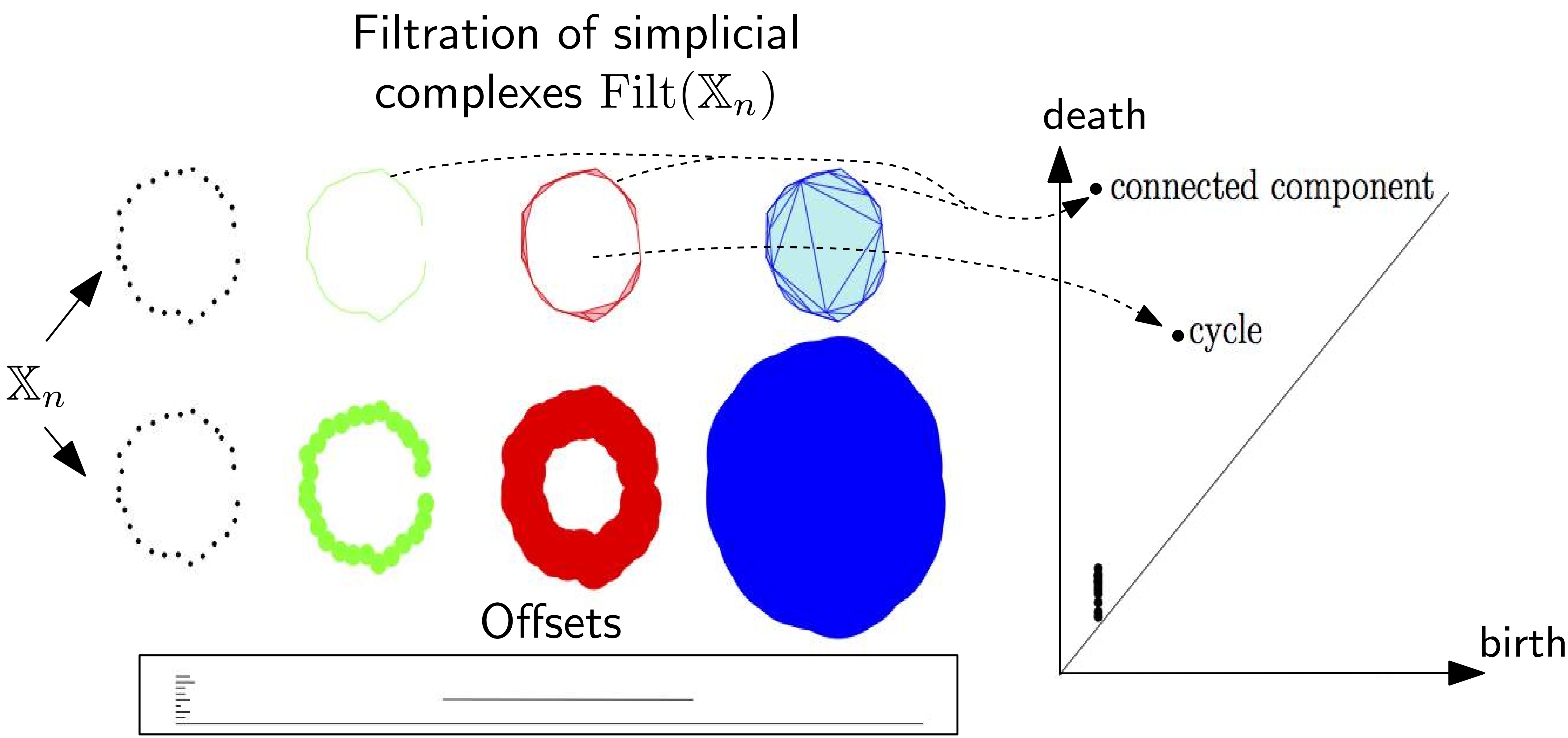
Barecodes and Persistence Diagrams

Filtration of simplicial complexes $\mathrm{Filt}(\mathbb{X}_n)$



Barcode

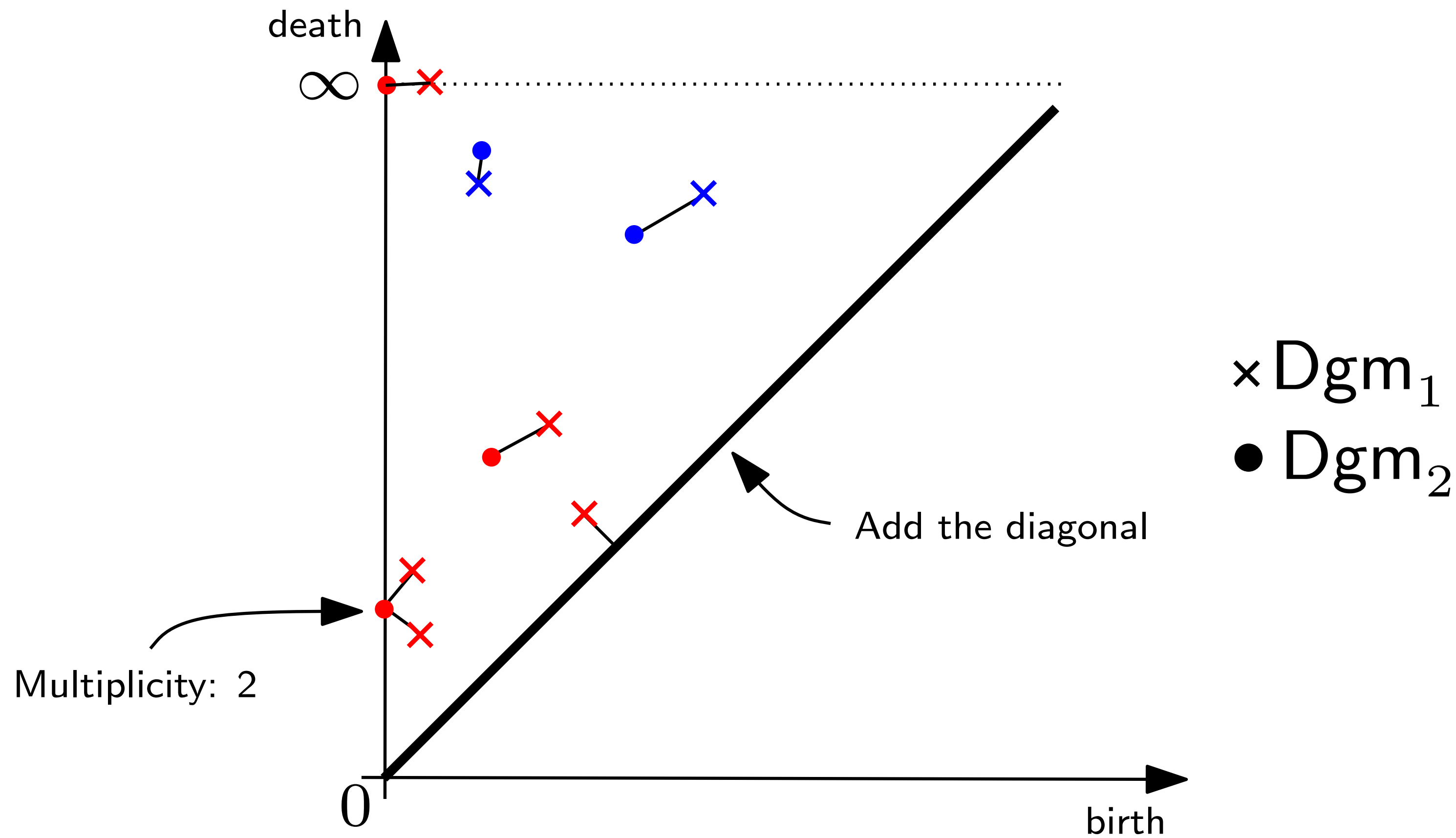
Barecodes and Persistence Diagrams



Barcode

$\mathrm{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)$
Persistence diagram of the filtration $\mathrm{Filt}(\mathbb{X}_n)$ built on $\mathbb{X}_n$.

Distance between persistence diagrams and stability



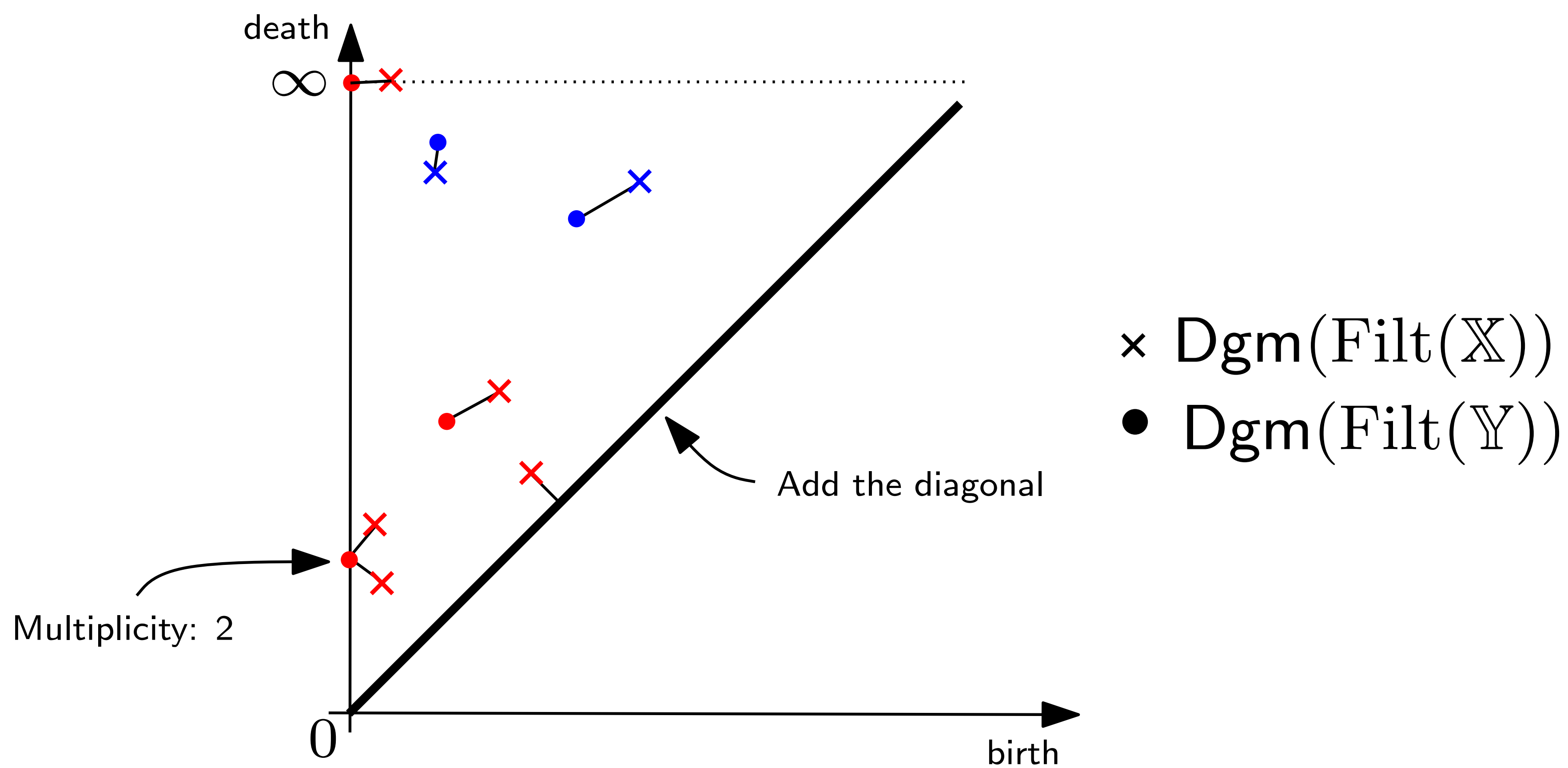
The bottleneck distance between two diagrams Dgm_1 and Dgm_2 is

$$\mathrm{d_b}(\mathsf{Dgm}_1, \mathsf{Dgm}_2) = \inf_{\gamma \in \Gamma} \sup_{p \in \mathsf{Dgm}_1} \|p - \gamma(p)\|_\infty$$

where Γ is the set of all the bijections between Dgm_1 and Dgm_2 and

$$\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|).$$

Distance between persistence diagrams and stability



Theorem [Chazal et al., 2012]: For any compact metric spaces $(\mathbb{X}, \rho)$ and $(\mathbb{Y}, \rho')$,

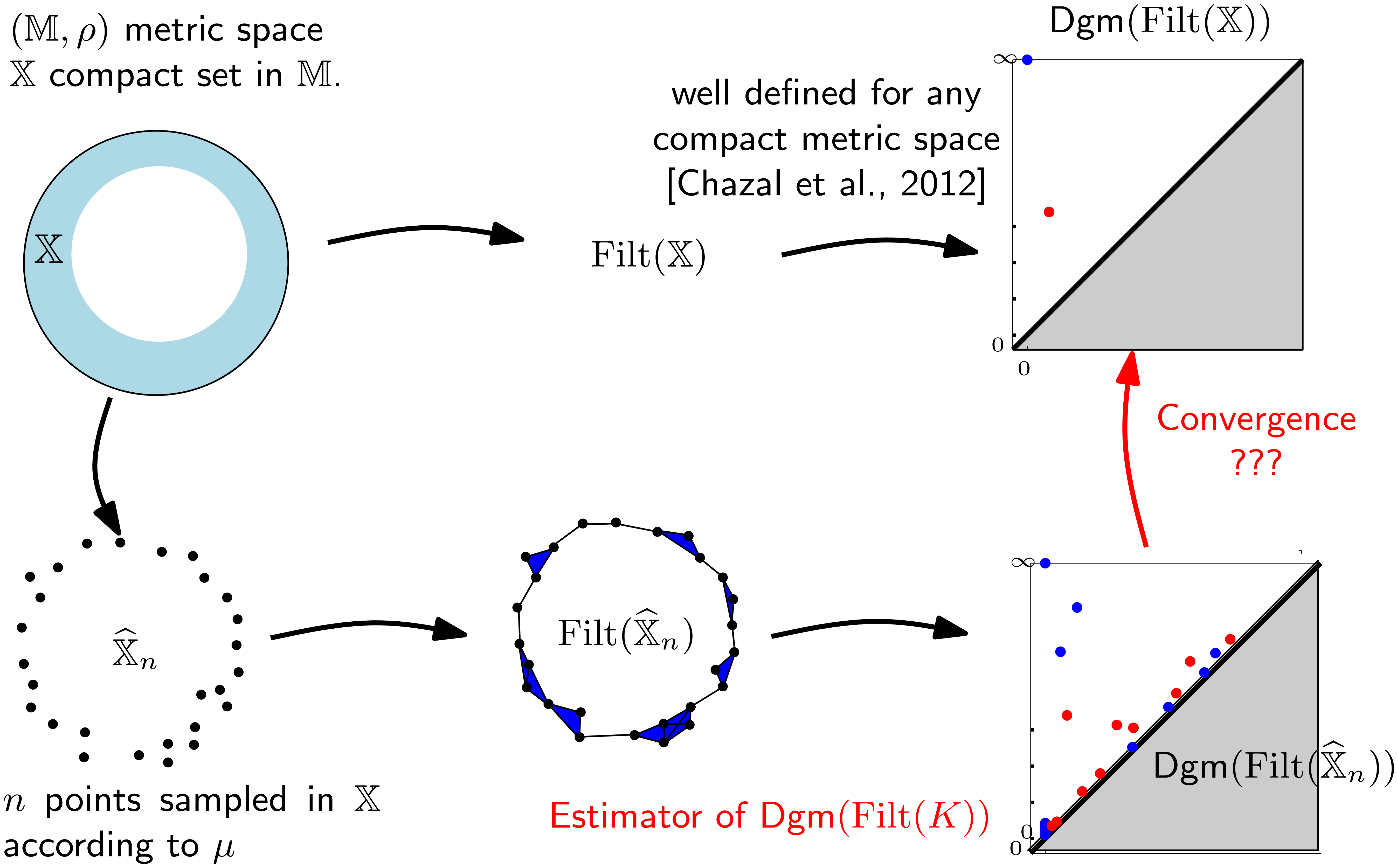
$$\mathrm{d_b}\left(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathsf{Dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leq 2\, \mathrm{d}_{\mathsf{GH}}\left(\mathbb{X}, \mathbb{Y}\right).$$

Consequently, if $\mathbb{X}$ and $\mathbb{Y}$ are embedded in the same metric space $(\mathbb{M}, \rho)$ then

$$\mathrm{d_b}\left(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X})), \mathsf{Dgm}(\mathrm{Filt}(\mathbb{Y}))\right) \leq 2\, \mathrm{d}_{\mathsf{H}}\left(\mathbb{X}, \mathbb{Y}\right).$$

Statistics
and
Persistent homology

Persistence diagram inference [Chazal 2015 JMLR]



Persistence diagram inference

For $a, b > 0$, μ satisfies the (a, b)-standard assumption on its support $\mathbb{X}_\mu$ if for any $x \in X_\mu$ and any $r > 0$:

$$\mu(B(x,r)) \geq \min(ar^b, 1).$$

$\mathcal{P}(a, b, \mathbb{M})$: set of all the probability measures satisfying the (a, b)-standard assumption on the metric space $(\mathbb{M}, \rho)$.

Theorem: For $a, b > 0$:

$$\sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{Dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_n)))\right] \leq C \left(\frac{\ln n}{n}\right)^{1/b}$$

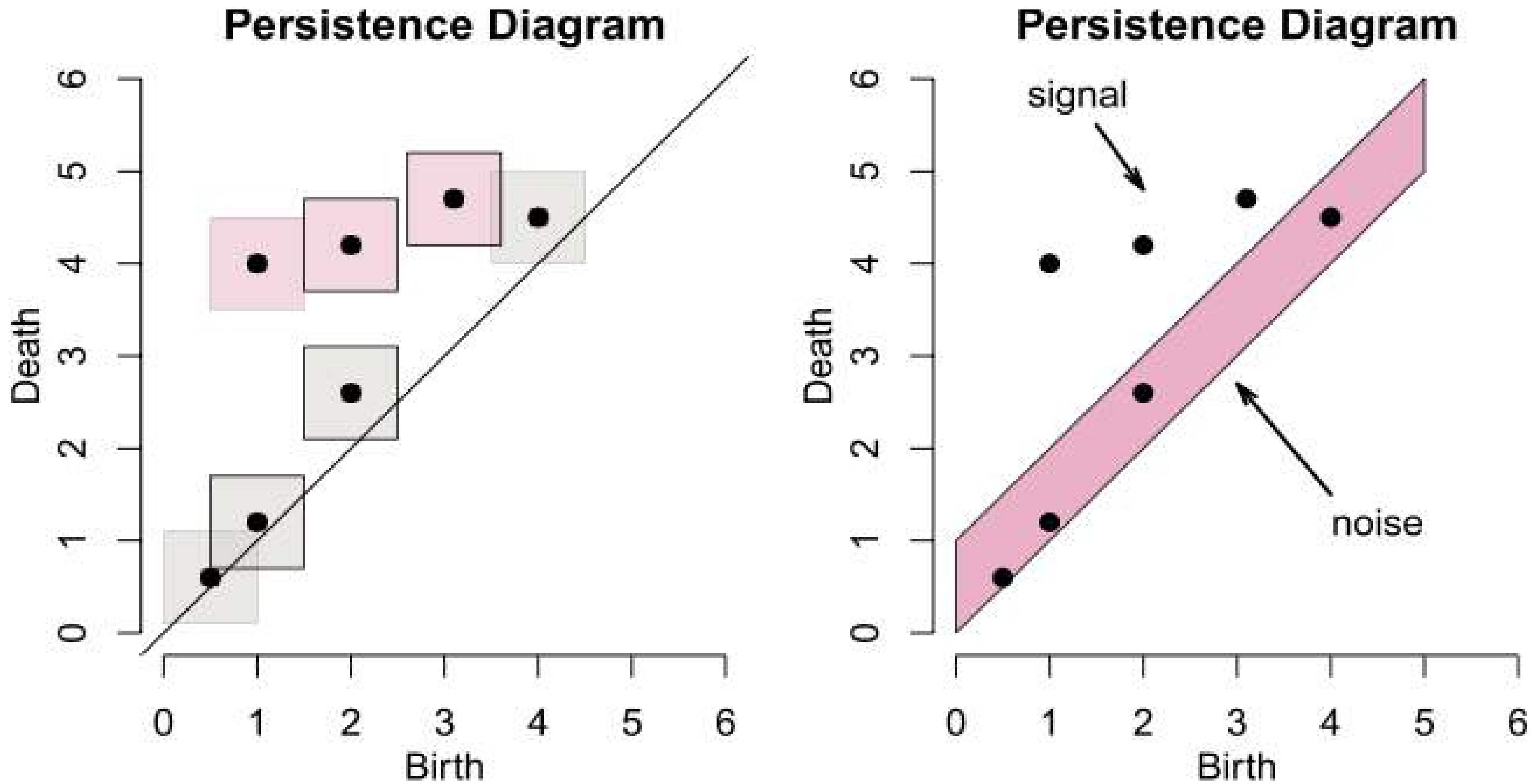
where C only depends on a and b.

Under additional technical hypotheses, for any estimator $\widehat{\mathsf{Dgm}}_n$ of $\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$:

$$\liminf_{n \to \infty} \sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{Dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathsf{Dgm}}_n)\right] \geq C' n^{-1/b}$$

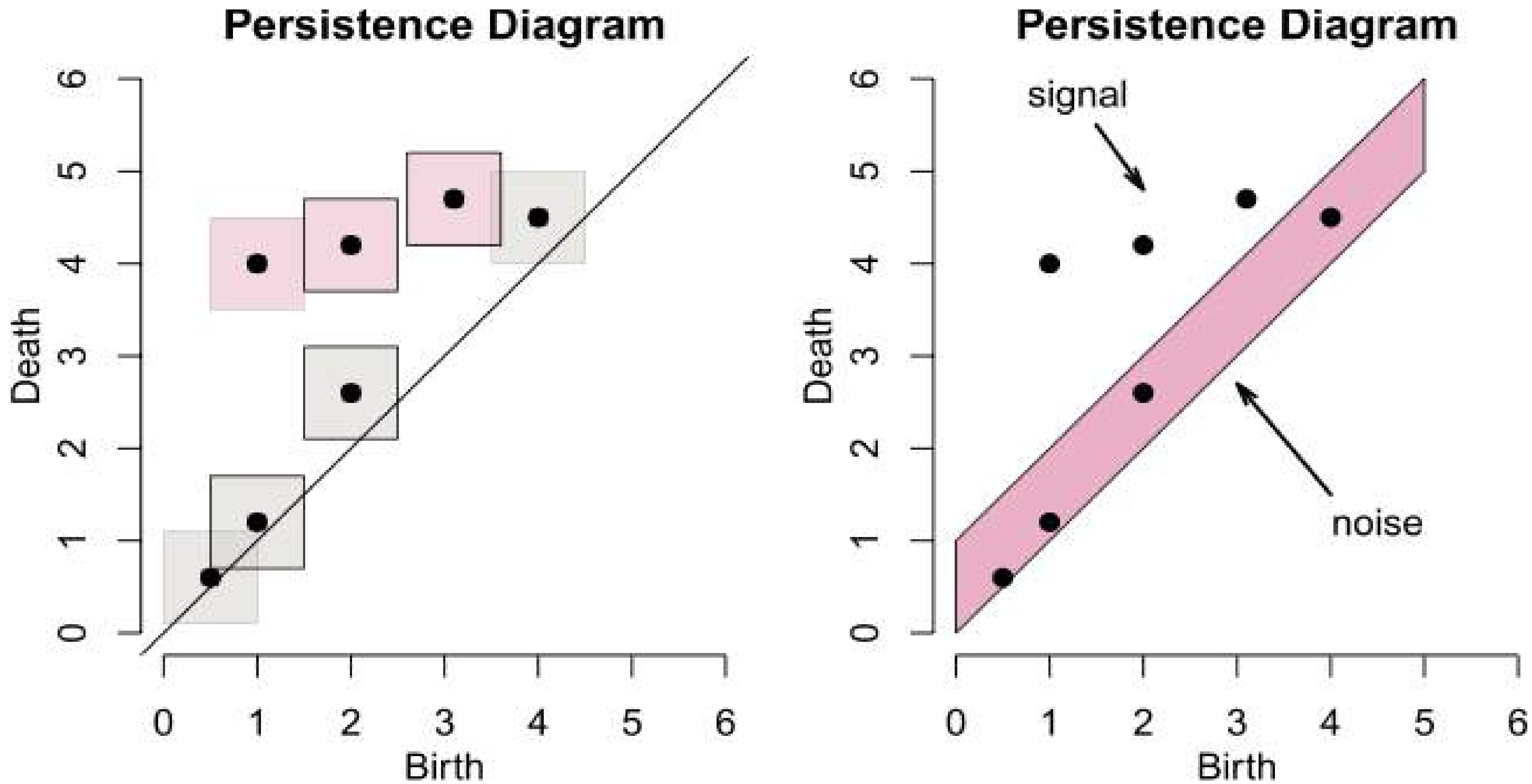
where C' is an absolute constant.

Confidence sets for persistence diagrams [Fasy 2014 AoS]



$$P\left(\mathsf{Dgm}(\mathrm{Filt}(K)) \in \hat{\mathcal{R}}\right) \geq 1-\alpha \qquad ??$$

Confidence sets for persistence diagrams [Fasy 2014 AoS]



$$P\left(\mathsf{Dgm}(\mathrm{Filt}(K)) \in \hat{\mathcal{R}}\right) \geq 1-\alpha \qquad ??$$

Using the Hausdorff stability, we can define confidence sets for persistence diagrams:

$$\mathrm{d_b}\left(\mathsf{Dgm}\left(\mathrm{Filt}(K)\right), \mathsf{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)\right) \leq \mathrm{d_H}(K, \mathbb{X}_n).$$

It is sufficient to find c_n such that

$$\limsup_{n\to\infty}\left(\mathrm{d_H}(K, \mathbb{X}_n) > c_n\right) \leq \alpha.$$

Confidence sets for persistence diagrams [Fasy 2014 AoS]

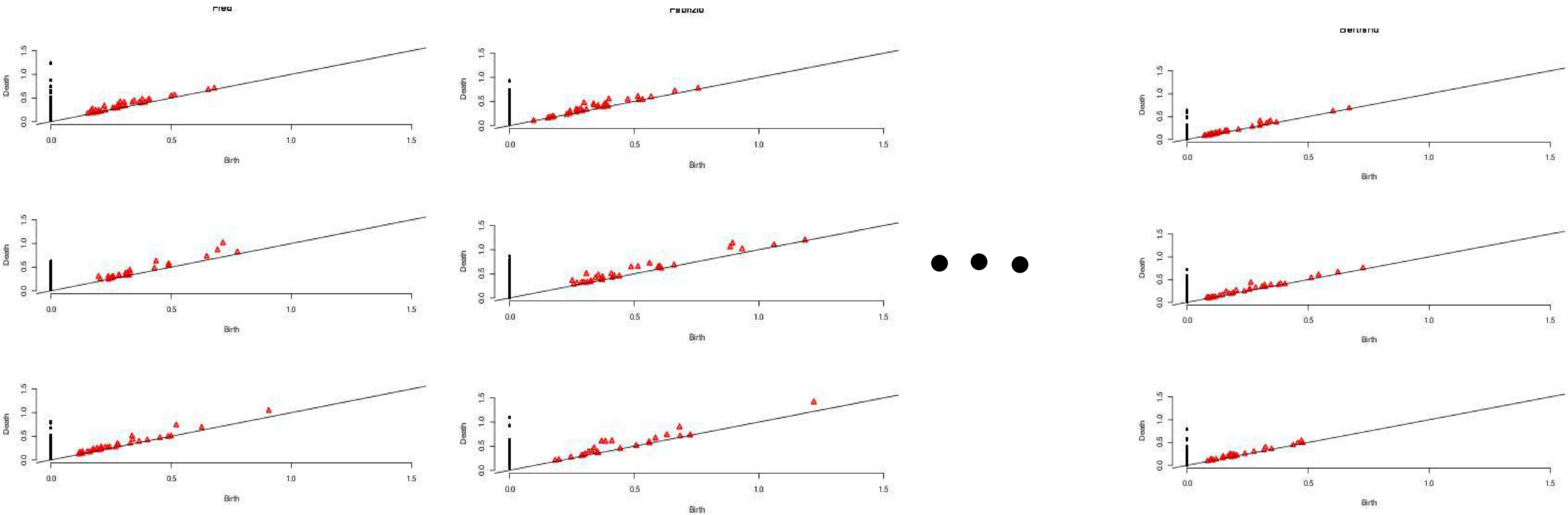
Subsampling method:

- N subsamples $\mathbb{X}^1_{b,n}, \ldots, \mathbb{X}^N_{b,n}$ of size b.
- Compute $T_j = \mathrm{d_H}\left(\mathbb{X}^j_{b,n}, \mathbb{X}_n\right)$, $j = 1, \ldots, N$.
- Compute $L_b(t) = \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{T_j > t}$,
- Take $c_b = 2L_b^{-1}(\alpha)$.

If P satisfies an (a,b) standard assumption then, for n large enough :

$$\begin{aligned} P\Big(W_\infty\left(\mathrm{Dgm}\left(\mathrm{Filt}(\mathbb{X}_\mu)\right), \mathrm{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)\right) > c_b\Big) &\leq P\Big(\mathrm{d_H}\left(\mathbb{X}_\mu, \mathbb{X}_n\right) > c_b\Big) \\ &\leq \alpha + O\left(\frac{b}{n}\right)^{1/4} \end{aligned}$$

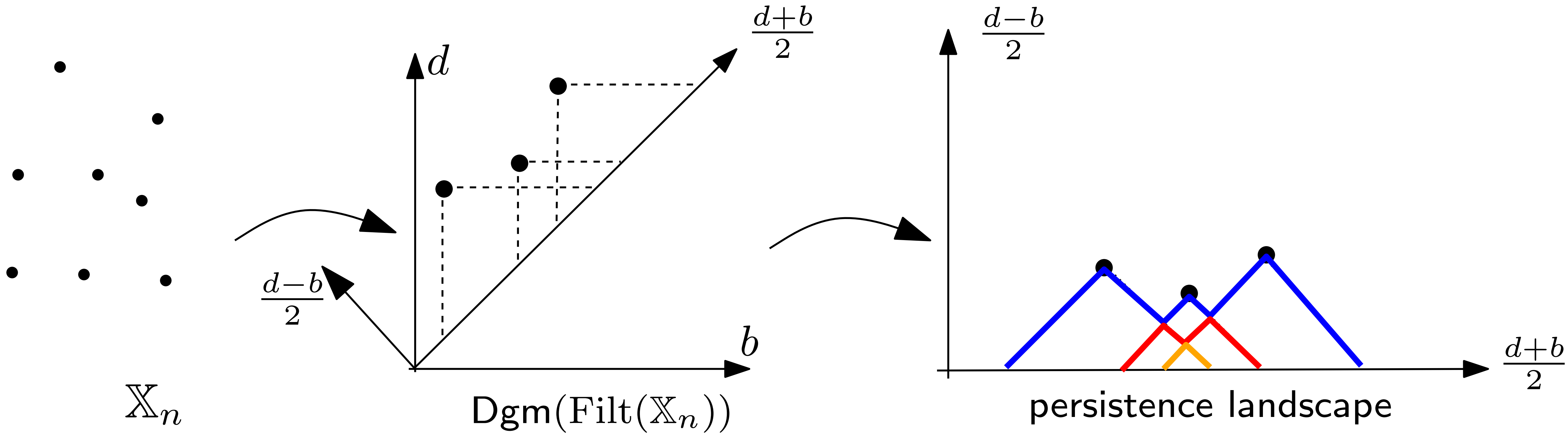
Central tendency for persistent homology



- Frechet mean [Turner 2014]

- Use an alternative descriptor of persistence : Persistence landscapes [Bubenik, 2015]

Persistence landscapes [Bubnik JMLR 2015]



$$\mathrm{Dgm} = \left\{ \left(\tfrac{d_i+b_i}{2}, \tfrac{d_i+b_i}{2}\right),\ i \in I \right\}$$

For $p = \left(\frac{b+d}{2}, \frac{d-b}{2}\right) \in$ Dgm,

$$\Lambda_p(t) \;=\; \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases}$$

Persistence landscape λ of Dgm:

$$\lambda(k,t) = \underset{p \in D}{\mathrm{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R},\ k \in \mathbb{N},$$

where kmax is k-th largest value in the set.

Stability: For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda(k,t) - \lambda'(k,t)| \le \mathrm{d_b}(\mathrm{Dgm}, \mathrm{Dgm}')$.

Subsampling methods for pers. homology [Chazal ICML 2015]

joint work with F. Chazal, B. Fasy, F. Lecci, A. Rinaldo and L. Wasserman

- Let $X = \{X_1, \cdots, X_m\}$ sampled from μ.
- λ_X: corresponding persistence landscape.
- Ψ_μ^m: the measure induced by $\mu^{\otimes m}$ on the space of persistence landscapes.
- We consider the point-wise expectations of the (random) persistence landscape under this measure:

$$\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)], t \in [0, T]$$

- For $S_1^m, \ldots, S_\ell^m$ some independent samples of size m from $\mu^{\otimes m}$, the empirical counterpart of $\mathbb{E}_{\Psi_\mu^m}[\lambda_X(t)]$ is

$$\overline{\lambda_\ell^m}(t) = \frac{1}{\ell} \sum_{i=1}^{\ell} \lambda_{S_i^m}(t), \qquad \text{for all } t \in [0, T],$$

Subsampling methods for pers. homology [Chazal ICML 2015]

Definition: The p-th Wasserstein distance between two measures μ, ν defined on $(\mathbb{M}, \rho)$ is

$$W_{\rho,p}(\mu, \nu) = \left(\inf_{\Pi} \int_{\mathbb{M}\times\mathbb{M}} [\rho(x, y)]^p d\Pi(x, y) \right)^{\frac{1}{p}},$$

where the infimum is taken over all measures on $\mathbb{M} \times \mathbb{M}$ with marginals μ and ν.

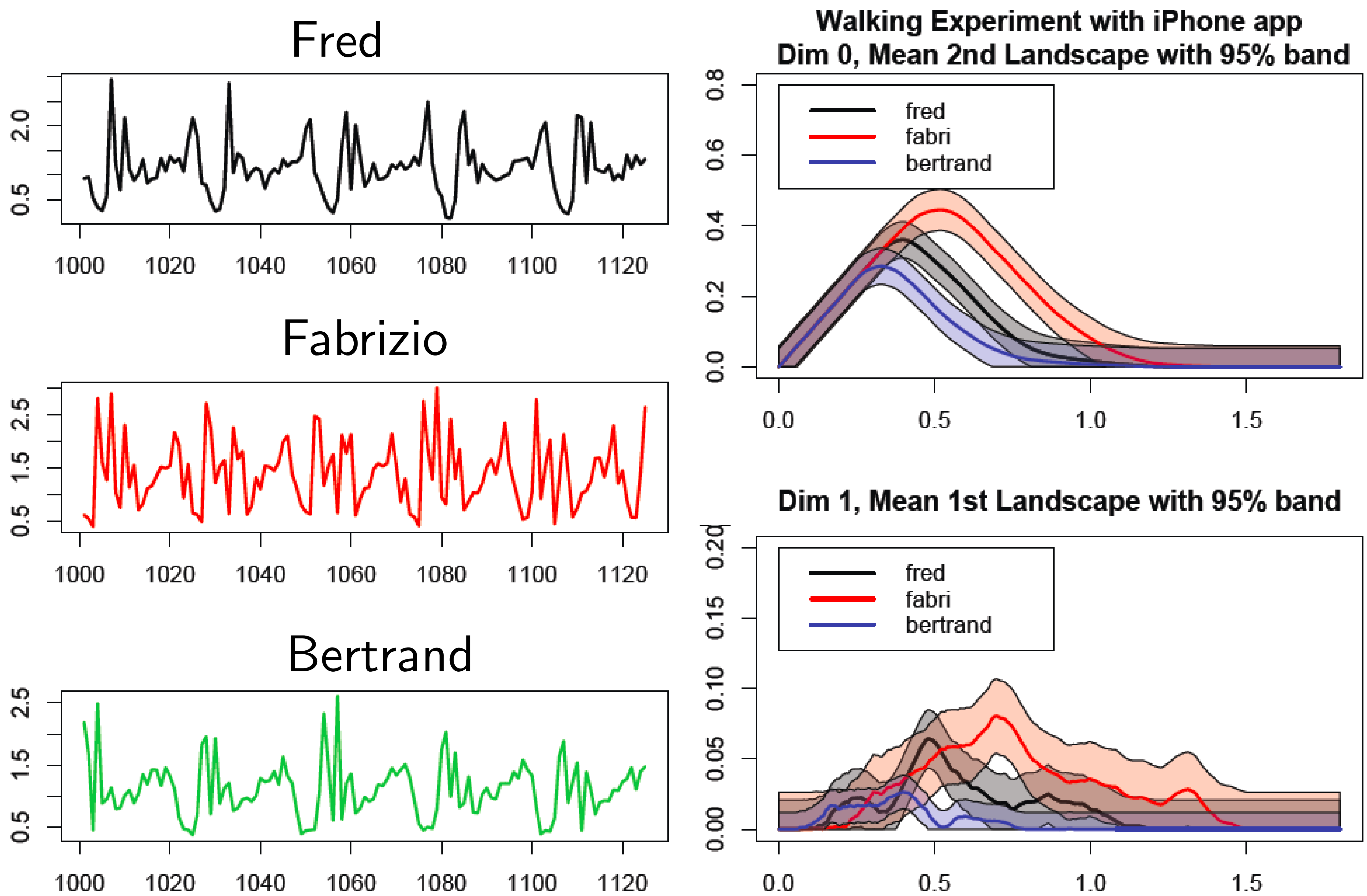
Stability of the average landscape:

Theorem: Let $X \sim \mu^{\otimes m}$ and $Y \sim \nu^{\otimes m}$, where μ and ν are two probability measures on $\mathbb{M}$. For any $p \geq 1$ we have

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2\, m^{\frac{1}{p}} W_{\rho,p}(\mu, \nu).$$

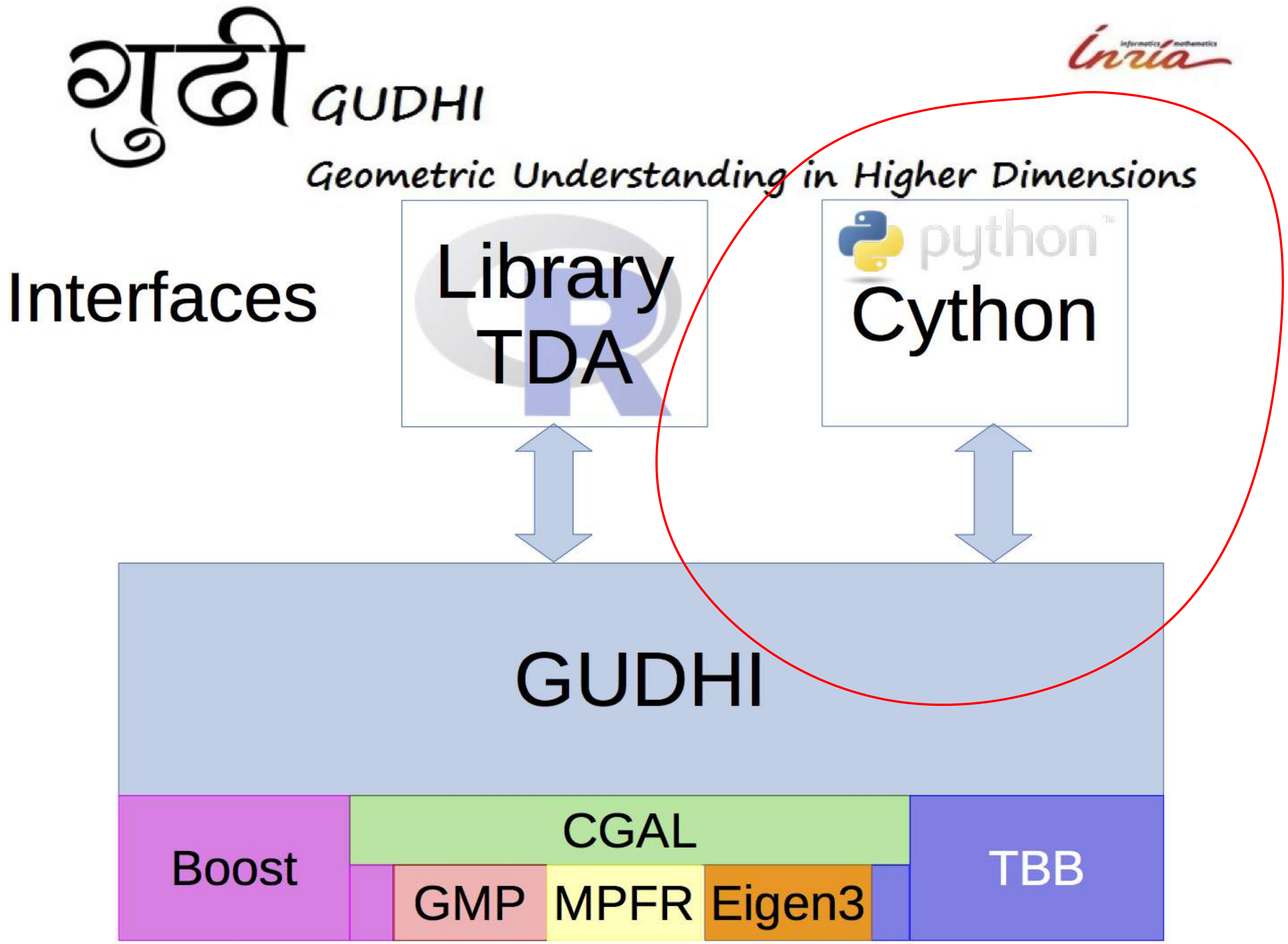
Subsampling methods for pers. homology [Chazal ICML 2015]

Application: Analysis of accelerometer data.



- topological features carry discriminative information
- no registration/calibration preprocessing step needed

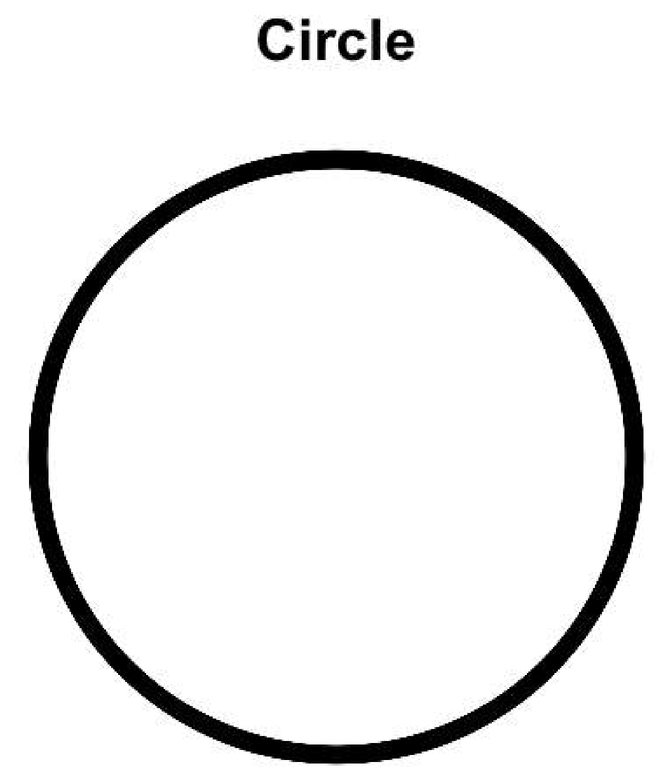
Commercial break: Gudhi with Statistical learning Python Libraries

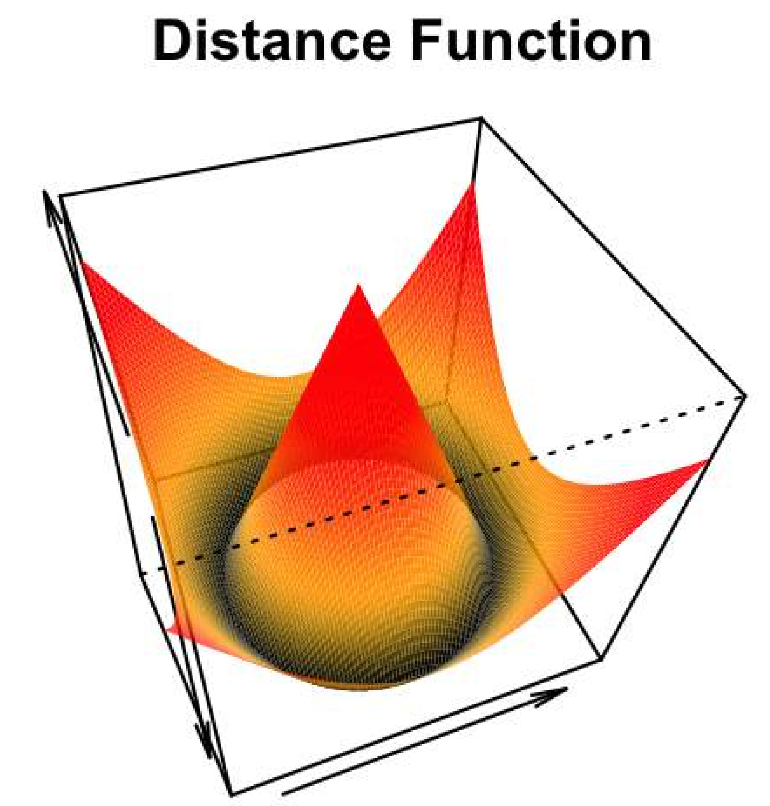


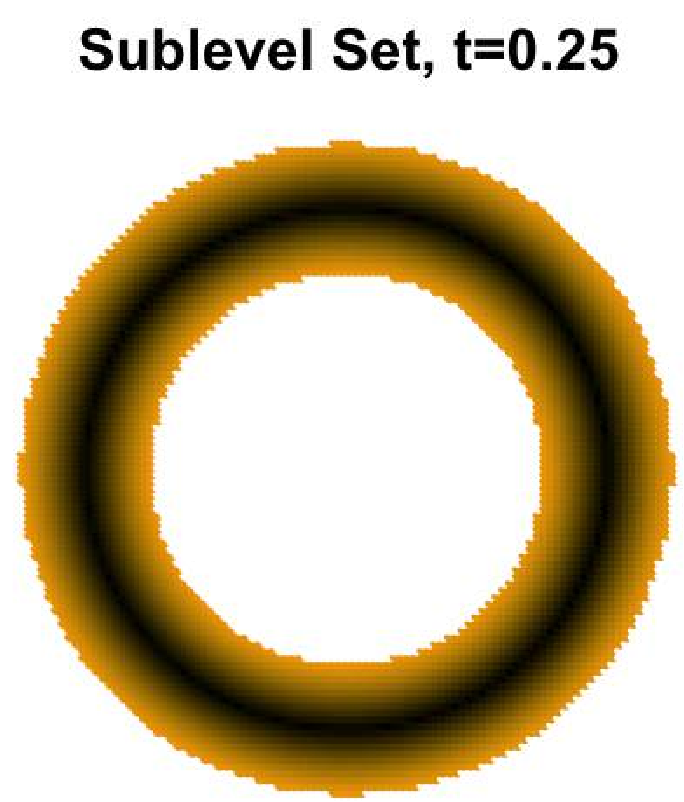
and coming soon : `Gudhi Stat` with more tools for statistics and TDA.

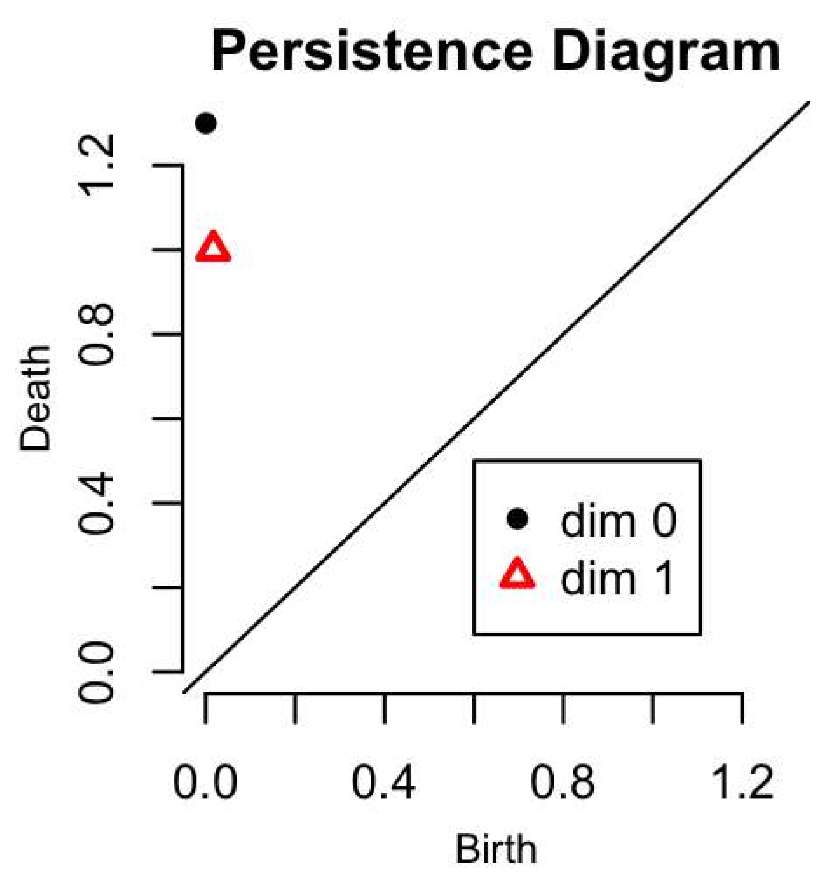
Robust TDA

Standard TDA methods are not robust to outliers







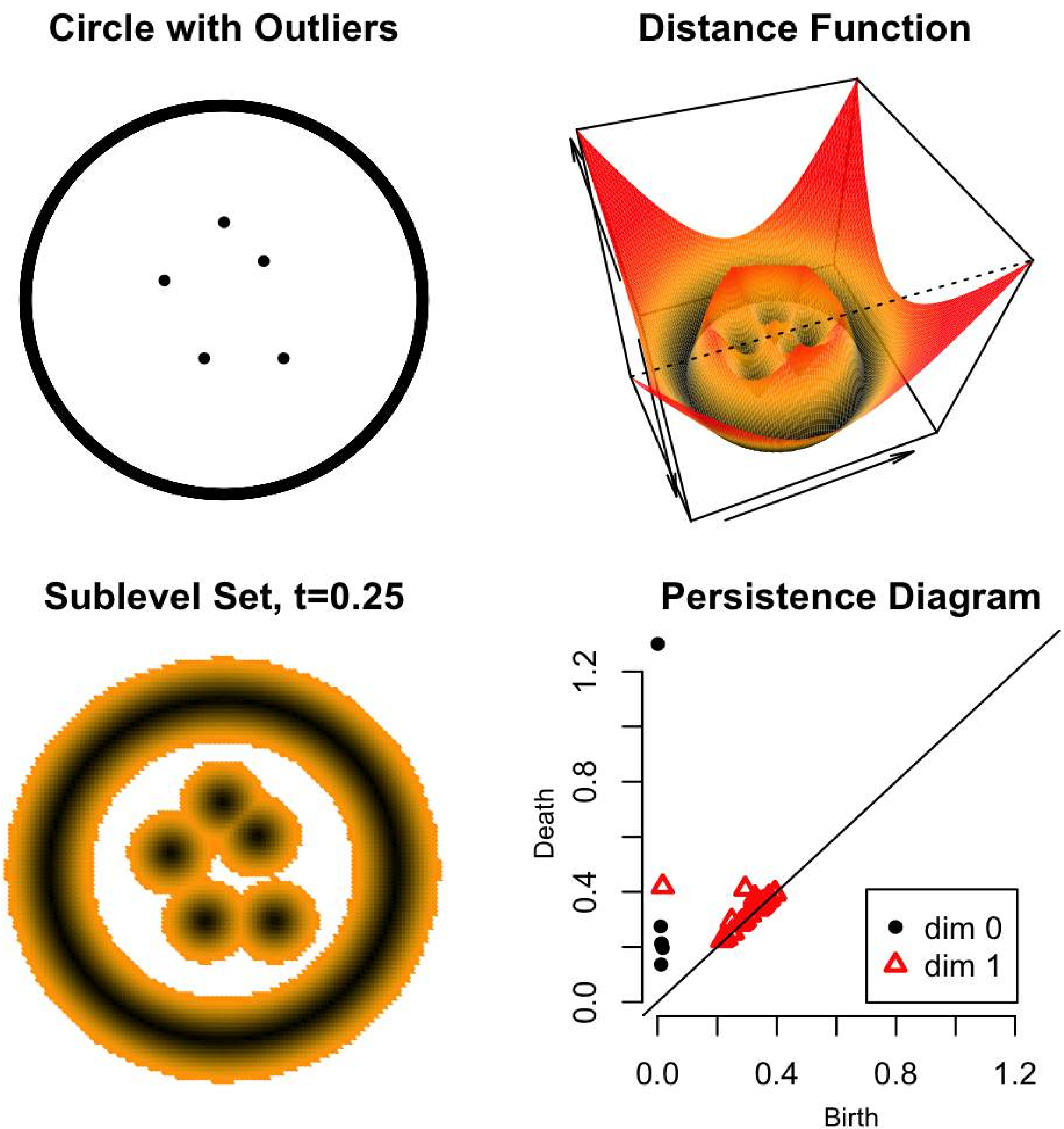


$$\mathbb{X}^r := \bigcup_{x \in \mathbb{X}} B(x, r) = d_{\mathbb{X}}^{-1}([0, r])$$

where the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is

$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\|$$

Standard TDA methods are not robust to outliers



$$\mathbb{X}^r := \bigcup_{x\in\mathbb{X}} B(x,r) = d_{\mathbb{X}}^{-1}([0,r])$$

where the distance function $d_{\mathbb{X}}$ to $\mathbb{X}$ is

$$d_{\mathbb{X}}(y) = \inf_{x\in\mathbb{X}} \|x-y\|$$

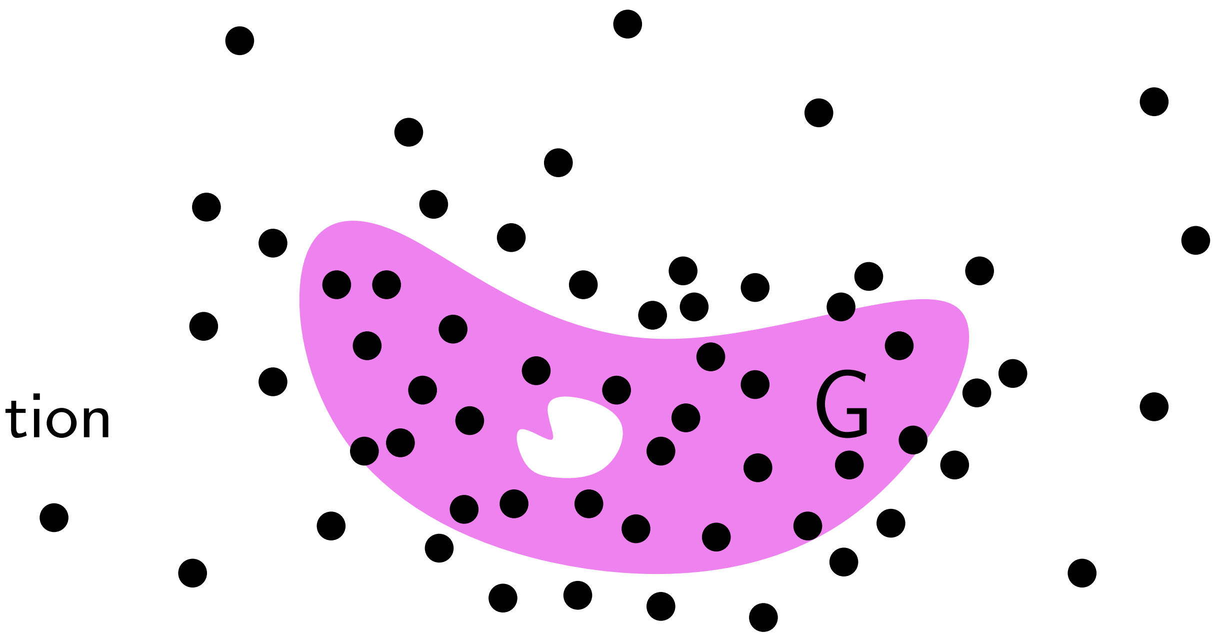
Some possible "noise models" for geometry

- Additive noise model

$$P = \mu \star \Phi$$

distribution with support the geometric shape G

noise distribution

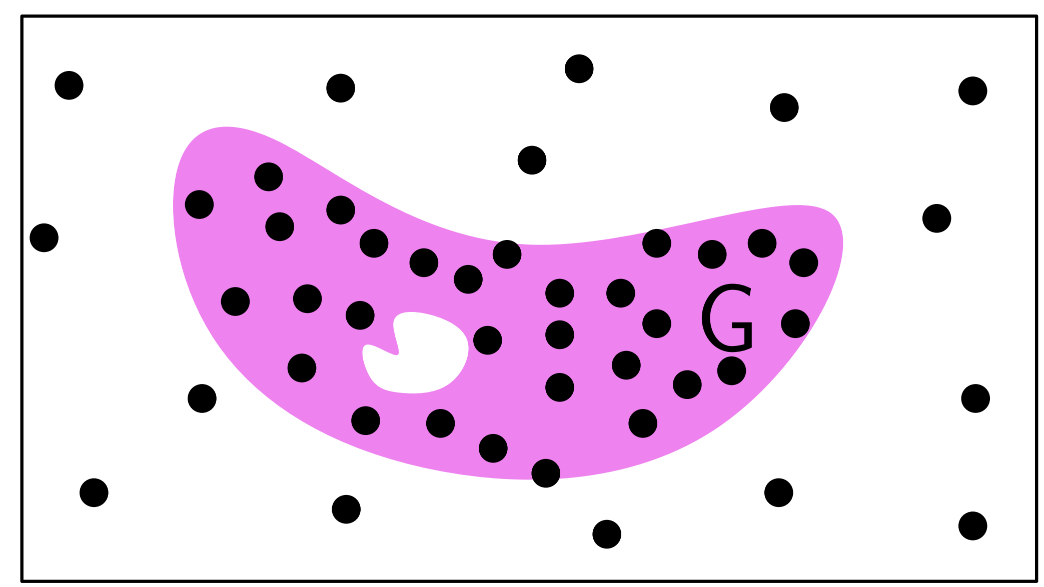


- Clutter noise model

$$P = \pi\mu + (1-\pi)U$$

distribution with support the geometric shape G

Uniform distribution on the box

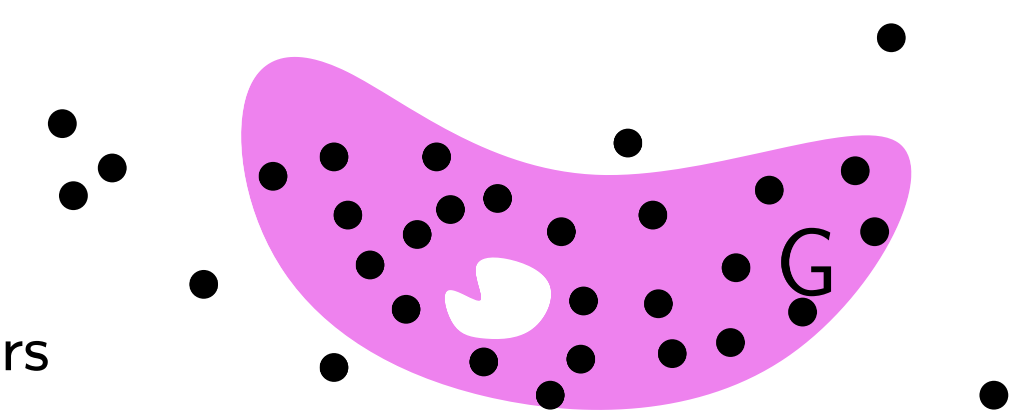


- A few outliers

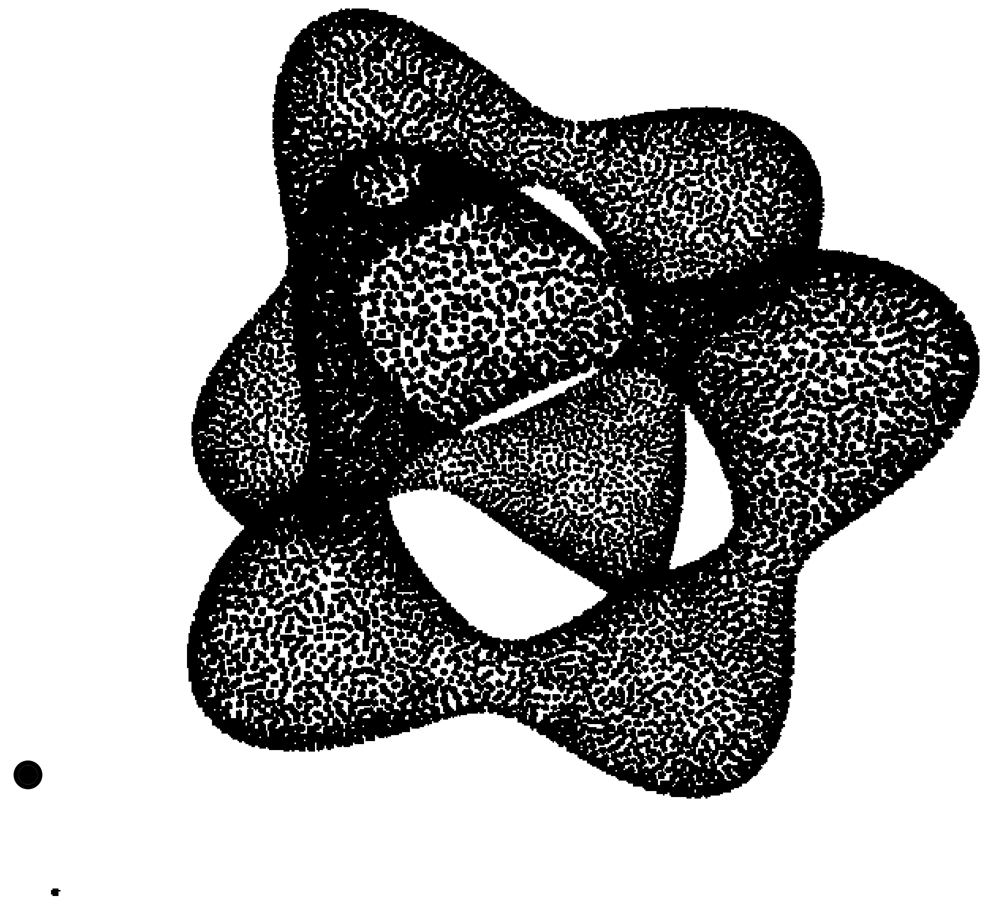
$$P = \pi\mu + (1-\pi)\Psi$$

distribution with support the geometric shape G

Distribution of outliers



Robust TDA with an alternative distance function ?

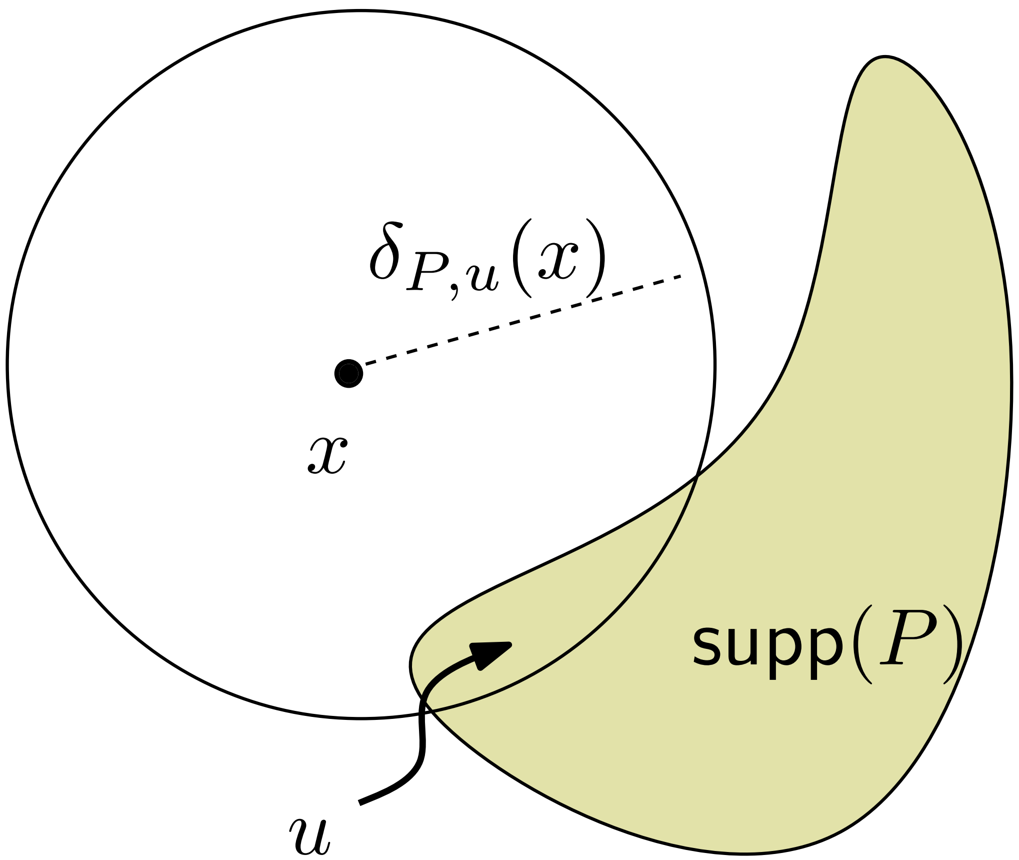


We would like to consider the sub levels of an alternative distance function related to the sampling measure, which support is $\mathbb{X}$, or close to $\mathbb{X}$.

Distance To Measure [Chazal 11 FoCM]

Preliminary distance function to a measure P:
Let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf \{ r > 0 : P(B(x,r)) \geq u \}$$



$\delta_{P,u}$ is the smallest distance needed to capture a mass of at least u.

$\delta_{P,u}$ is the quantile function at u of the r.v.

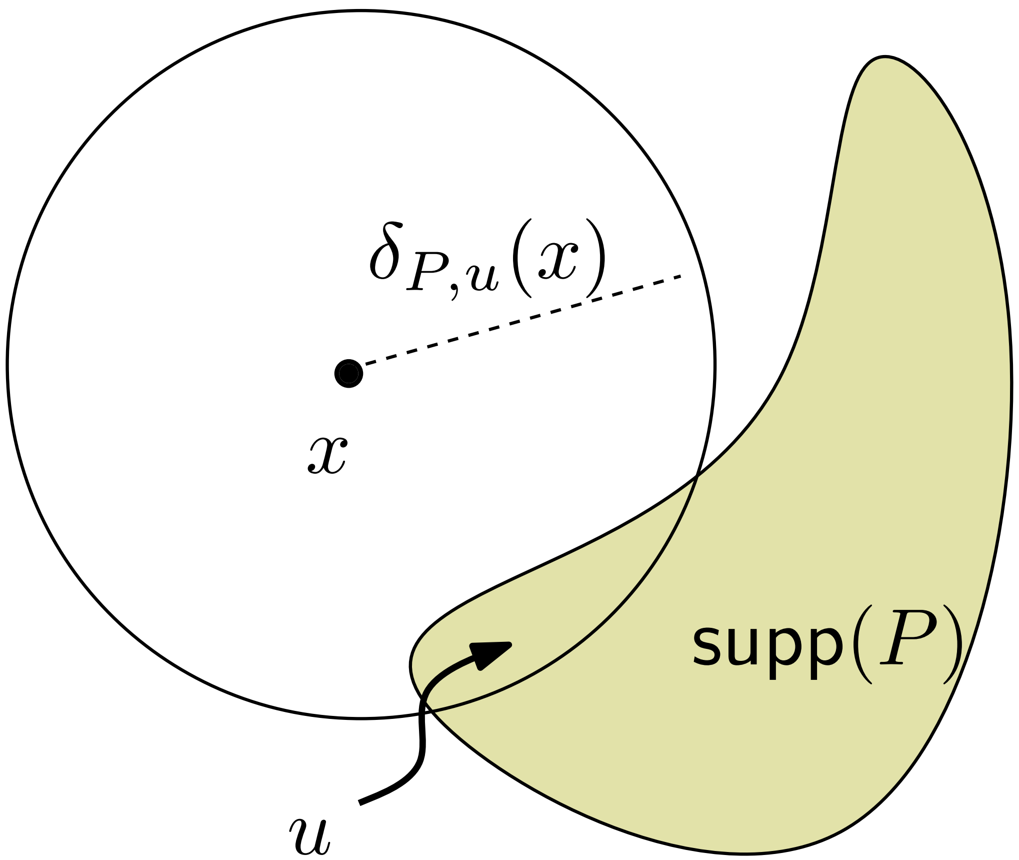
$$\|x - X\|$$

where $X \sim P$.

Distance To Measure [Chazal 11 FoCM]

Preliminary distance function to a measure P:
Let $u \in]0,1[$ be a positive mass, and P a probability measure on $\mathbb{R}^d$:

$$\delta_{P,u}(x) = \inf\{r > 0 : P(B(x,r)) \geq u\}$$



Definition: Given a probability measure P on $\mathbb{R}^d$ and $m > 0$, the distance function to the measure P (DTM) is defined by

$$\boxed{d_{P,m} : x \in \mathbb{R}^d \mapsto \left(\frac{1}{m}\int_0^m \delta_{P,u}^2(x)\,du\right)^{1/2}}$$

Distance To Measure [Chazal 11 FoCM]

Properties of the DTM :

- Stability under Wassertein perturbations:

$$\|d_{P,m} - d_{Q,m}\|_\infty \leq \frac{1}{\sqrt{m}} W_2(P, Q)$$

- The function $x \mapsto d^2_{P,m}(x)$ is semiconcave, this is ensuring strong regularity properties on the geometry of its sublevel sets.

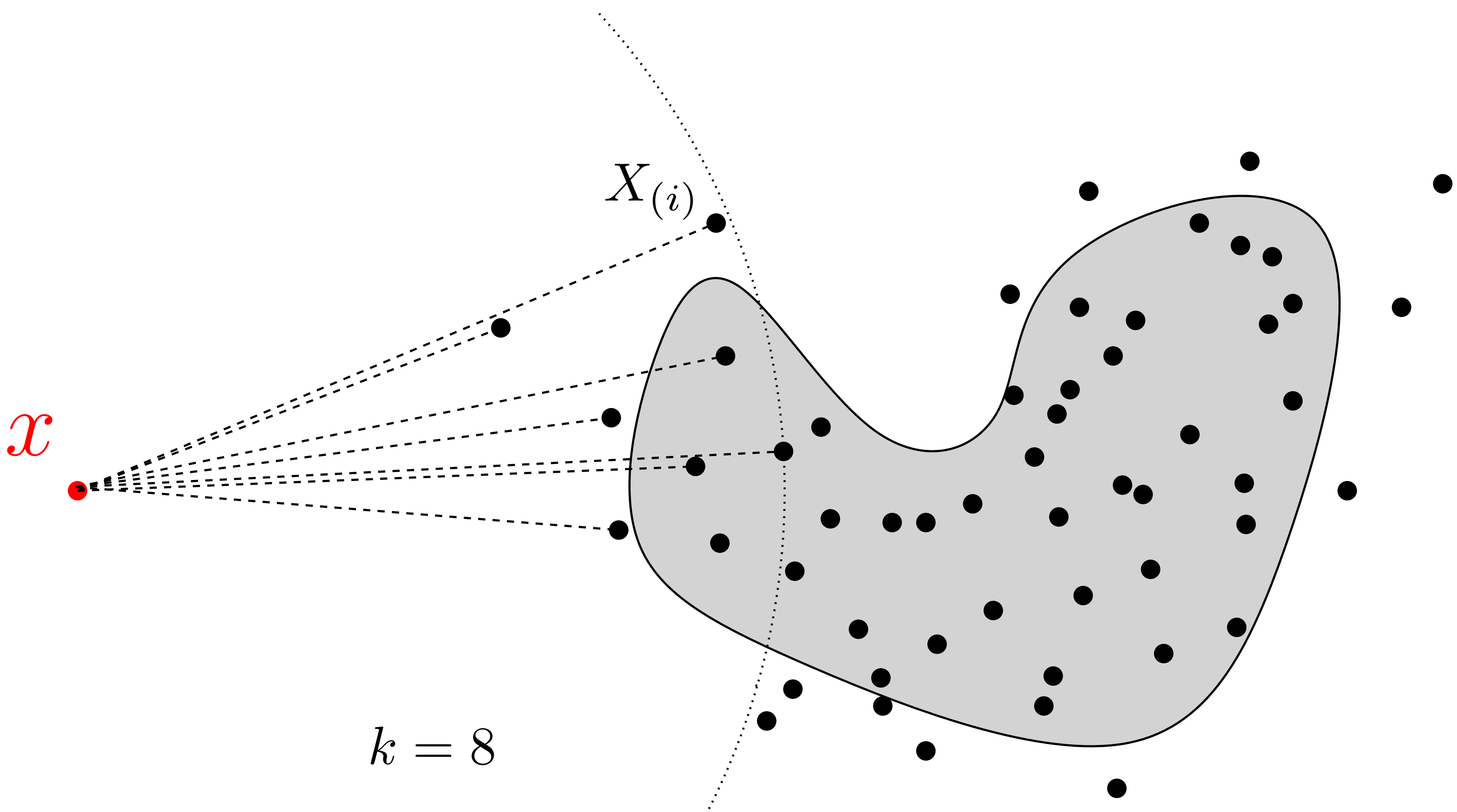
- Consequently, if $\tilde{P}$ is a probability distribution close to P for Wasserstein distance W_2, then the sublevel sets of $d_{\tilde{P},m}$ provide a topologically correct approximation of the support of P.

Distance to The Empirical Measure (DTEM)

Let $X_1, \ldots, X_n$ sample according to P and let P_n be the empirical measure. Then

$$d^2_{P_n, \frac{k}{n}}(x) = \frac{1}{k} \sum_{i=1}^{k} ||x - X_{(i)}||^2$$

where $||X_{(1)} - x|| \geq ||X_{(2)} - x|| \geq \cdots \geq ||X_{(k)} - x| \cdots \geq ||X_{(n)} - x||$



Geometric inference with the DTM

Theorem: [Chazal et al., 2011]
Let μ be a measure that has dimension at most $k > 0$ with compact support G such that $\text{reach}_\alpha(G) \geq R > 0$ for some $\alpha > 0$.
Let ν be another measure and ε be an upper bound on the uniform distance between d_G and d_{ν,m_0}. Then, for any $r \in [4\varepsilon/\alpha^2, R - 3\varepsilon]$ and any $\eta \in]0, R[$, the r-sublevel sets of d_{μ,m_0} and the η-sublevel sets of d_G are homotopy equivalent as soon as:

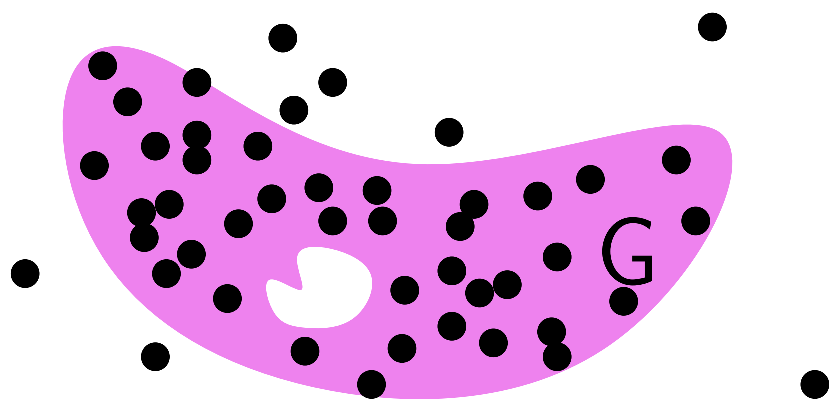
$$W_2(\mu,\nu) \leq \frac{R\sqrt{m_0}}{5 + 4/\alpha^2} - C(\mu)^{-1/k} m_0^{1/k+1/2}.$$

In practice : $X_1 \dots X_n$ sampled according to P.

Assume $W_2(P,\mu)$ small.

$P_n = \sum_{i=1}^n \delta_{X_i}$: empirical measure.

Than for n large enough, $W_2(P_n,\mu)$ is small and the sublevel sets of $d_{P_n,m}$ provide a topologically correct approximation of G.



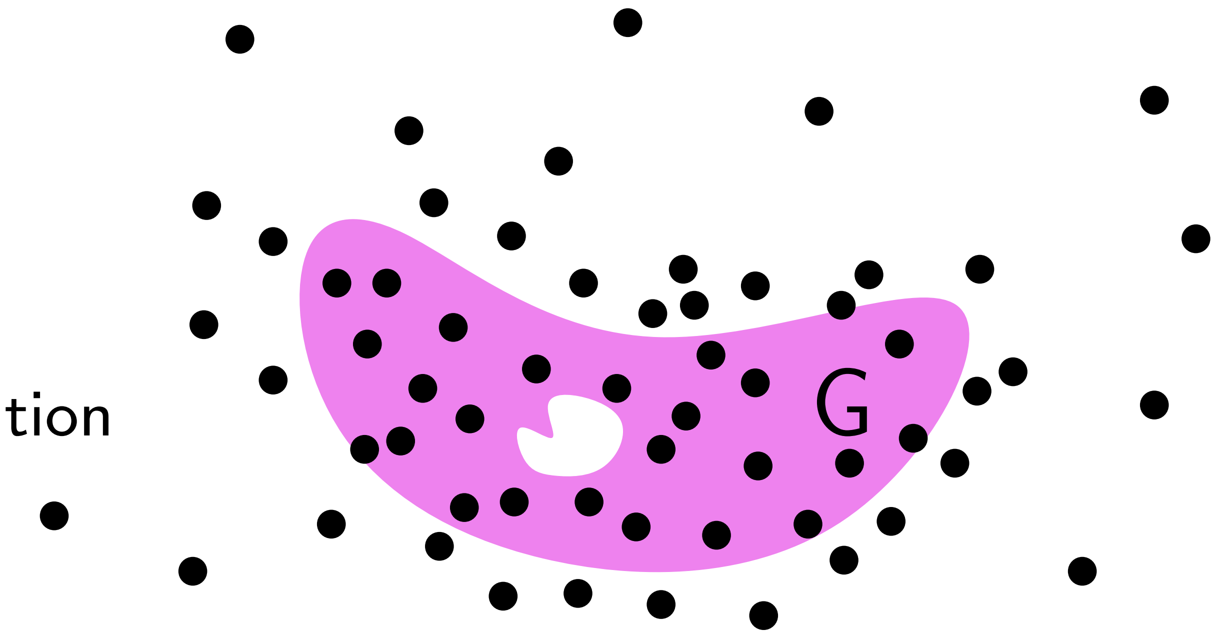
Wasserstein deconvolution and DTM denoising

Additive noise model

$$P = \mu \star \Phi$$

distribution with support the geometric shape G

noise distribution



In this case, $W_2(P, \mu)$ can be large.

Ideally we would like to denoise directly $d_{P_n,n}$, but this can be hardly achieved because the DTM is not a linear functional of the measure.

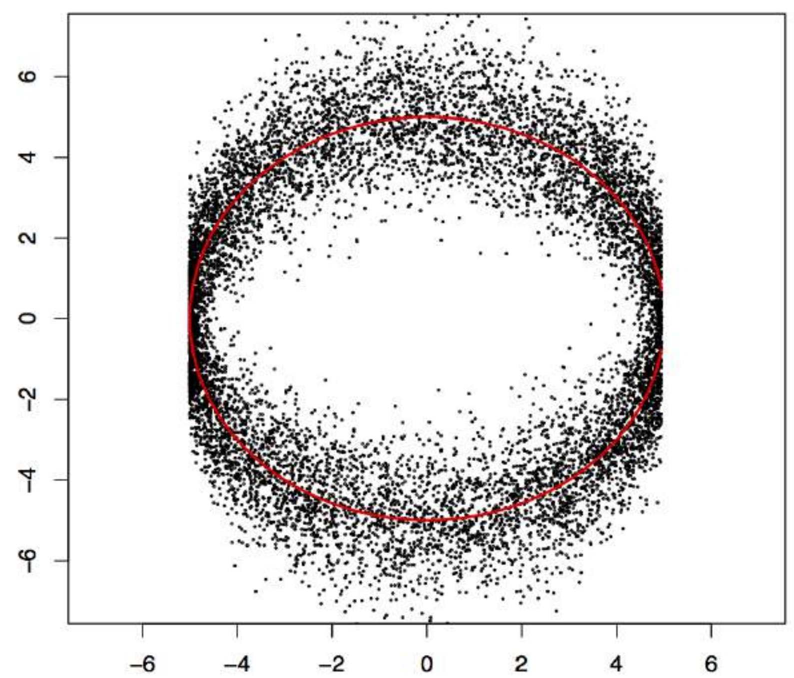
Alternative approach : deconvolve the observed measure [Caillerie EJS 2011]

$X_1, \ldots, X_n \longrightarrow$ Deconvolved measure $\tilde{\mu}_n \longrightarrow d_{\tilde{\mu}_n,m}$

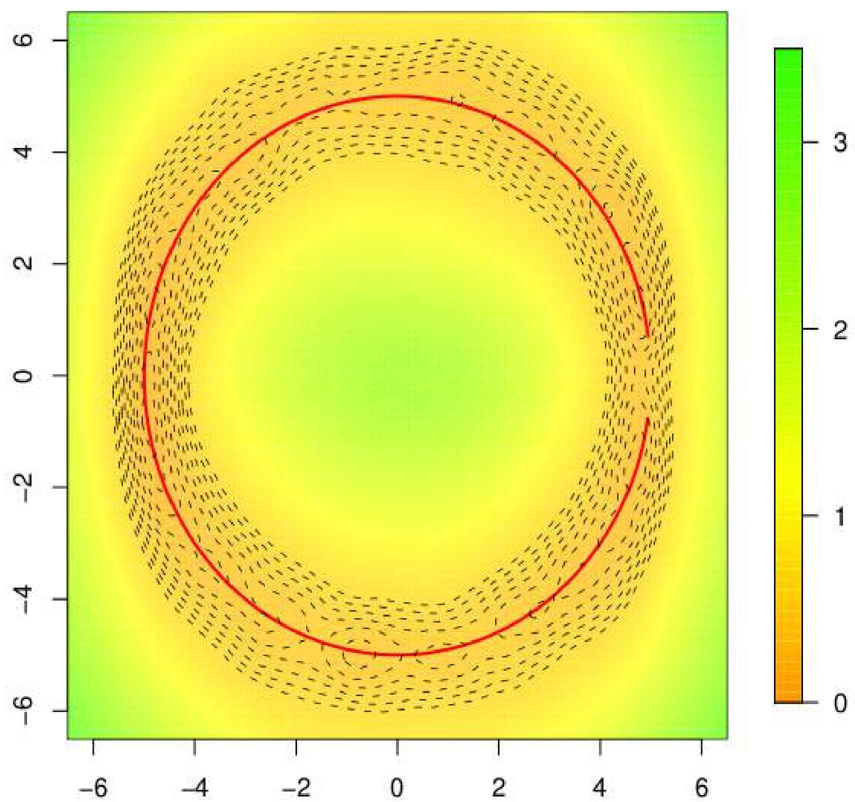
$$W_2(P_n, \mu) \geq W_2(\tilde{\mu}_n, \mu) \geq \sqrt{m}\|d_{\tilde{\mu}_n,m} - d_\mu\|_\infty$$

$\cancel{(\to 0)}$ $\quad\quad$ $(\to 0)$

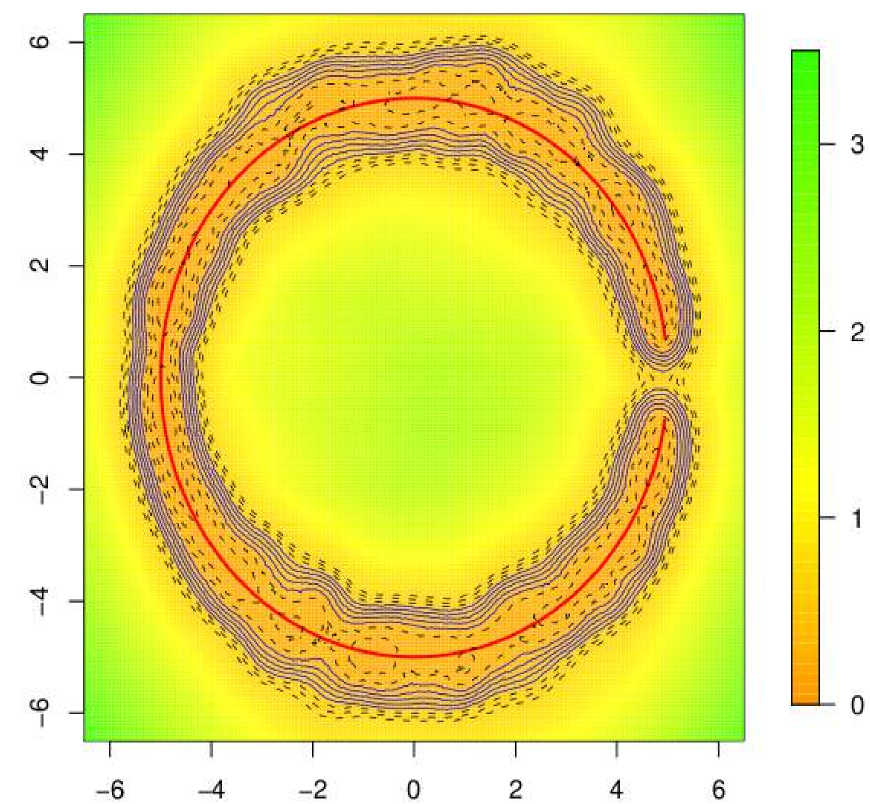
Wasserstein deconvolution and DTM denoising



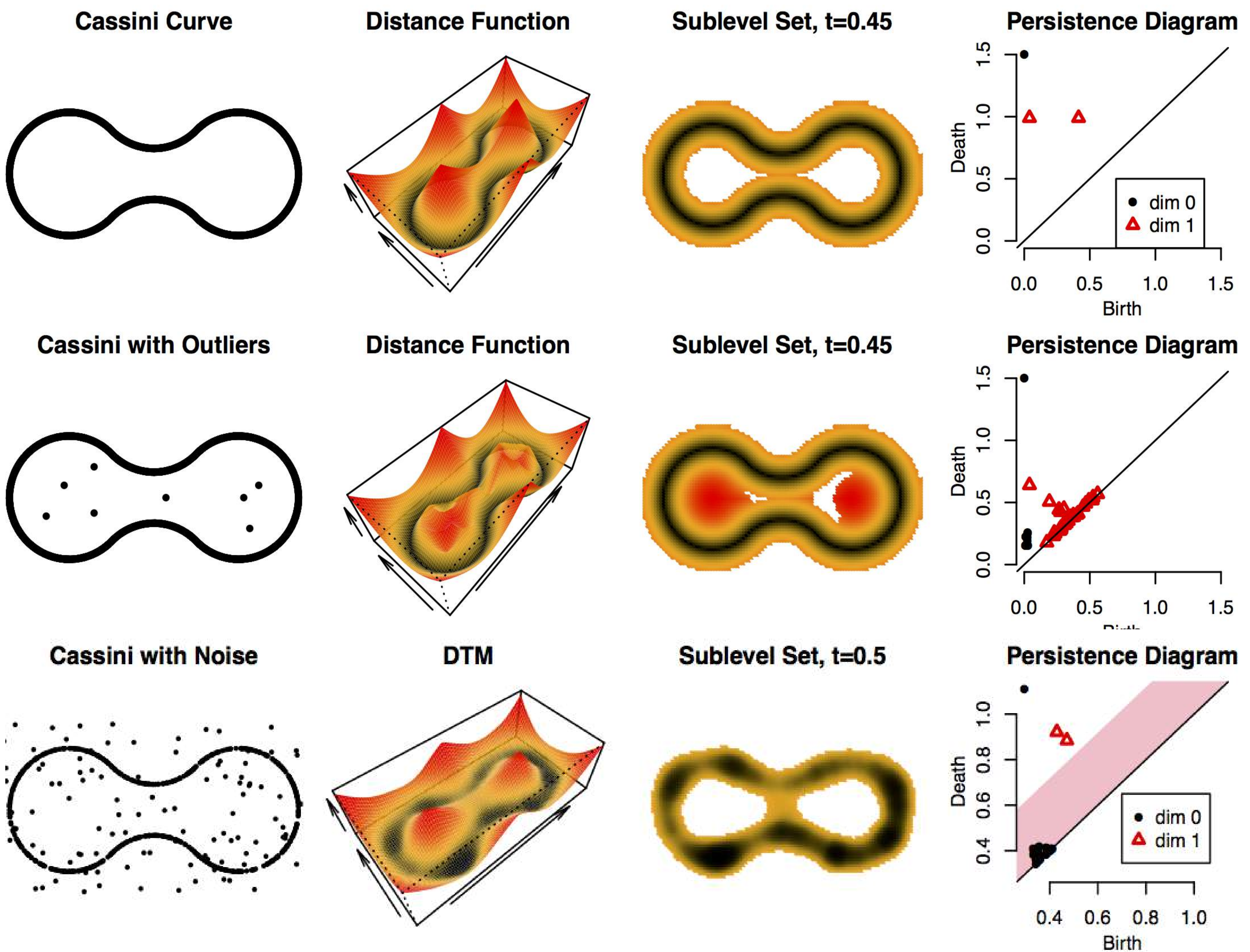
Distance to the empirical measure



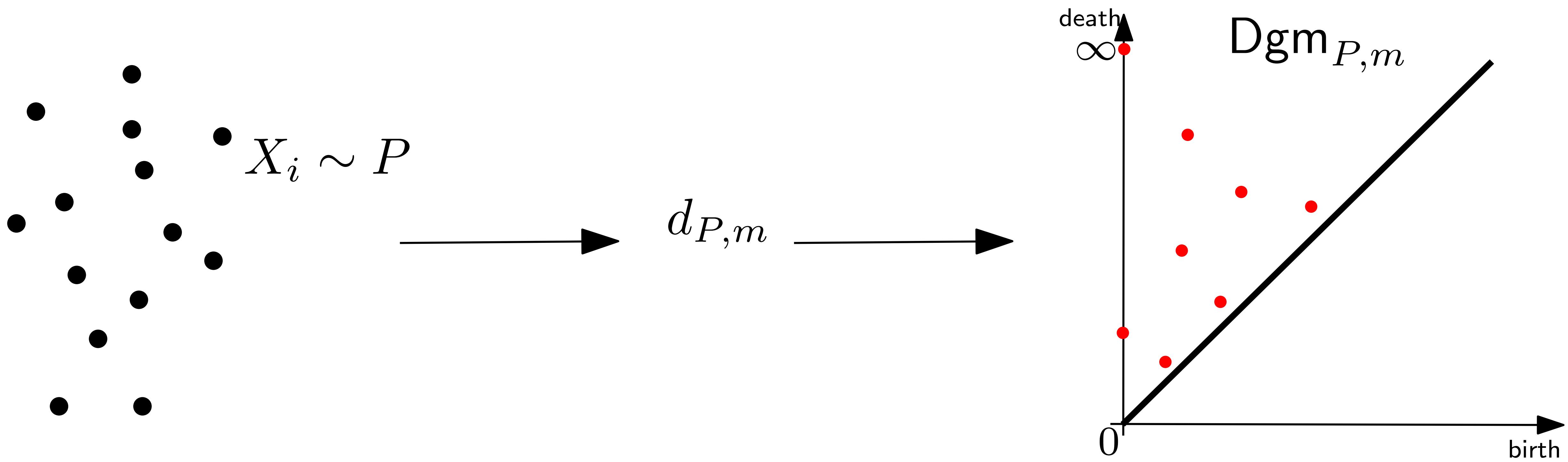
Distance to the estimator



DTM and persistent homology



DTM and persistent homology



$$\mathrm{d_b}\left(\mathrm{Dgm}_{P,m}, \mathrm{Dgm}_{Q,m}\right) \leq \|d_{P,m} - d_{Q,m}\|_\infty \leq \frac{1}{\sqrt{m}} W_2(P,Q)$$

Take $Q = P_n$ ...

Wasserstein Stability of the DTM [Chazal et al., 2012]

Stability of Persistent homology [Cohen-Steiner et al.,2005, Chazal et al., 2012]

Estimation of the DTM via the empirical DTM

[Chazal EJS 17, Chazal JMLR 17]

Quantity of interest:

$$d^2_{P_n,\frac{k}{n}}(x) - d^2_{P,\frac{k}{n}}(x)$$

- Observe that

$$d^2_{P,m}(x) = \frac{1}{m}\int_0^m F_x^{-1}(u)du$$

  where F_x is the cdf of $\|x - X\|^2$ with $X \sim P$.

- The distance to the empirical measure is the empirical counter part of the distance to P:

$$d_{P_n,m}(x)^2 = \frac{1}{m}\int_0^m F_{x,n}^{-1}(u)du$$

  where $F_{x,n}$ is the cdf of $\|x - X\|^2$ with $X \sim P_n$.

- Finally we get that

$$d^2_{P_n,\frac{k}{n}}(x) - d^2_{P,\frac{k}{n}}(x) = \frac{1}{m}\int_0^m \left\{F_{x,n}^{-1}(u) - F_x^{-1}(u)\right\} du$$

Estimation of the DTM via the empirical DTM

[Chazal EJS 17, Chazal JMLR 17]

Quantity of interest:

$$d^2_{P_n,\frac{k}{n}}(x) - d^2_{P,\frac{k}{n}}(x)$$

Two complementary approaches of the problem:

- Asymptotic approach : $\frac{k_n}{n} = m$ is fixed and n tends to infinity.

- Non asymptotic approach : n is fixed, and we want a tight control over the fluctuations of the empirical DTM, in function of k, which can be taken very small.

We **do not use Wasserstein stability** for either of the two approaches. Wasserstein rates of convergence [Fournier and Guillin, 2013 ;Dereich et al., 2013] do not provide tight rates for the DTM in this context.

Bootstrap and significance of topological features

Aim : studying the persistent homology of the sub-levels of the DTM and providing confidence regions.

Two alternative boostrap methods :

- by bootstrapping the DTM
- Bottleneck Bootstrap

Bootstrap and significance of topological features

Bootstrapping the DTM

$$P$$

$$P_n$$

Bootstrap and significance of topological features

Bootstrapping the DTM

$$P \qquad\qquad P_n$$

$$P_n \qquad\qquad P_n^*$$

Bootstrap and significance of topological features

Bootstrapping the DTM

$$
\begin{tikzcd}
P \arrow[r, leftrightarrow] & P_n \\
P_n \arrow[r, leftrightarrow] & P_n^*
\end{tikzcd}
$$

Bootstrapping the DTM

$$\begin{tikzcd} P \arrow[r, leftrightarrow] & P_n \\ P_n \arrow[r, leftrightarrow] & P_n^* \end{tikzcd}$$

$$\Phi(P) \longleftrightarrow \Phi(P_n)$$

$$\Phi(P_n) \longleftrightarrow \Phi(P_n^*)$$

Bootstrap and significance of topological features

Bootstrapping the DTM

For $m \in (0,1)$, define c_α by

$$\mathbb{P}\left(\sqrt{n}||d^2_{P,m} - d^2_{P_n,m}||_\infty > c_\alpha\right) = \alpha.$$

Let $X_1^*, \ldots, X_n^*$ be a sample from P_n, and let P_n^* be the corresponding (bootstrap) empirical measure.

We consider the bootstrap quantity $d_{P_n^*,m}(x)$ of $d_{P_n,m}$.

The bootstrap estimate $\hat{c}_\alpha$ is defined by

$$\mathbb{P}\left(\sqrt{n}||d^2_{P_n,m} - d^2_{P_n^*,m}||_\infty > \hat{c}_\alpha \,|\, X_1, \ldots, X_n\right) = \alpha$$

where $\hat{c}_\alpha$ can be approximated by Monte Carlo.

Theorem: If F_x^{-1} is regular enough, the DTM is Hadamard differentiable at P. Consequently, the bootstrap method for the DTM is asymptotically valid.

Bootstrap and significance of topological features

Bootstrapping the DTM

Dgm : persistence diagram of the sub-levels of $d_{P,m}$
$\widehat{\text{Dgm}}$: persistence diagram of the sub-levels of $d_{P_n,m}$.
Let

$$\mathcal{C}_n = \left\{ E \in \mathcal{D}\text{iag} : \ \mathrm{d_b}(\widehat{\text{Dgm}}, E) \le \frac{\hat{c}_\alpha}{\sqrt{n}} \right\},$$

where $\mathcal{D}$iag is the set of all the persistence diagrams.

Then,

Bootstrap estimate

$$\mathbb{P}(\text{Dgm} \in \mathcal{C}_n) = \mathbb{P}\left(\mathrm{d_b}(\text{Dgm}, \widehat{\text{Dgm}}) \le \frac{\hat{c}_\alpha}{\sqrt{n}} \right) \ge \mathbb{P}\left(\|d^2_{P,m} - d^2_{P_n,m}\|_\infty \le \frac{\hat{c}_\alpha}{\sqrt{n}} \right)$$

Bootstrap and significance of topological features

The Bottleneck Bootstrap

Dgm : persistence diagram of the sub-levels of $d_{P,m}$
$\widehat{\mathsf{Dgm}}$: persistence diagram of the sub-levels of $d_{P_n,m}$.
$\widehat{\mathsf{Dgm}}^*$: persistence diagram of the sub-levels of $d_{P_n^*,m}$.

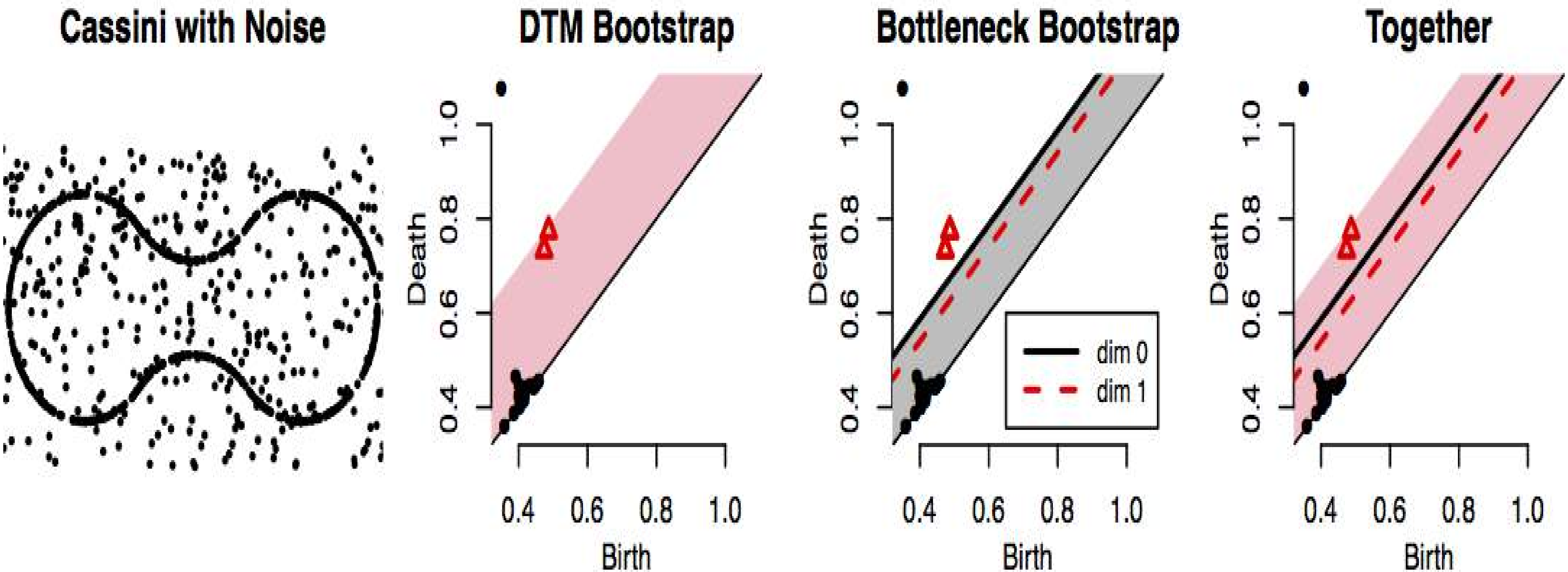
We directly bootstrap in the set of the persistence diagram by considering the random quantity $\mathrm{d_b}(\widehat{\mathsf{Dgm}}^*, \widehat{\mathsf{Dgm}})$. We define $\hat{t}_\alpha$ by

$$\mathbb{P}\left(\sqrt{n}\mathrm{d_b}(\widehat{\mathsf{Dgm}}^*, \widehat{\mathsf{Dgm}}) > \hat{t}_\alpha \,|\, X_1, \ldots, X_n\right) = \alpha.$$

The quantile $\hat{t}_\alpha$ can be estimated by Monte Carlo.

Bootstrap and significance of topological features

For both methods we can identify significant features by putting a band of size $2\hat{c}_\alpha$ or $2\hat{t}_\alpha$ around the diagonal:



In practice, the bottleneck bootstrap can lead to more precise inferences because in many cases the stability result is not sharp enough:

$$\mathrm{d_b}(\widehat{\mathsf{Dgm}}, \mathsf{Dgm}) \leq \|d_{P,m} - d_{P_n,m}\|_\infty.$$

Concluding remarks

- TDA methods focus on the topological properties (homology / persistent homology) of a shape.
- TDA methods can be used
  - as an "exploratory method", in particuar when the point cloud is sampled on (close to) a real geometric object
  - as a "feature extraction" procedure, next these extracted features can be used for learning purposes.
- TDA is an emerging field, at the interface maths, computer sciences, statistics.
- Many topics about the statistical analysis of TDA
- Applications in many fields of sciences (medecine, biology, dynamic systems, astronomy, dynamical systems, physics ...)