Stochastic Geometry Conference

Nantes, April 6 2016

A Statistical Approach to Topological Data Analysis

Bertrand MICHEL

Laboratoire de Statistique Théorique et Appliquée
Université Pierre et Marie Curie

I - Introduction : Statistics and Topological Data Analysis

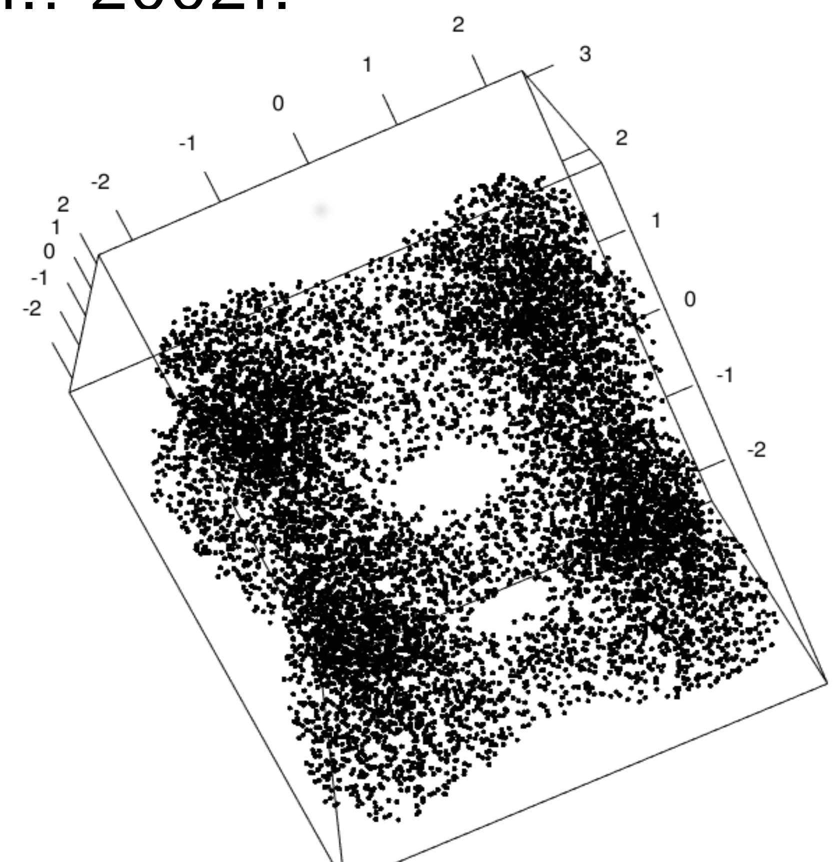
Topological data analysis and topological inference

- **Geometric inference**, **algebraic topology tools** and **computational topology** have recently witnessed important developments with regards to data analysis, giving birth to the field of **topological data analysis** (TDA).

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

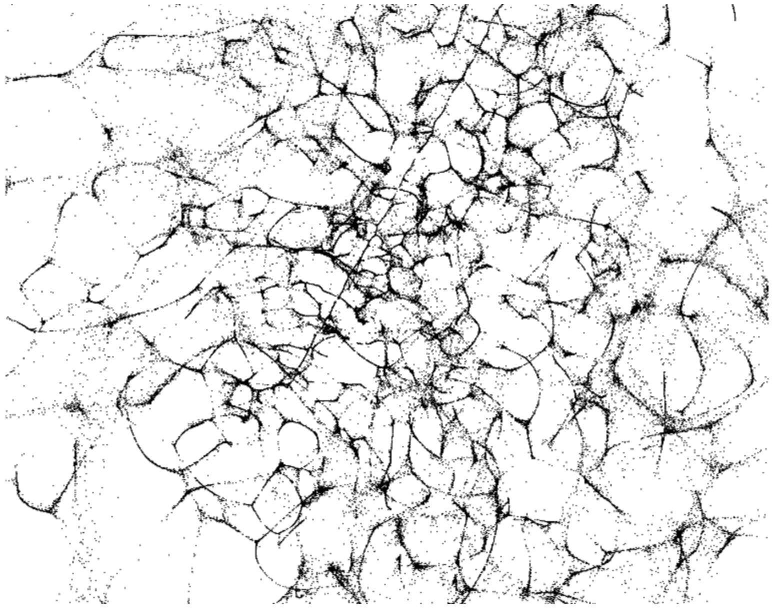
- The two popular methods in TDA : **Mapper algorithm** [Singh et al., 2007] and **persistent homology** [Edelsbrunner et al., 2002].

- **Topological inference** methods aim to infer topological properties of an unknown topological space $\mathbb{X}$, typically from a point cloud $\mathbb{X}_n$ "close" to $\mathbb{X}$.

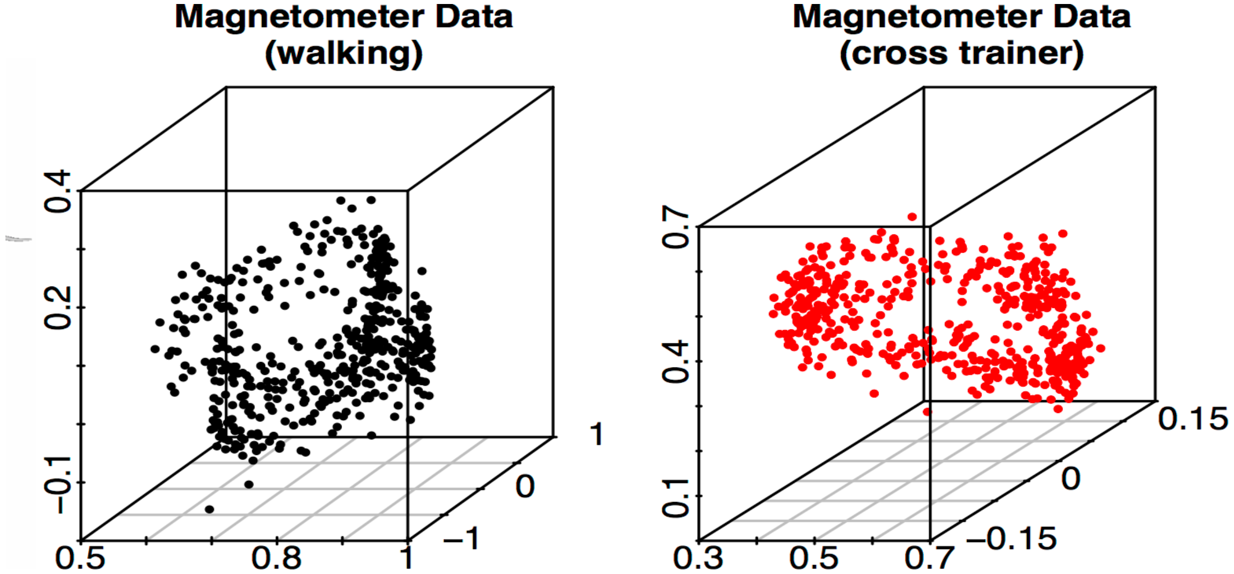


Application fields of TDA methods

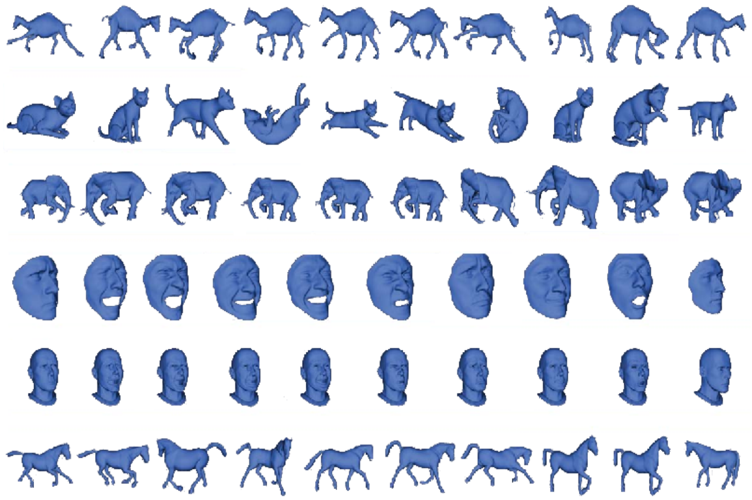
[distribution of galaxies]

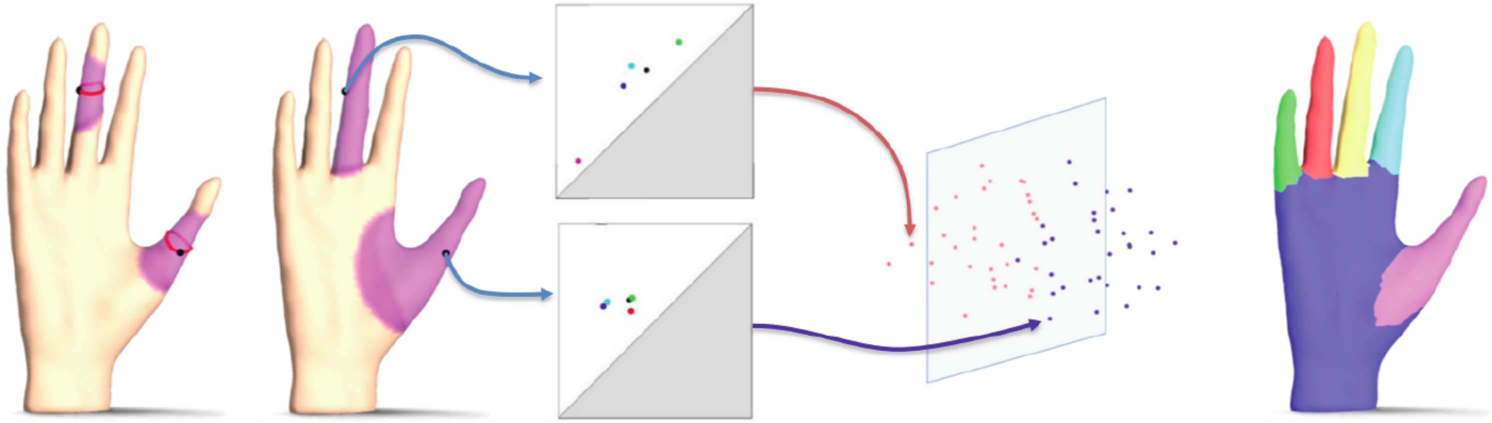


[Sensor Data]



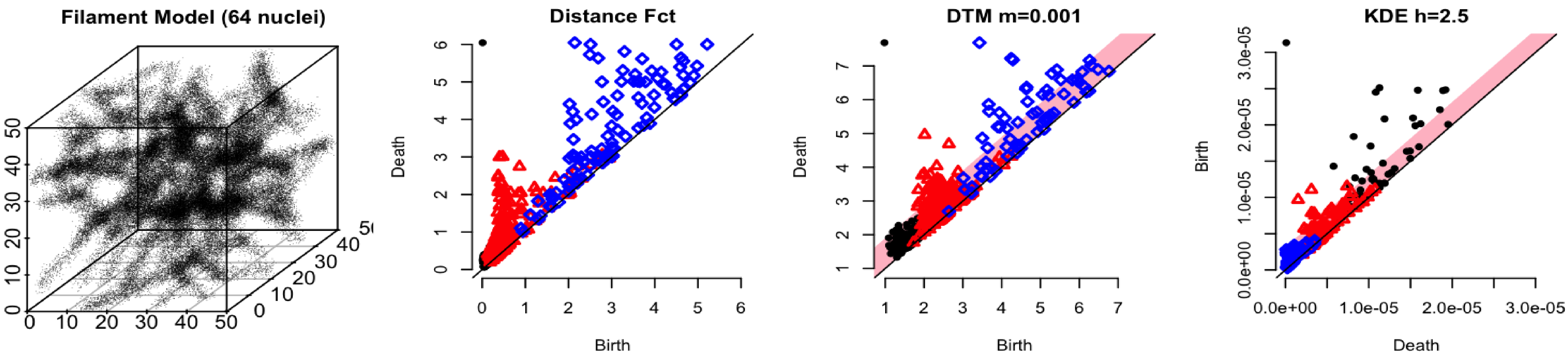
[3D shape database]





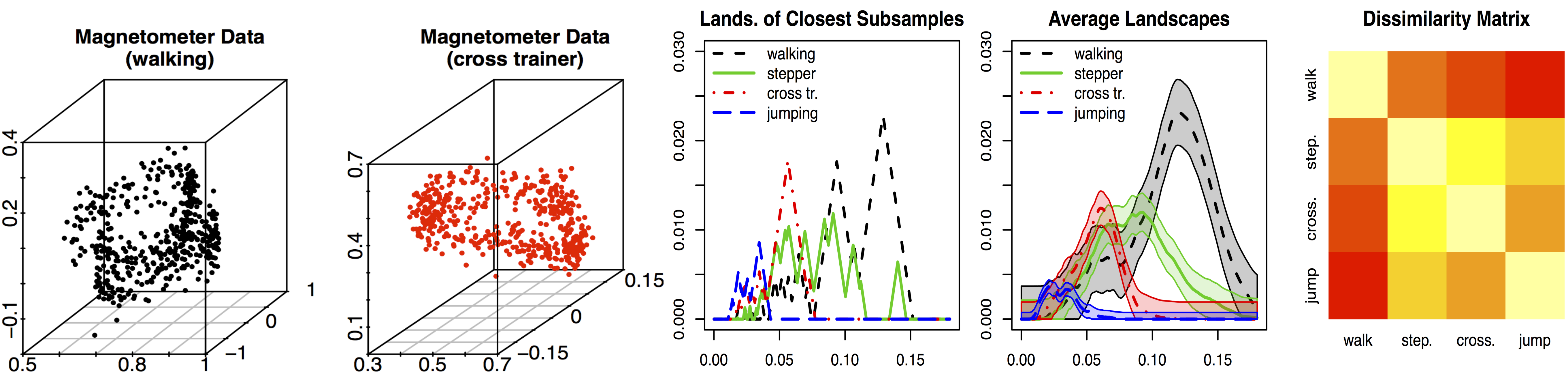
Topological data analysis methods can be used:

- For **exploratory analysis**, visualization:



[Chazal et al., 2014b]

- For **feature extraction** in supervised settings (prediction) :



[Chazal et al., 2015a]

Statistics and TDA

Until very recently, TDA and topological inference mostly relied on deterministic approaches. Alternatively, 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.

Non exhaustive list of questions for a statistical approach to TDA :

- proving consistency of TDA methods.
- providing confidence regions for topological features and discussing the significance of the estimated topological quantities.
- selecting relevant scales at which the topological phenomenon should be considered.
- dealing with outliers and providing robust methods for TDA.
- ...

II- Homology and Persistent homology

Basic tools for TDA : Offsets and Simplicial Complexes

Point clouds in themselves do not carry any non trivial topological or geometric structure.

For a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ (or in a metric space),
the α-offset of $\mathbb{X}_n$ is defined by

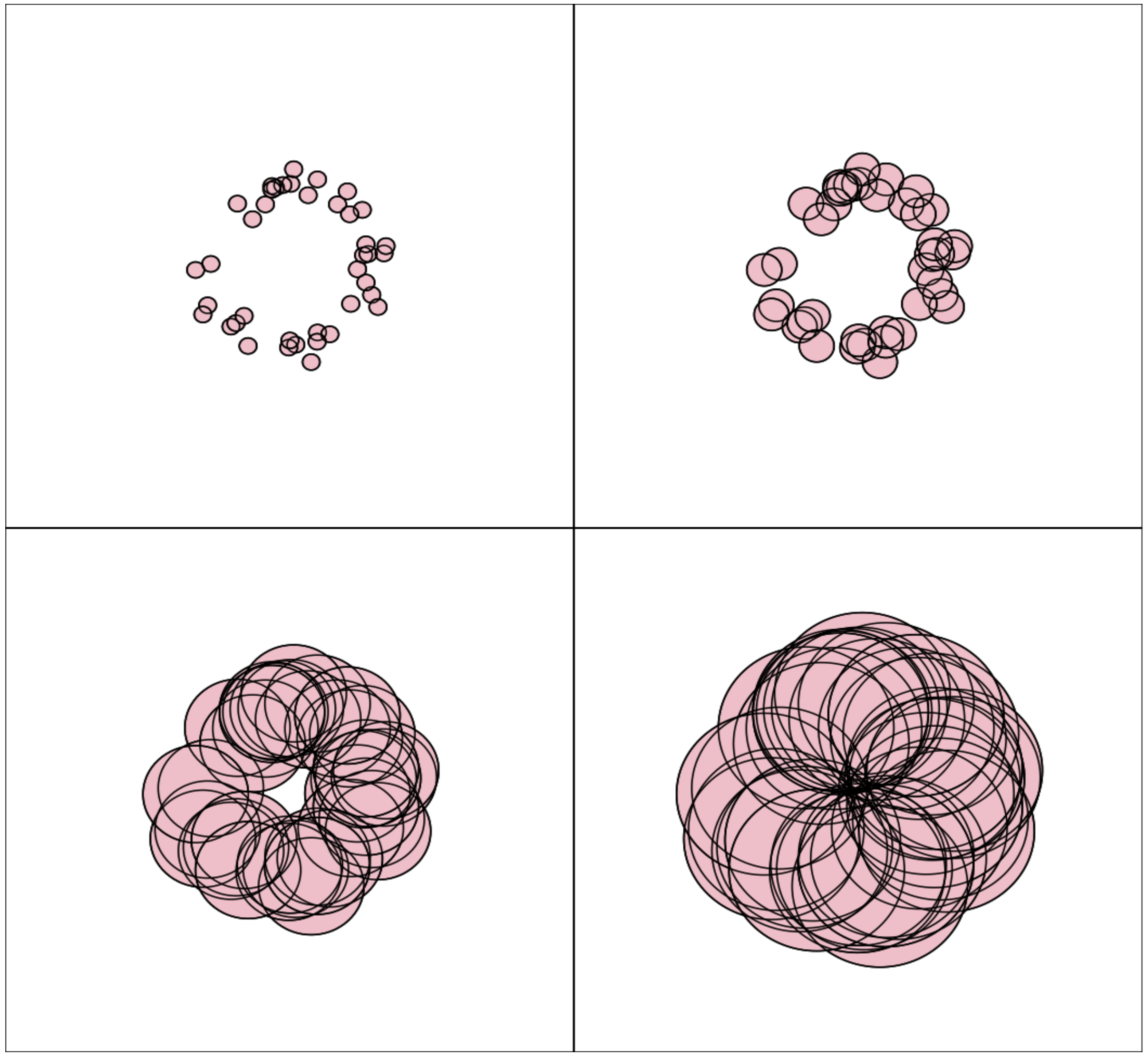
$$\mathbb{X}_n^\alpha = \bigcup_{x \in \mathbb{X}_n} B(x, \alpha).$$

More generally, for any compact set $\mathbb{X}$,

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

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

$$d_{\mathbb{X}}(y) = \inf_{x \in \mathbb{X}} \|x - y\| \qquad \text{(in } \mathbb{R}^d\text{)}$$



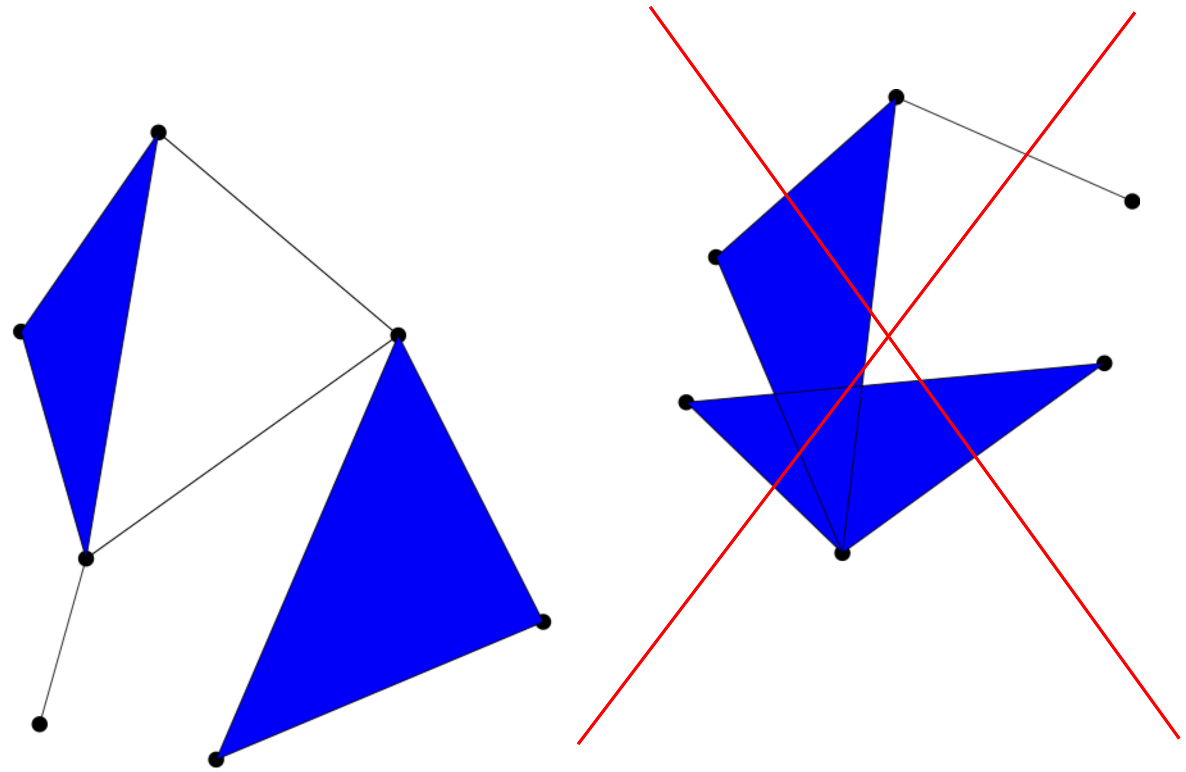
General idea: deduce from $(\mathbb{X}_n^\alpha)_{r>0}$ some topological and geometric information of an underlying object.

Basic tools for TDA : Offsets and Simplicial Complexes

Non-discrete sets such as offsets, and also continuous mathematical shapes like curves, surfaces cannot easily be encoded as finite discrete structures.

A geometric simplicial complex $\mathcal{C}$ is a set of simplices such that:

- Any face of a simplex from $\mathcal{C}$ is also in $\mathcal{C}$.
- The intersection of any two simplices s_1, $s_2 \in \mathcal{C}$ is either a face of both s_1 and s_2, or empty.

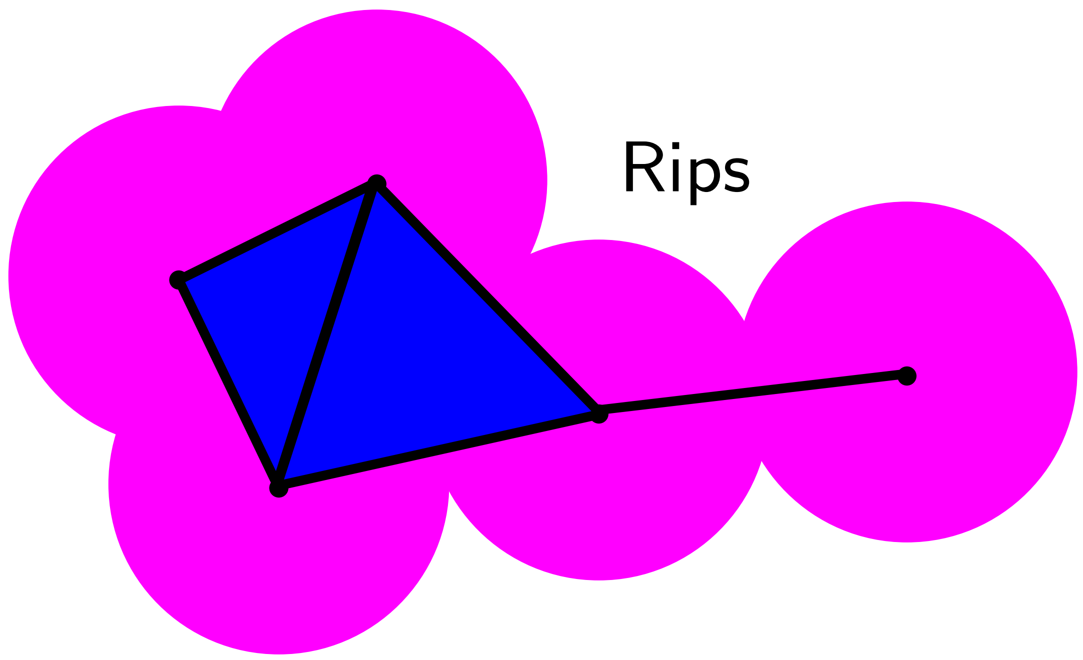


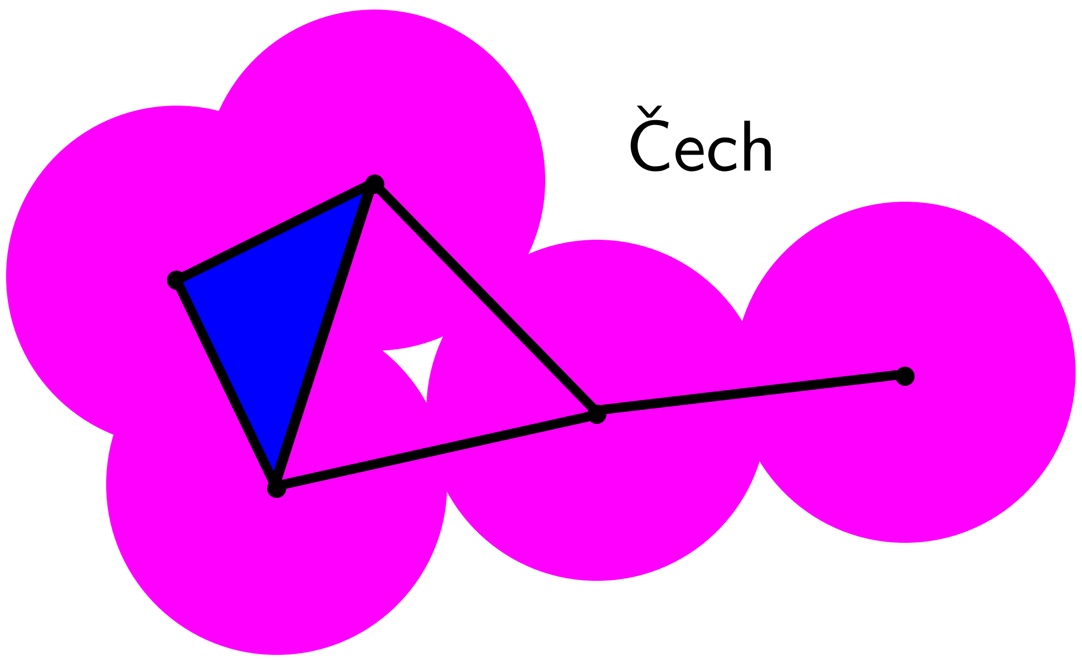
Basic tools for TDA : Offsets and Simplicial Complexes

Examples:

- A simplex $[x_0, x_1, \cdots, x_k]$ is in the Čech complex $\check{\mathbb{C}}\mathrm{ech}_\alpha(\mathbb{X}_n)$ if and only if $\bigcap_{j=0}^{k} B(x_j, \alpha) \neq \emptyset$.
- A simplex $[x_0, x_1, \cdots, x_k]$ is in the Rips complex $\mathbb{R}\mathrm{ips}_\alpha(\mathbb{X}_n)$ if and only if $\|x_j - x_{j'}\| \leq \alpha$ for all $j, j' \in \{1, \ldots, k\}$.

Can be also defined for a set of points in any metric space or for any compact metric space.

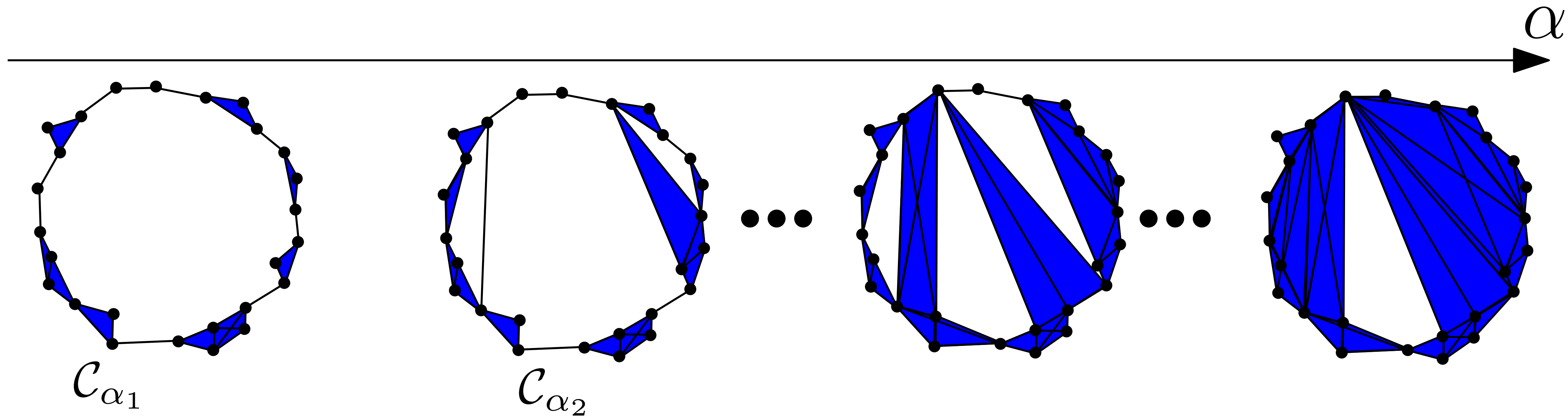




Nerve Theorem : the offsets $\mathbb{X}_n^\alpha$ of a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$ are homotopy equivalent to the Čech complex $\check{\mathbb{C}}\mathrm{ech}_\alpha(\mathbb{X}_n)$

Filtrations of simplicial complexes

Given a point cloud $\mathbb{X}_n$ in $\mathbb{R}^d$, we generally define a **filtration** of (nested simplicial) complexes by considering all the possibles scale parameters α : $(\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$

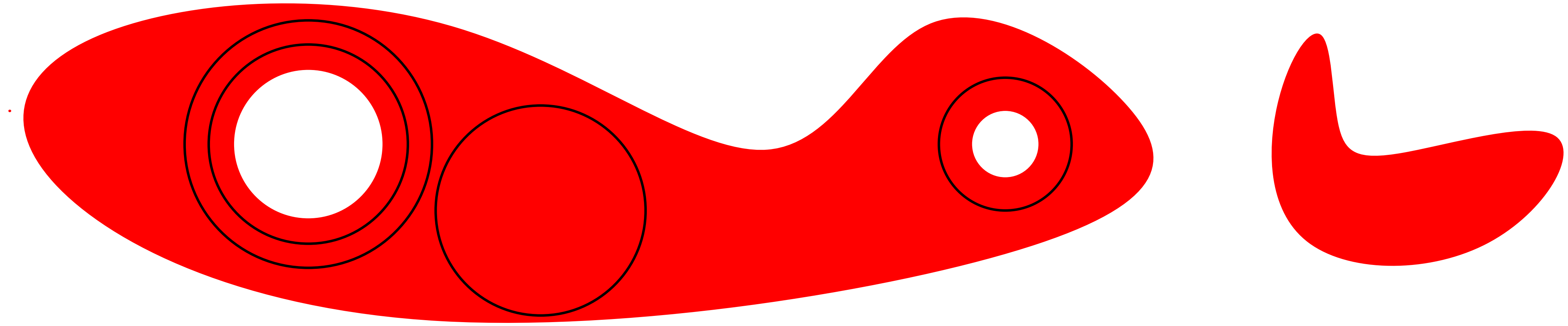


Homology inference

- **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.

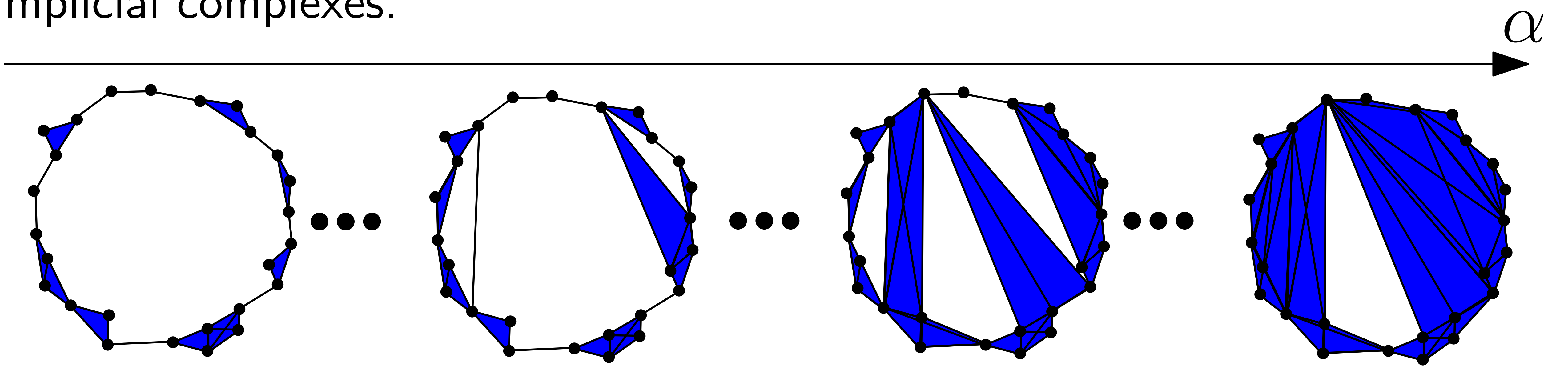
- **Computational Topology** : Betti numbers can be computed on simplicial complexes.



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

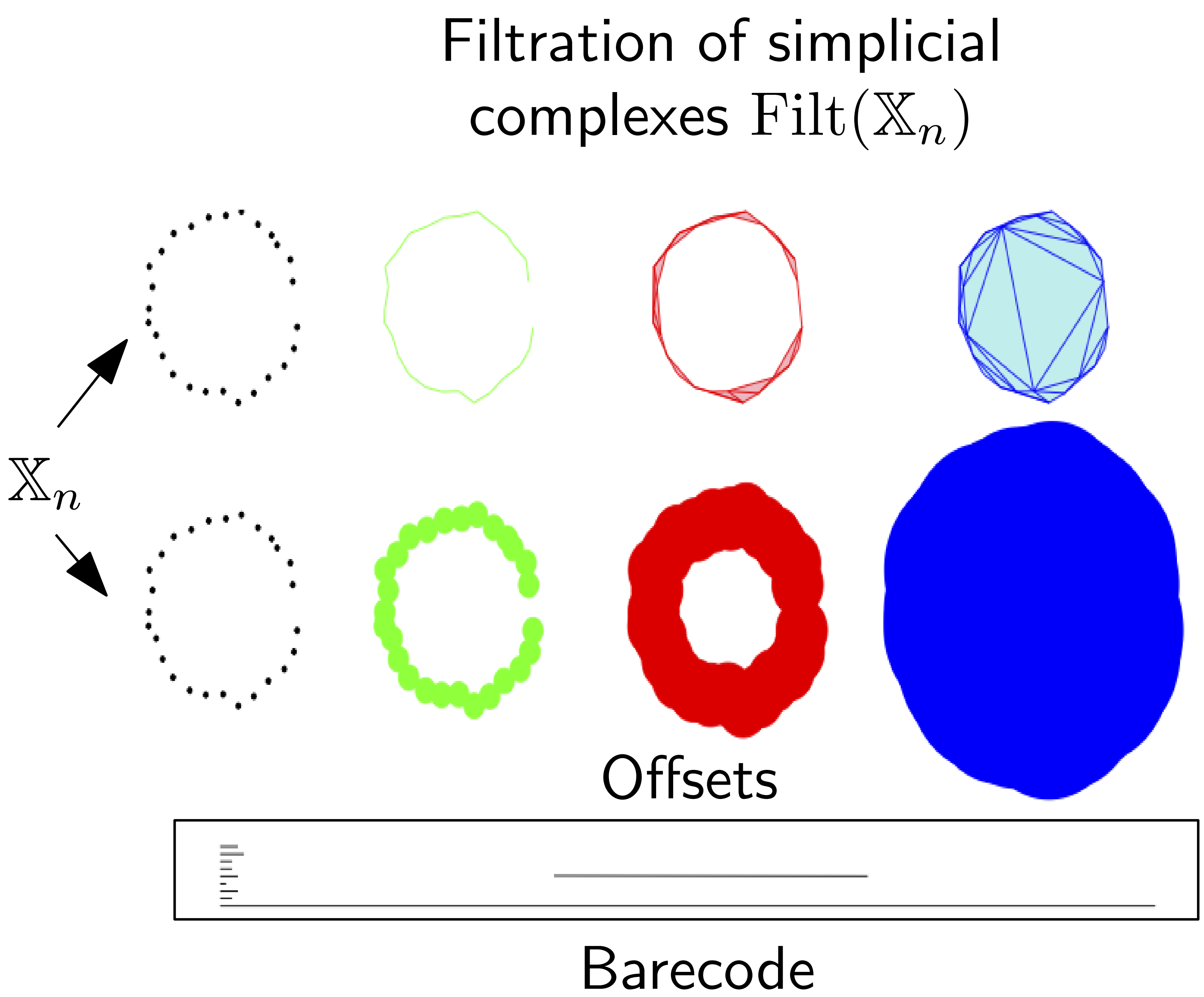
Starting from a point cloud $\mathbb{X}_n$, let $\mathrm{Filt} = (\mathcal{C}_\alpha)_{\alpha \in \mathcal{A}}$ be a fitration of nested simplicial complexes.



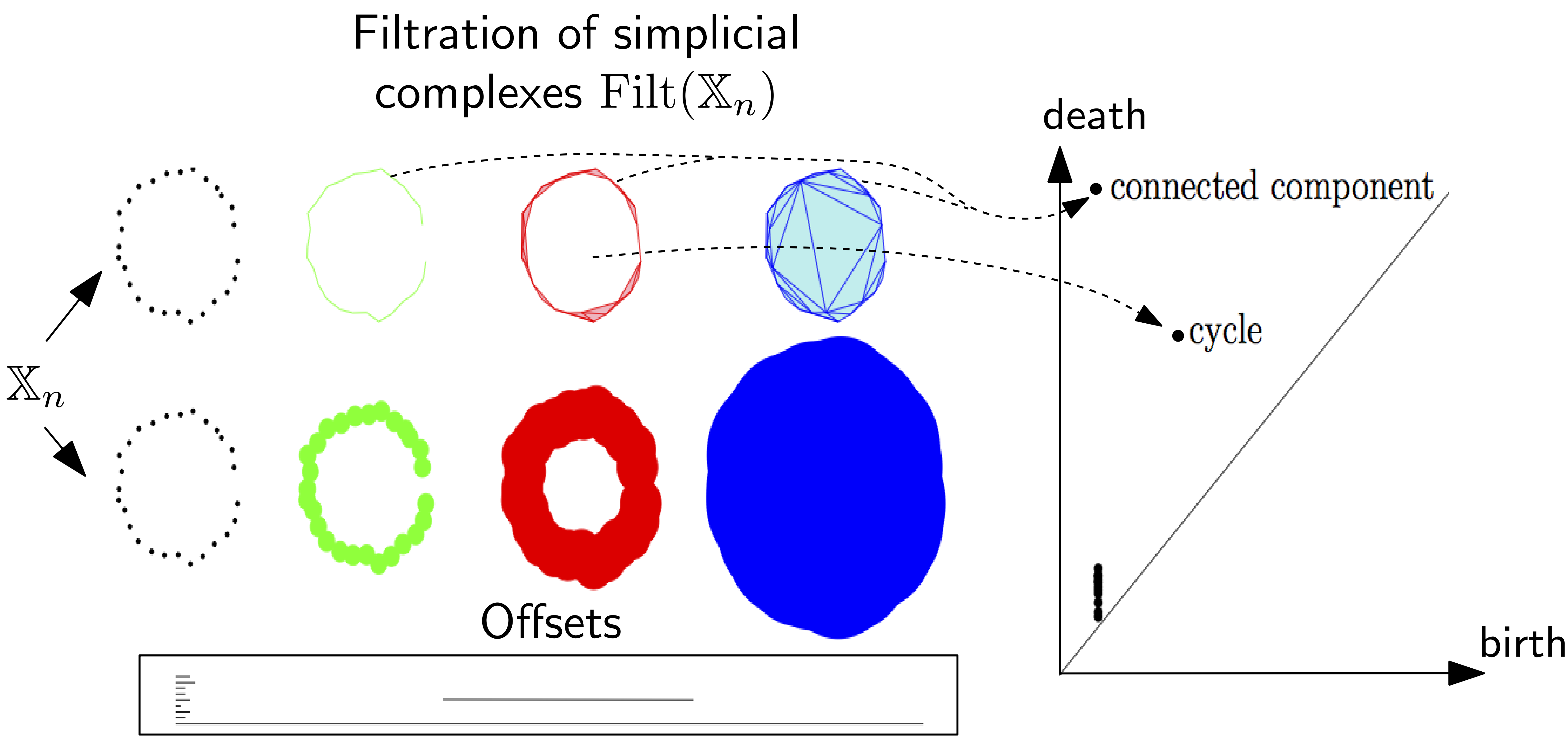
Persistent homology: identification of "persistent" topological features along the filtration.

- multiscale information ;
- more stable and more robust ;
- (but does not answer the scale selection problem...)

Barecodes and Persistence Diagrams

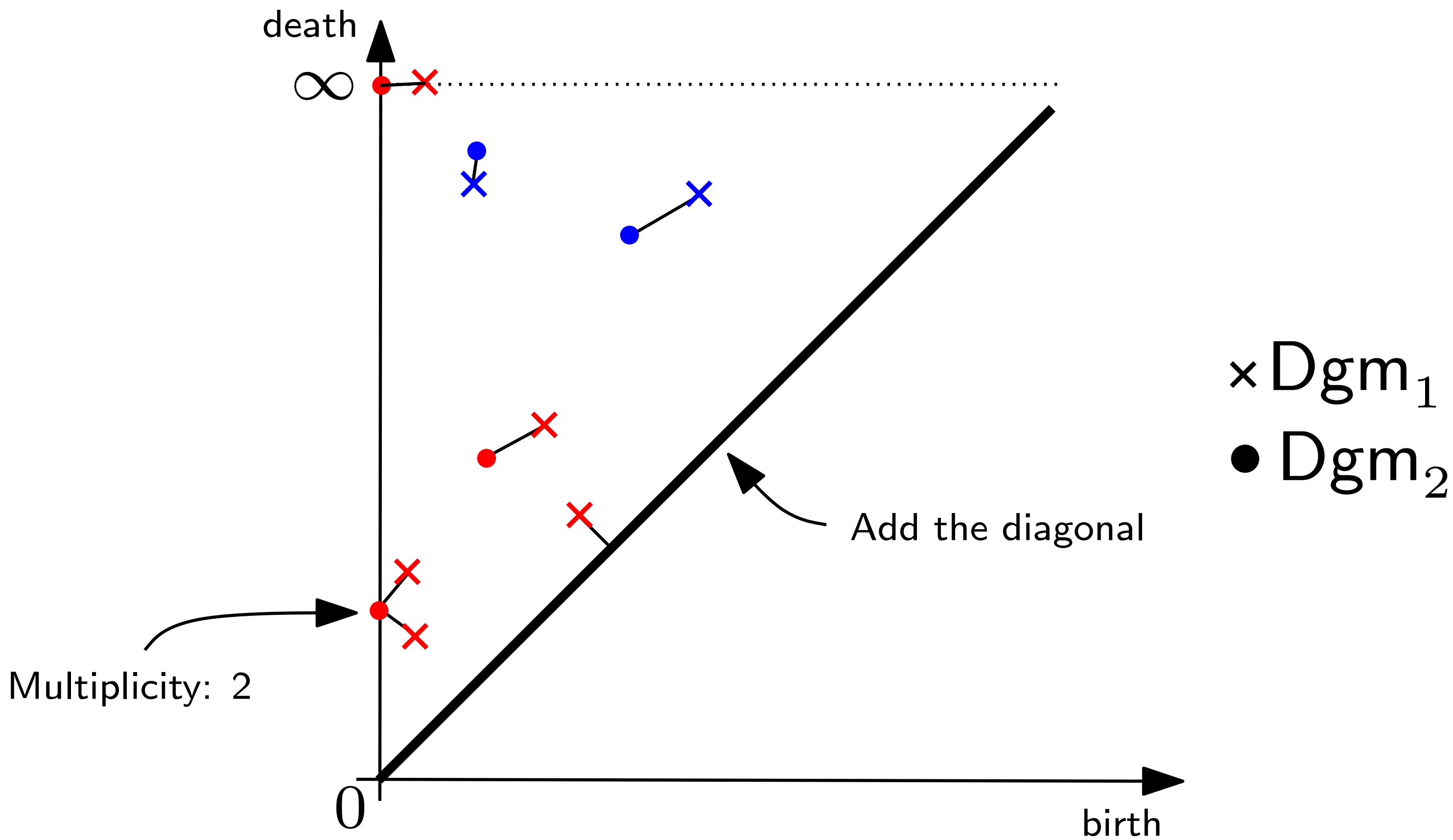


Barecodes and Persistence Diagrams



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

Distance between persistence diagrams and stability



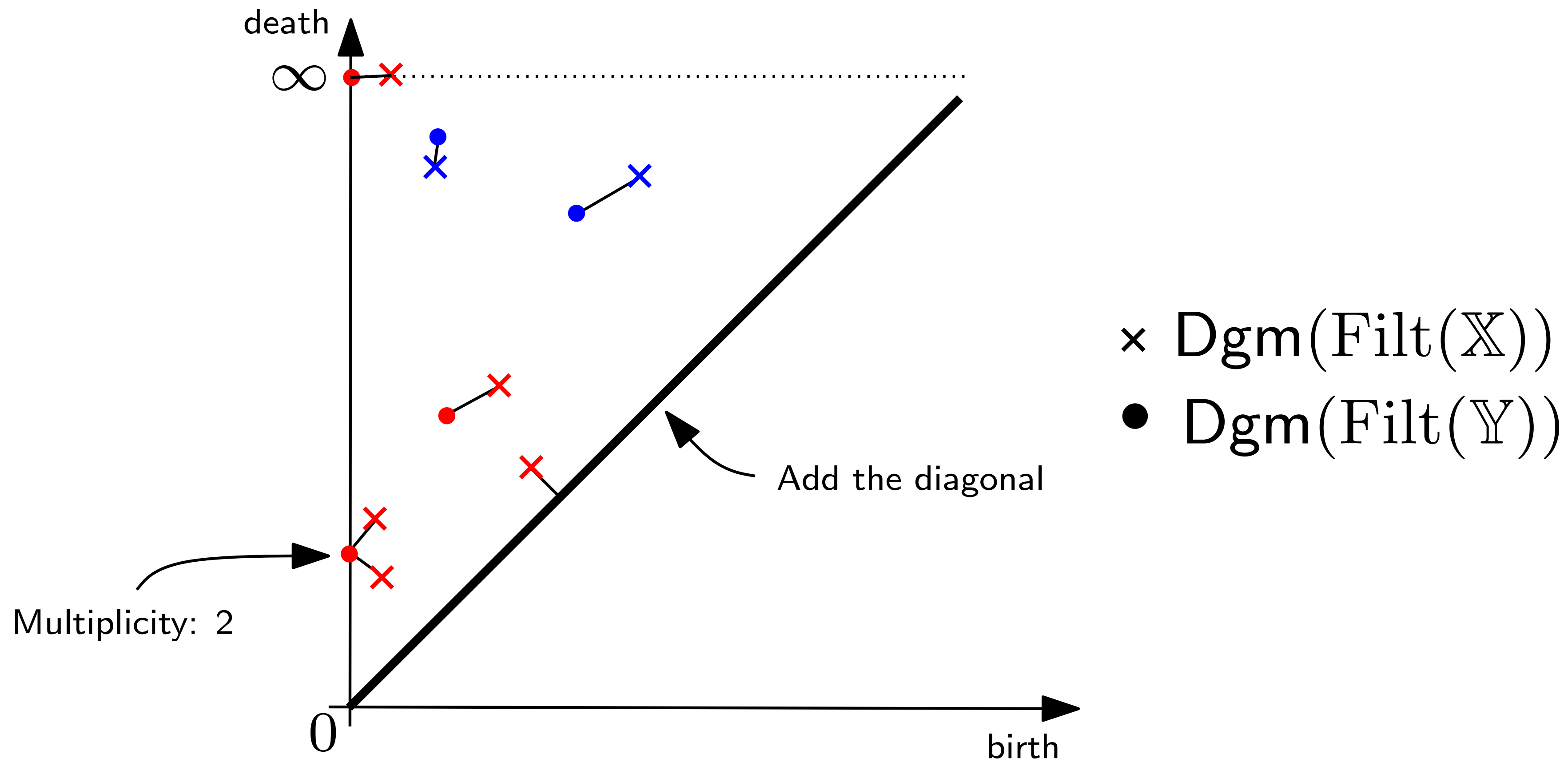
The bottleneck distance between two diagrams Dgm_1 and Dgm_2 is

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

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 $(\mathbb{Y}, \rho')$,

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

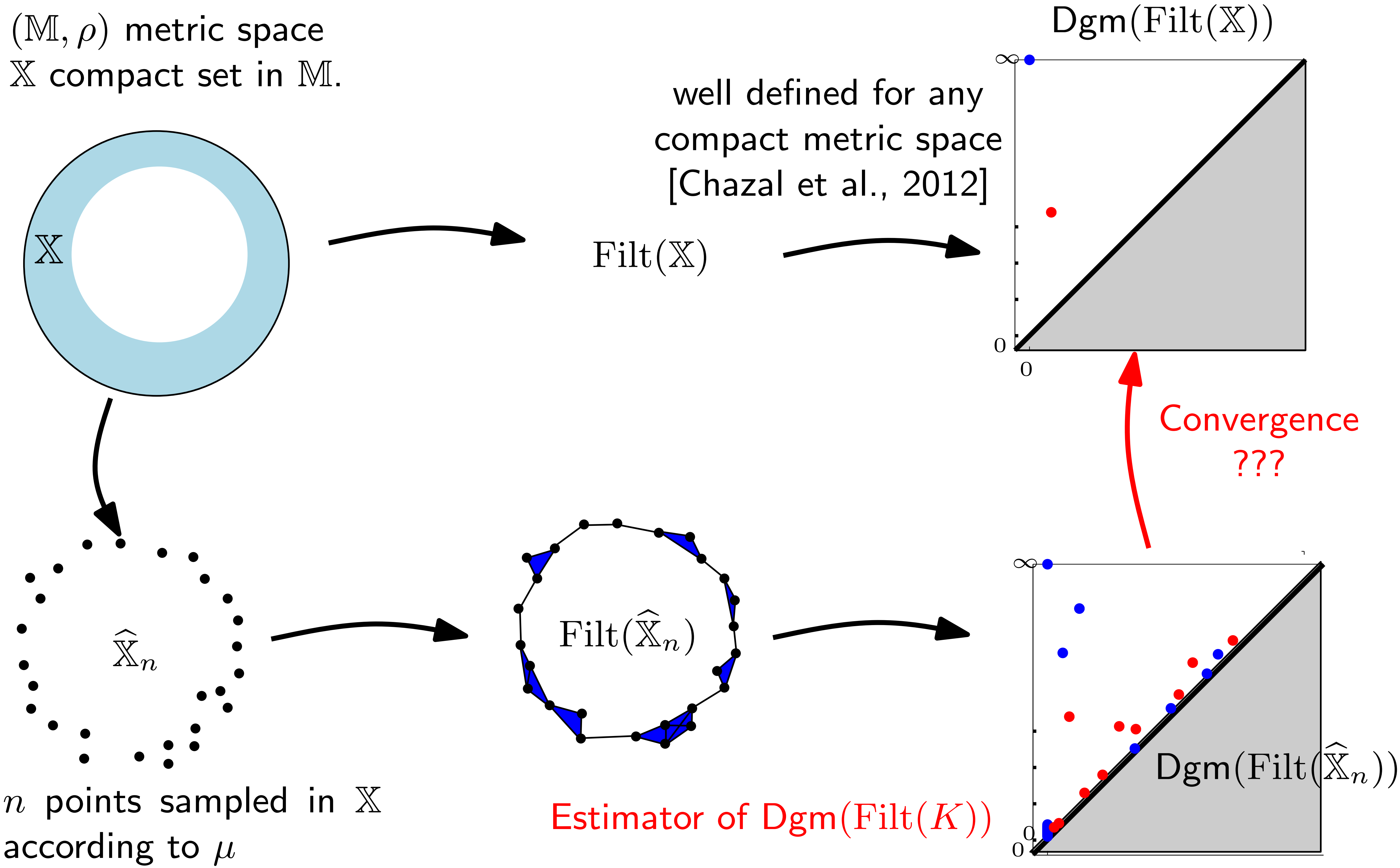
Consequently, if $\mathbb{X}$ and $\mathbb{Y}$ are embedded in the same metric space $(\mathbb{M}, \rho)$ then

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

III - Statistics
and
Persistent homology

Persistence diagram inference [Chazal et al., 2014b]

Joint work with F. Chazal, M. Glisse and C. Labruère.



Persistence diagram inference [Chazal et al., 2014a]

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 $a, b > 0$:

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

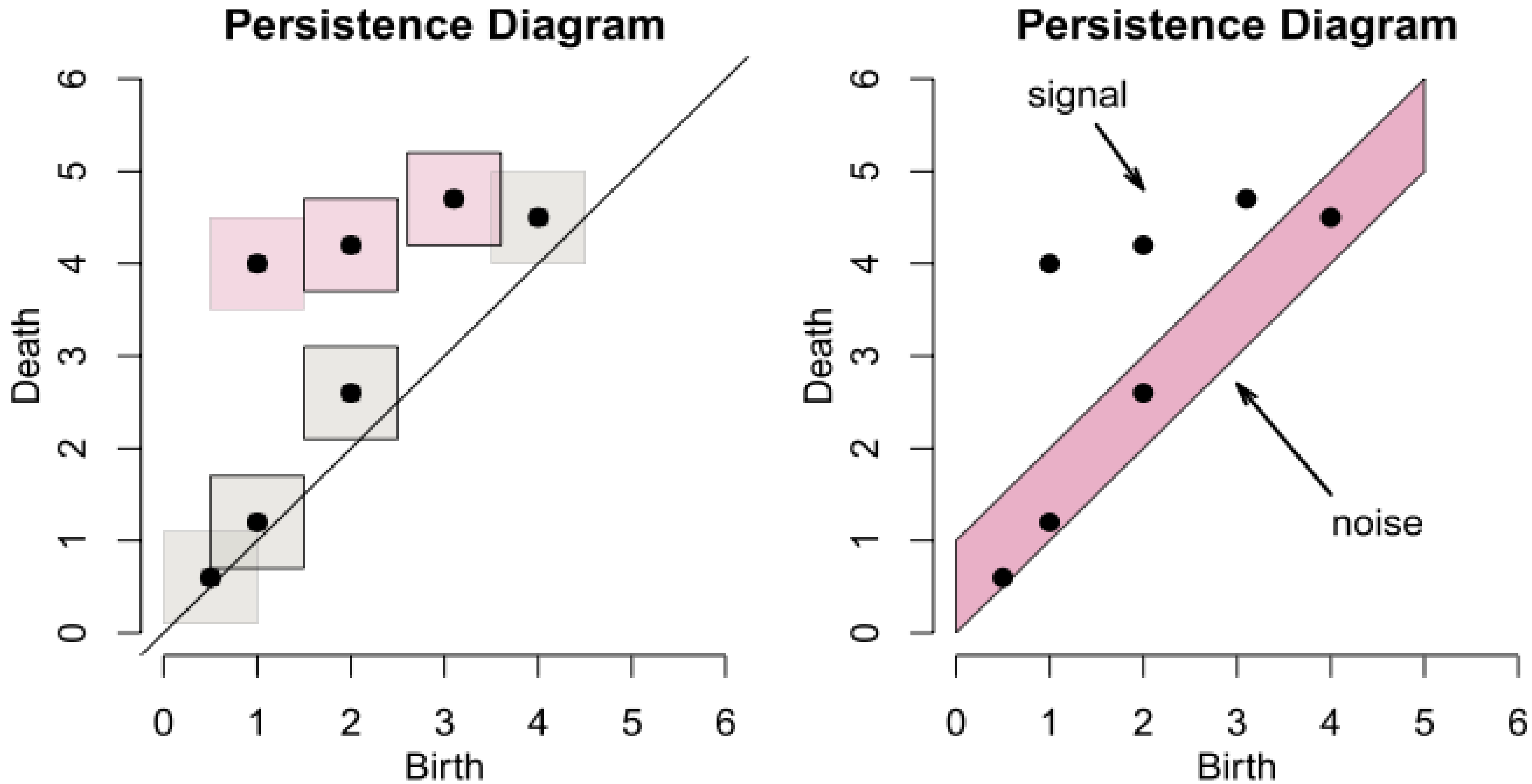
where C only depends on a and b.

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

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

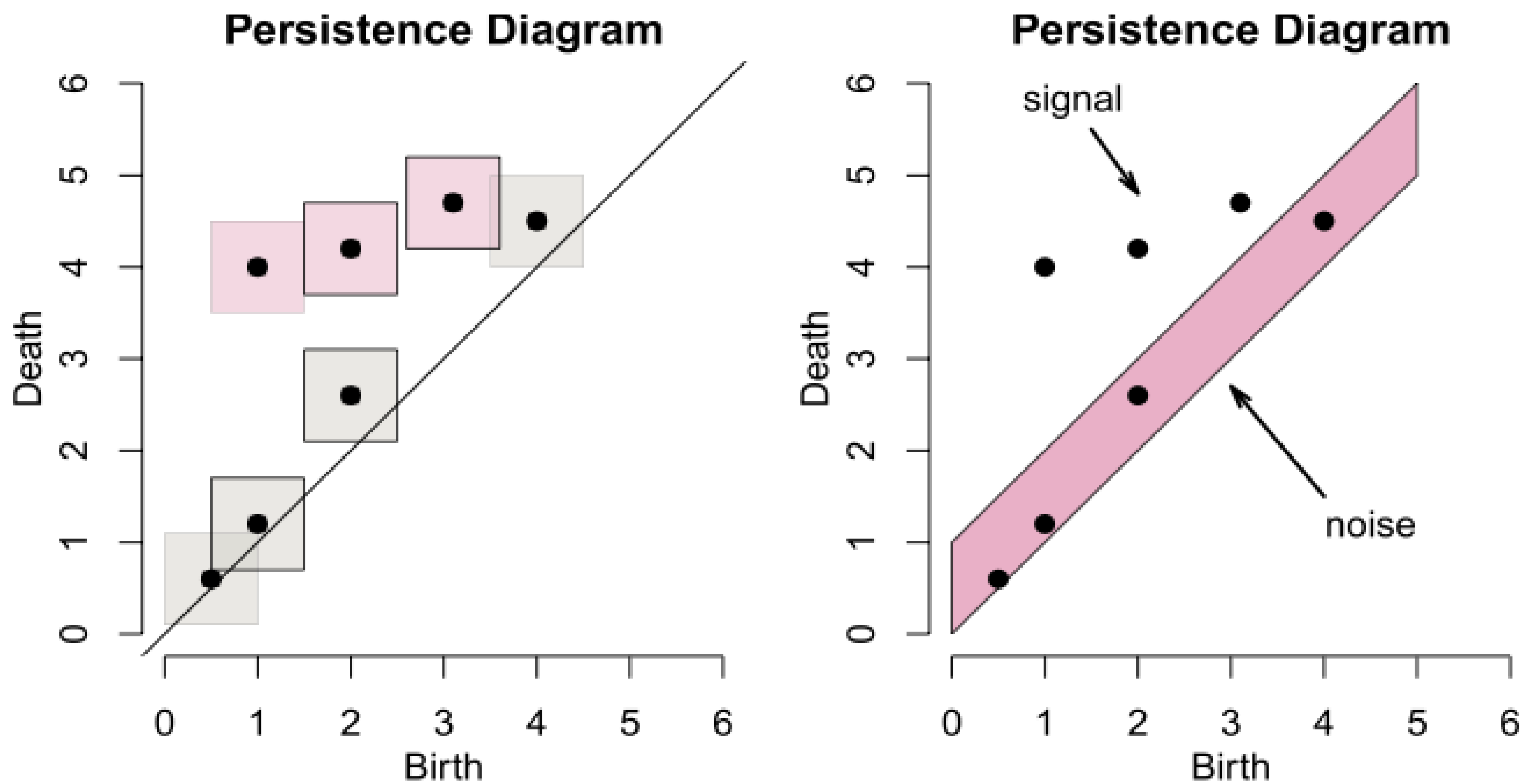
where C' is an absolute constant.

Confidence sets for persistence diagrams [Fasy et al., 2014]



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

Confidence sets for persistence diagrams [Fasy et al., 2014]



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

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

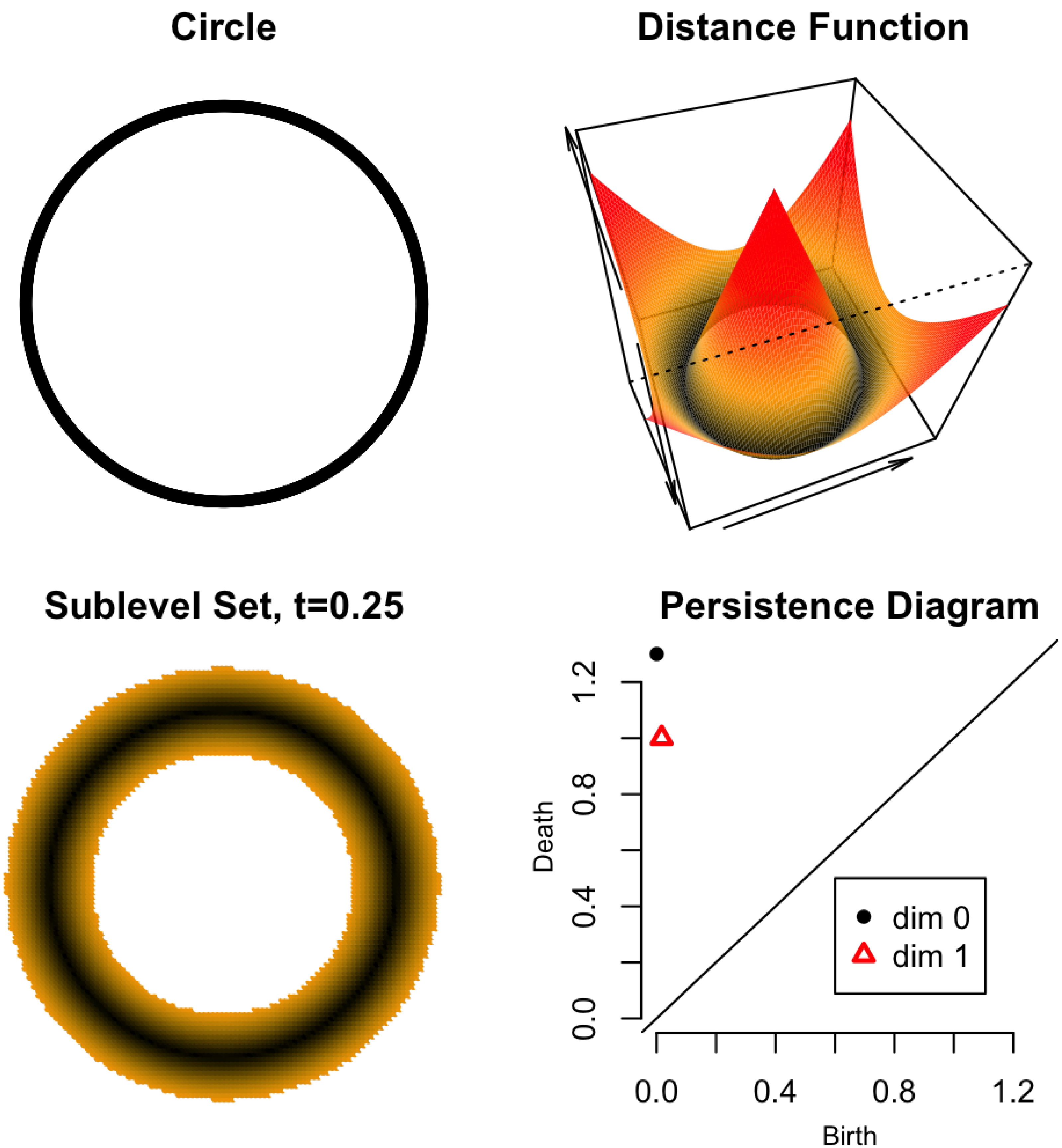
$$\mathrm{d_b}\left(\mathsf{Dgm}\left(\mathrm{Filt}(K)\right), \mathsf{Dgm}\left(\mathrm{Filt}(\mathbb{X}_n)\right)\right) \leq \mathrm{d_H}(K, \mathbb{X}_n).$$

It is sufficient to find c_n such that

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

IV - Robust distance functions for TDA and geometric inference

Standard TDA methods are not robust to outliers

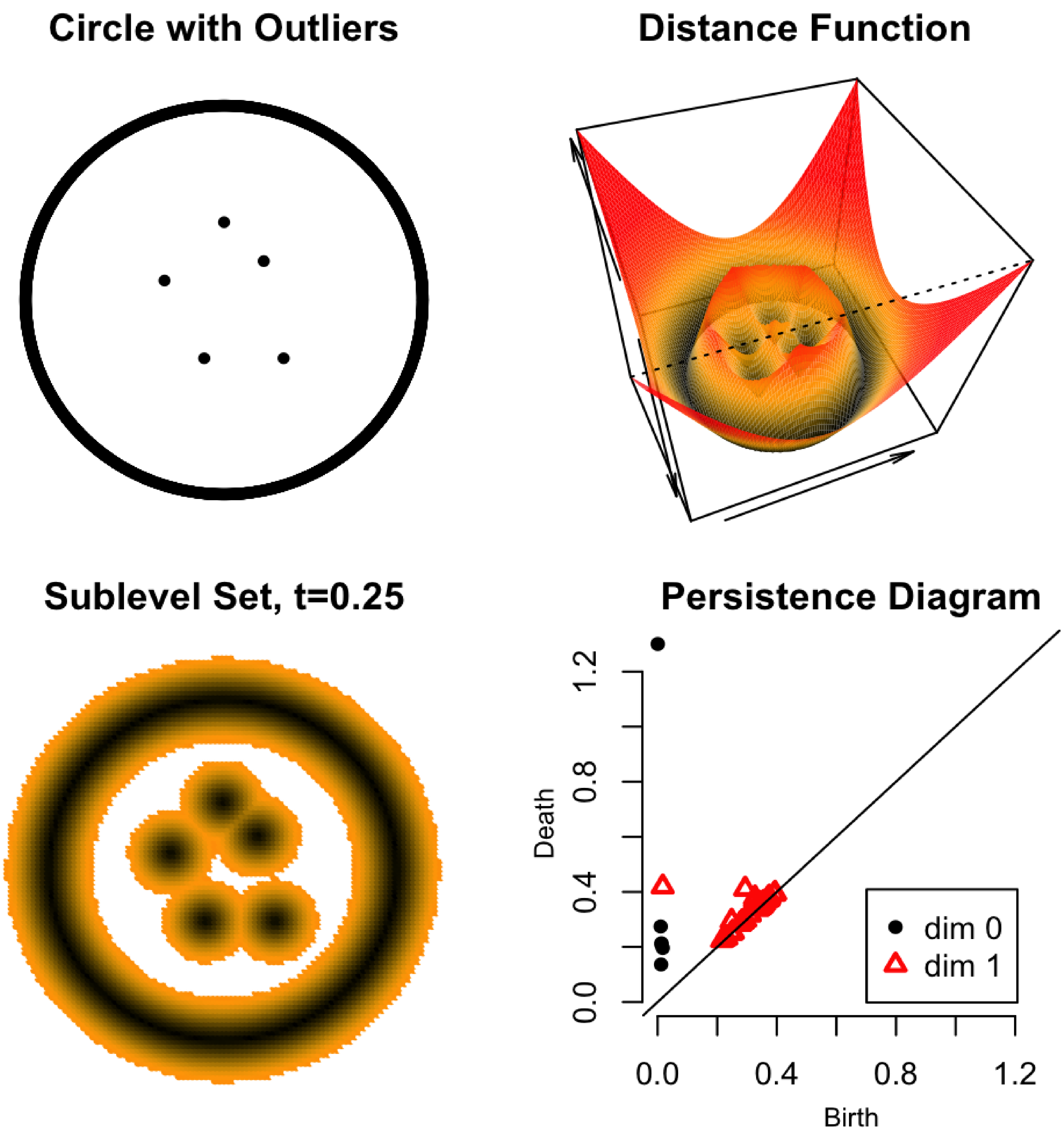


$$\mathbb{X}^r := \bigcup_{x \in \mathbb{X}} B(x, r) = d_{\mathbb{X}}^{-1}([0, r])$$

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

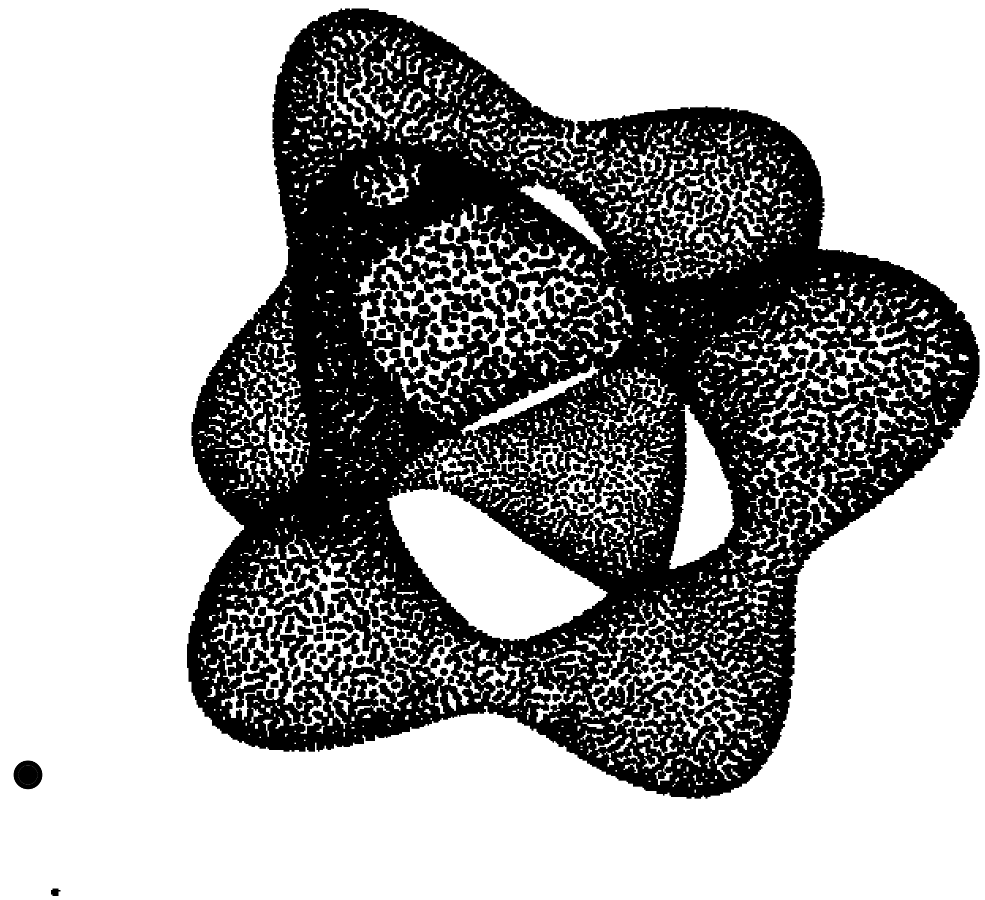


$$\mathbb{X}^r := \bigcup_{x \in \mathbb{X}} B(x, r) = d_{\mathbb{X}}^{-1}([0, r])$$

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

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

Robust TDA with an alternative distance function ?

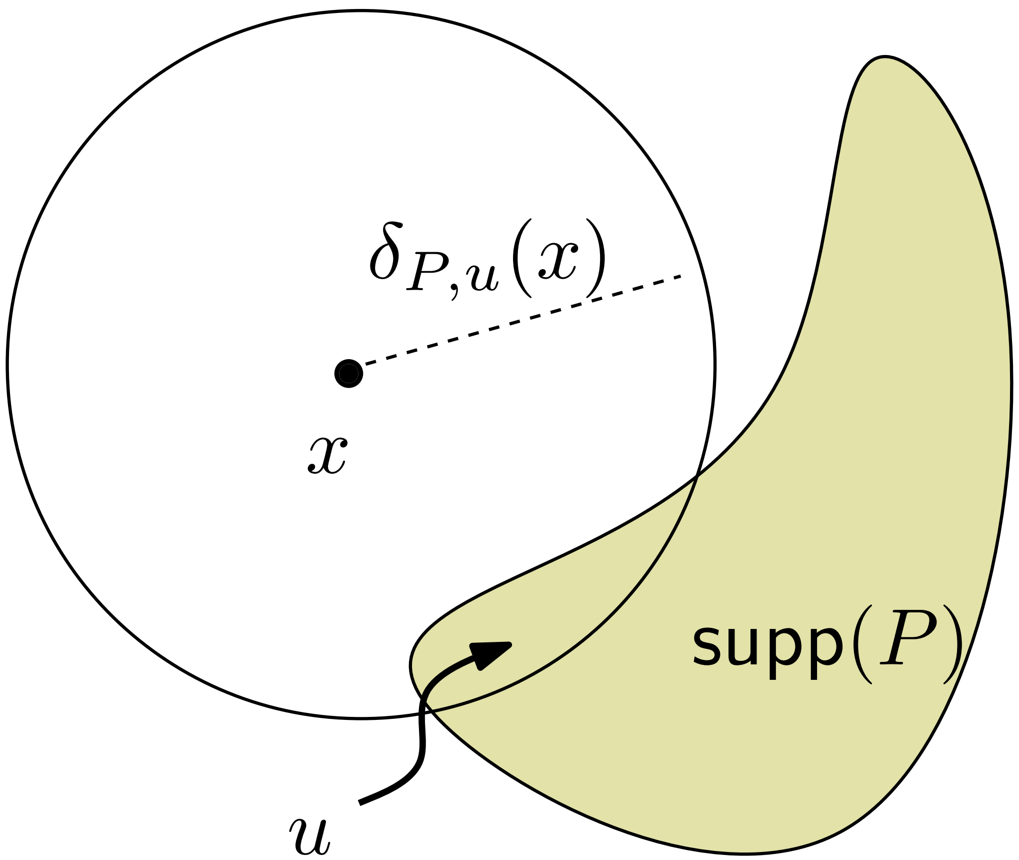


We would like to consider the sub levels of an alternative distance function related to the sampling measure, which support is $\mathbb{X}$, or close to $\mathbb{X}$.

Distance To Measure [Chazal et al., 2011]

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

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



$\delta_{P,u}$ 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.

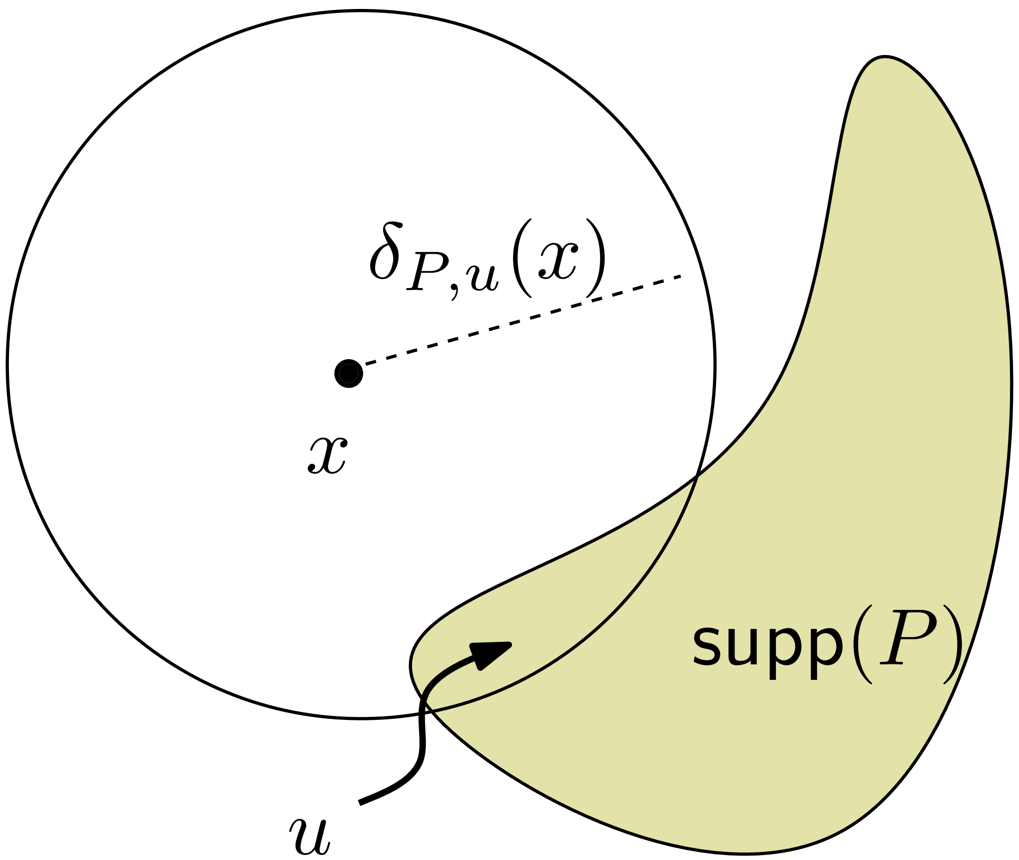
$$\|x - X\|$$

where $X \sim P$.

Distance To Measure [Chazal et al., 2011]

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

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



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

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

Distance To Measure [Chazal et al., 2011]

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.

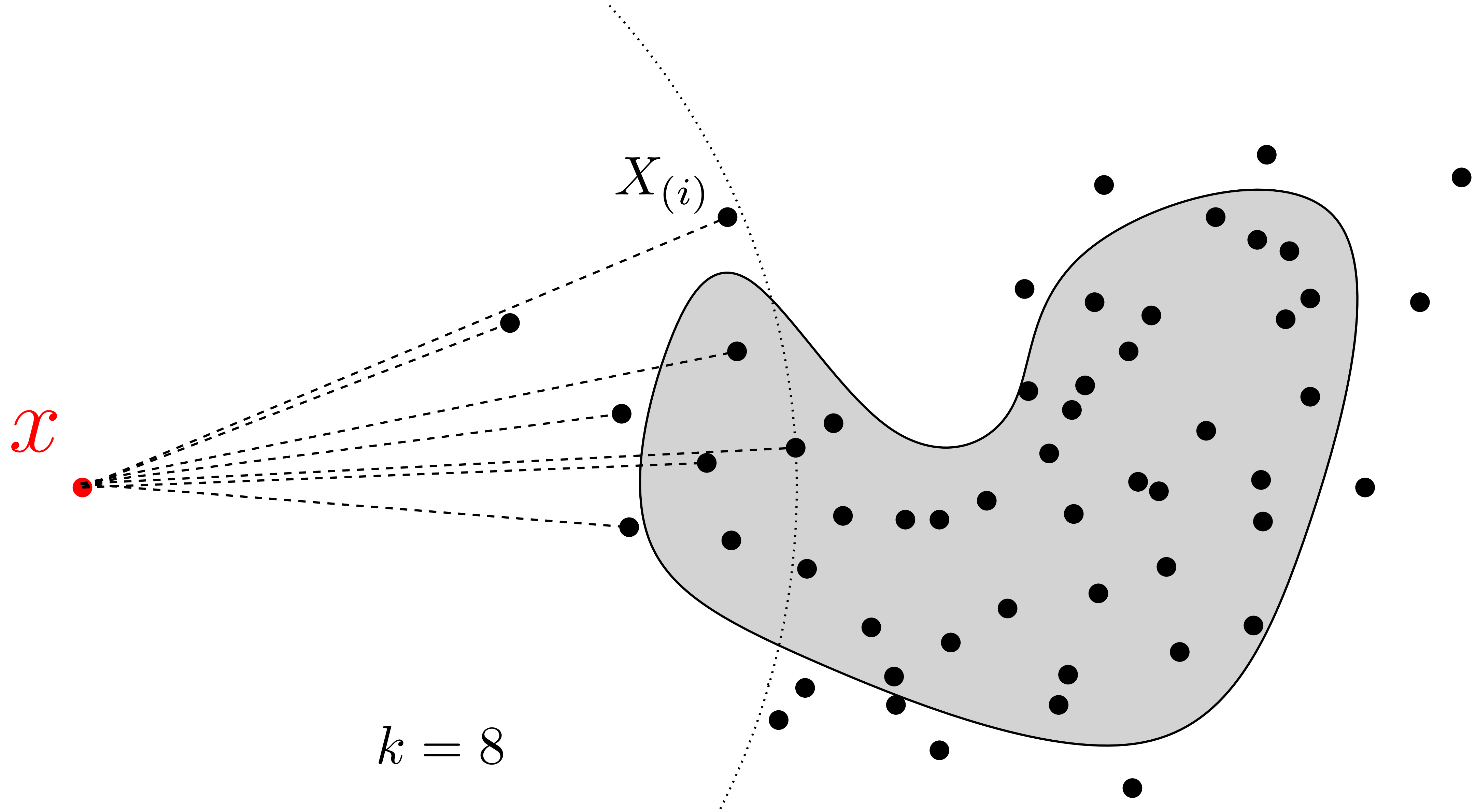
- Consequently, if $\tilde{P}$ is a probability distribution close to P for Wasserstein distance W_2, then the sublevel sets of $d_{\tilde{P},m}$ provide a topologically correct approximation of the support of P.

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

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

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



Estimation of the DTM via the empirical DTM

[Chazal et al., 2014b] and [Chazal et al., 2015b]

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 et al., 2014b] and [Chazal et al., 2015b]

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.

Functional convergence [Chazal et al., 2014b]

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

Modulus of continuity $\tilde{\omega}_x$ **of** F_x^{-1} : for any $v \in (0,1]$

$$\tilde{\omega}_x(v) := \sup_{(u,u') \in [0,1]^2,\, u \neq u',\, \|u-u'\| \leq v} |F_x^{-1}(u) - F_x^{-1}(u')|.$$

Theorem: Let P be a measure on $\mathbb{R}^d$ with compact support. Let $\mathcal{D}$ be a compact domain on $\mathbb{R}^d$ and $m \in (0,1)$. Assume that there exists an uniform upper bound $\omega_{\mathcal{D}}$ on the modulus of continuity for the family $(F_x^{-1})_{x \in \mathcal{D}}$ satisfying

$$\lim_{u \to 0} \omega_{\mathcal{D}}(u) = \omega_{\mathcal{D}}(0) = 0.$$

Then $\sqrt{n}(d^2_{P_n,m} - d^2_{P,m})$converges in distribution to $\mathbb{B}$ on $\mathcal{D}$, where $\mathbb{B}$ is a centered Gaussian process with covariance kernel

$$\kappa(x,y) = \frac{1}{m^2} \int_0^{F_x^{-1}(m)} \int_0^{F_y^{-1}(m)} \left(\mathbb{P}\left[B(x,\sqrt{t}) \cap B(y,\sqrt{s}) \right] - F_x(t) F_y(s) \right) ds\, dt.$$

Fluctuations of the DTEM [Chazal et al., 2015b]

joint work with F. Chazal and P. Massart

Theorem: Let x be a fixed observation point in $\mathbb{R}^d$. Assume that $\omega_x : (0,1] \to \mathbb{R}^+$ is an upper bound on the modulus of continuity of F_x^{-1}. Let $k < \frac{n}{2}$. For any $\lambda > 0$:

$$P\left(\left|d^2_{P_n, \frac{k}{n}}(x) - d^2_{P, \frac{k}{n}}(x)\right| \geq \lambda\right) \leq 2\exp\left(-\frac{n}{8}\frac{\frac{k}{n}}{\left[\omega_x\left(\frac{k}{n}\right)\right]^2}\lambda^2\right) + \ldots$$

Assume moreover that $\omega_x(u)/u$ is a non increasing function, then

$$\mathbb{E}\left(\left|d^2_{P_n, \frac{k}{n}}(x) - d^2_{P, \frac{k}{n}}(x)\right|\right) \leq \frac{C}{\sqrt{n}}\sqrt{\frac{n}{k}}\omega_x\left(\frac{k}{n}\right).$$

renormalization by the mass proportion

localization at the origin

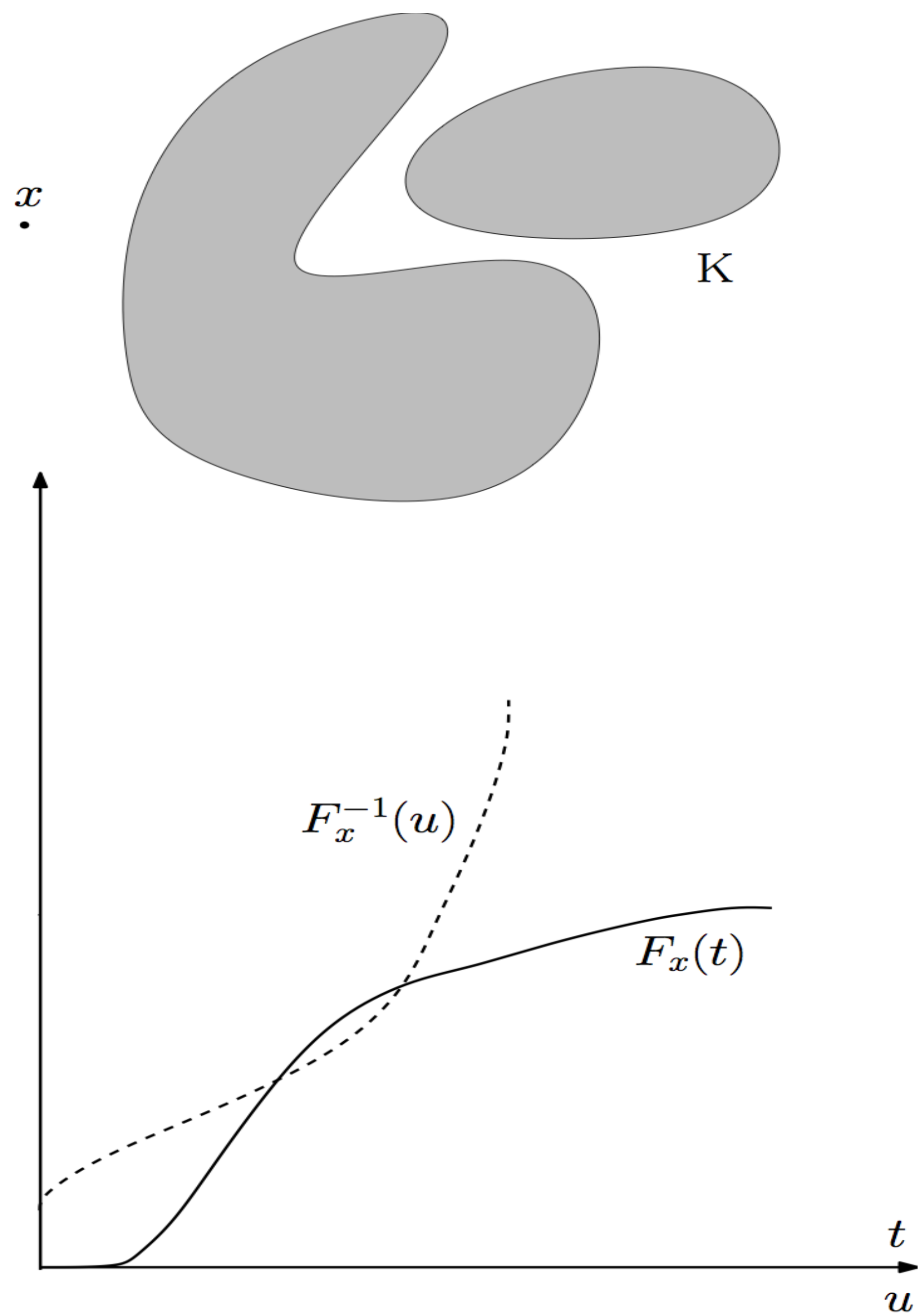
$$\mathbb{E}\left(\left|d^2_{P_n, \frac{k}{n}}(x) - d^2_{P, \frac{k}{n}}(x)\right|\right) \leq C\frac{n}{k}\frac{1}{\sqrt{n}}\sqrt{\frac{k}{n}}\omega_x\left(\frac{k}{n}\right)$$

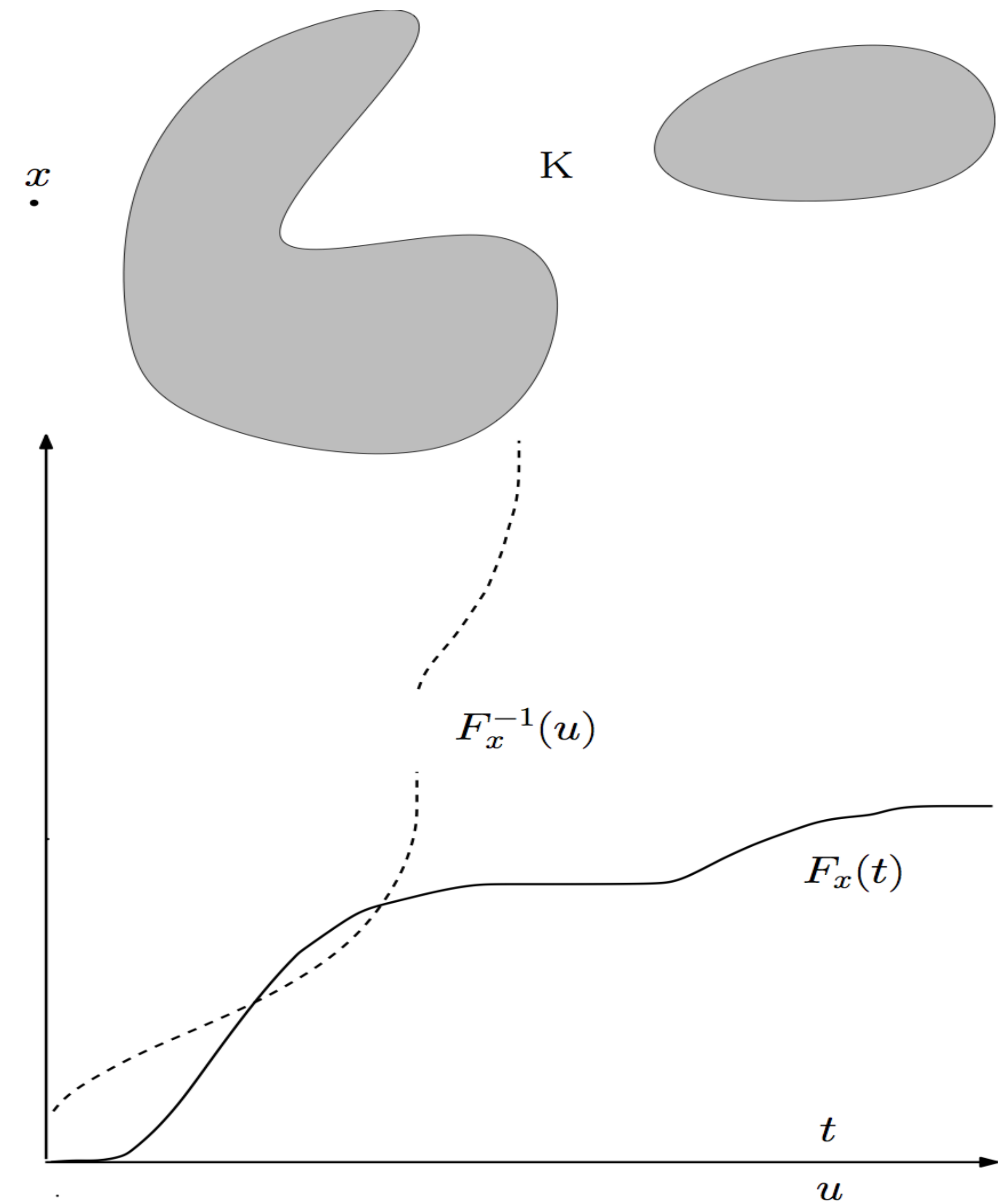
parametric rate of convergence

statistical complexity of the problem

Fluctuations of the DTEM [Chazal et al., 2015b]

The quantile function F_x^{-1} carries some geometric information.
For instance $\omega(0^+) = 0$ means that the support of dF_x is a closed interval.





Bootstrap and significance of topological features [Chazal et al., 2014b]

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 [Chazal et al., 2014b]

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 distance to measure function is Hadamard differentiable at P. Consequently, the bootstrap method for the DTM is asymptotically valid.

Bootstrap and significance of topological features [Chazal et al., 2014b]

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

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

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

Then,

Bootstrap estimate

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

Bootstrap and significance of topological features [Chazal et al., 2014b]

The Bottleneck Bootstrap

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

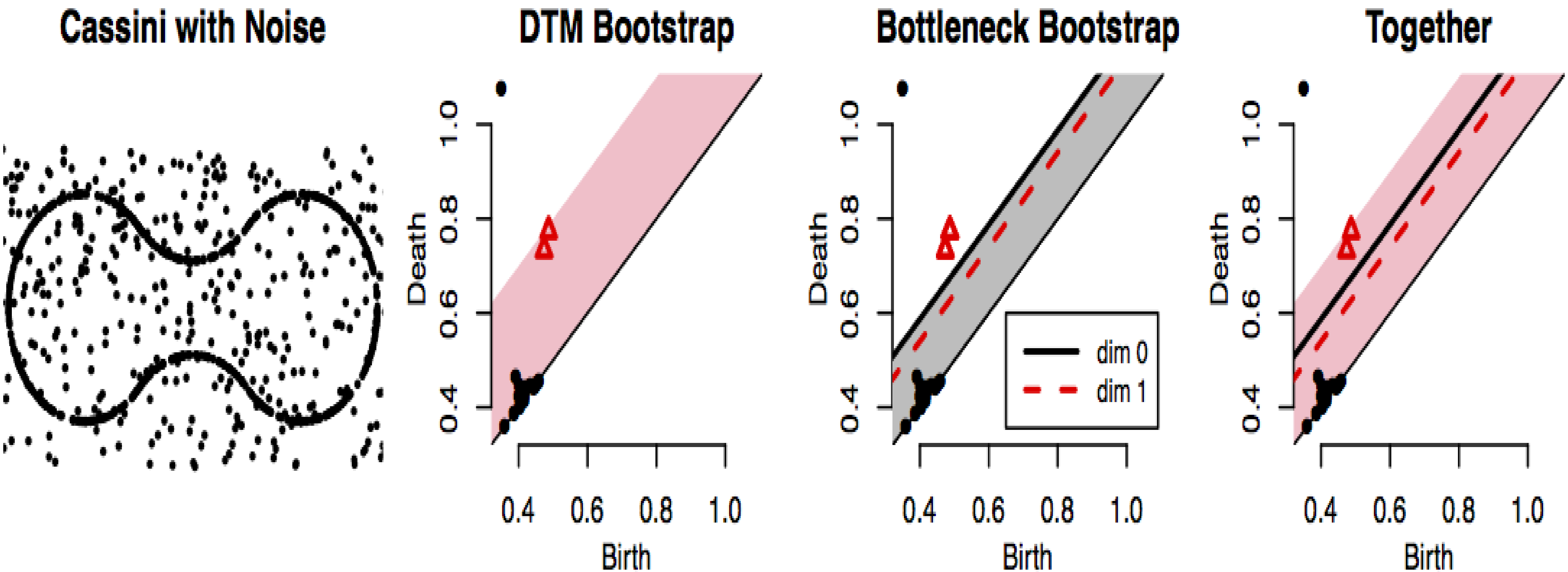
We directly bootstrap in the set of the persistence diagram by considering the random quantity $\mathrm{d_b}(\widehat{\mathrm{Dgm}}^*, \widehat{\mathrm{Dgm}})$. We define $\hat{t}_\alpha$ by

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

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

Bootstrap and significance of topological features [Chazal et al., 2014b]

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 following stability result is not sharp

$$\mathrm{d_b}(\widehat{\mathsf{Dgm}}, \mathsf{Dgm}) \leq \|d^2_{P,m} - d^2_{P_n,m}\|_\infty.$$

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 particular 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 ...)
- TDA methods need to bring together Geometric Inference, Computational Topology and Geometry, Statistics and Learning methods.

Thank you !

References

[Balakrishnan et al., 2012] Balakrishnan, S., Rinaldo, A., Sheehy, D., Singh, A., and Wasserman, L. A.. Minimax rates for homology inference. *Journal of Machine Learning Research - Proceedings Track*, 22:64–72.

[Bubenik, 2015] Bubenik, P. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102.

[Caillerie and Michel, 2011] Caillerie, C. and Michel, B. (2011). Model selection for simplicial approximation. *Foundations of Computational Mathematics*, 11(6):707–731.

[Carlsson, 2009] Carlsson, G. Topology and data. *AMS Bulletin*, 46(2):255–308.

[Chazal and Lieutier, 2007] Chazal, F. and Lieutier, A. Stability and computation of topological invariants of solids in {\ Bbb R}ˆ n. *Discrete & Computational Geometry*, 37(4):601–617.

[Chazal et al., 2012] Chazal, F., de Silva, V., Glisse, M., and Oudot, S. The structure and stability of persistence modules. *arXiv preprint arXiv:1207.3674*.

[Chazal et al., 2014a] Chazal, F., Glisse, M., Labruère, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysis. To appear in *Journal of Machine Learning Research*.

[Chazal et al., 2014b] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014a). Robust topological inference: Distance to a measure and kernel distance. *ArXiv preprint 1412.7197*.

References

[Chazal et al., 2015a] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology. To appear in *Proceedings of the 32 st International Conference on Machine Learning (ICML-15)*.

[Chazal et al., 2015b] Chazal, F., Massart, P., and Michel, B. (2015b). Rates of convergence for robust geometric inference. ArXiv preprint 1505.07602.

[Dereich et al., 2013] Dereich, S., Scheutzow, M., and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. H. Poincaré Probab. Statist.*, 49:1183–1203.

[Edelsbrunner et al., 2002] Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533.

[Federer, 1959] Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, pages 418–491.

[Fasy et al., 2014] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

[Fournier and Guillin, 2013] Fournier, N. and Guillin, A. (2013). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, pages 1–32.

[Genovese et al., 2012] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2012). Manifold estimation and singular deconvolution under hausdorff loss. *Ann. Statist.*, 40:941–963.

References

[Massart, 2007] Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume Lecture Notes in Mathematics 1896. Springer-Verlag.

[Niyogi et al., 2008] Niyogi, P., Smale, S., and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441.

[Singh et al., 2007] Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100. Citeseer.

[Singh et al., 2009] Singh, A., Scott, C., and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, 37(5B):2760–2782.

[Turner et al., 2014] Turner, K., Mileyko, Y., Mukherjee, S. and Harer, J. (2014) Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70 .

Topological invariants

How topological spaces can be compared from a topological point of view ?

For comparing topological spaces, we consider topological invariants (preserved by homeomorphism) : numbers, groups, polynomials.

Topological invariants

How topological spaces can be compared from a topological point of view ?

For comparing topological spaces, we consider topological invariants (preserved by homeomorphism) : numbers, groups, polynomials.

Homotopy is weaker than homeomorphism but is preserves many topological invariants.

- Two continous functions $f : X \to Y$ and $g : X \to Y$ are **homotopic** if there exists a continous application $H : X \times [0, 1] \to Y$ such that $H(\cdot, 0) = f$ and $H(\cdot, 1) = g$.
- Two topological spaces X and Y are **homotopic** if there exists two continous applications $f : X \to Y$ and $g : Y \to X$ such that
  - $g \circ f$ is homotopic to id_X;
  - $f \circ g$ is homotopic to id_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 $\mathbb{Y}$.

Topological Stability and Regularity

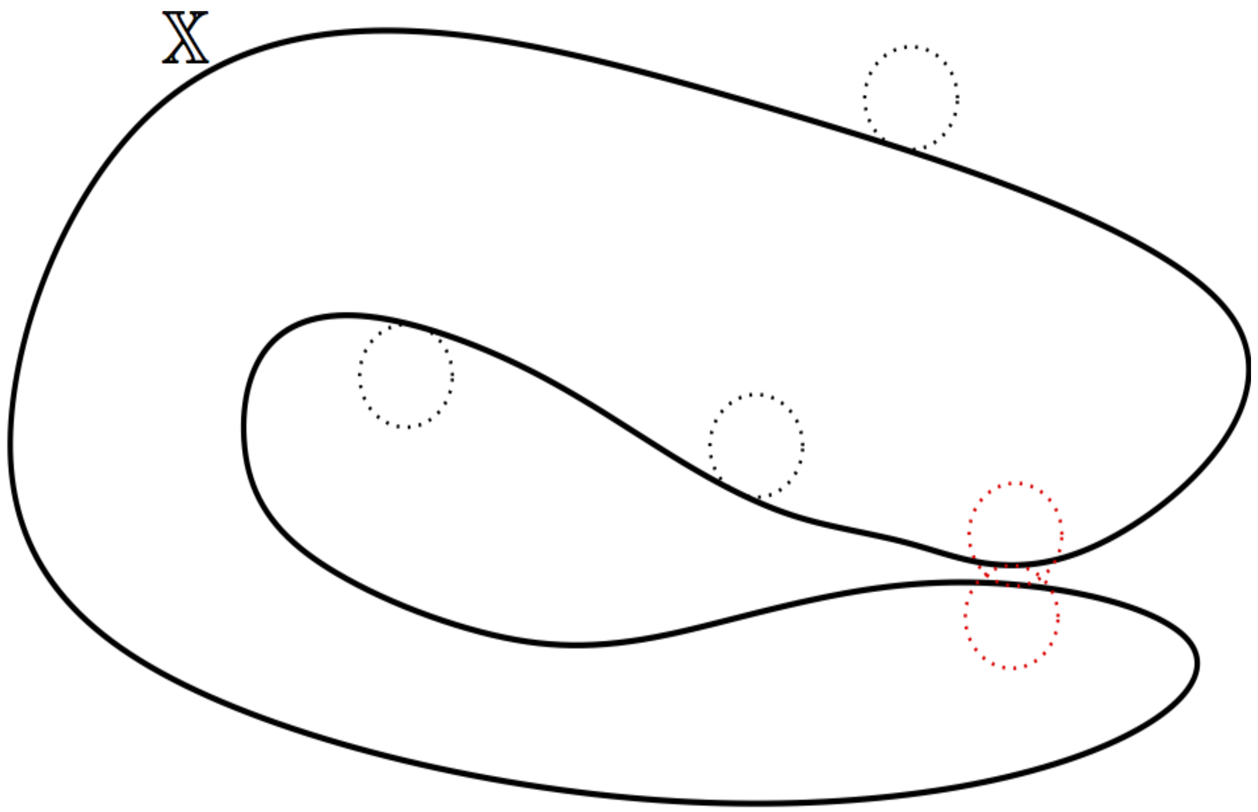
Topological inference : under "regularity assumptions", topological properties of $\mathbb{X}$ can be recovered from (the off-sets) of a close enough object $\mathbb{Y}$.

- 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 $\mathbb{Y}$.

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

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 $d_H(\mathbb{X}, \mathbb{Y}) < \varepsilon$, $\mathrm{wfs}(\mathbb{X}) > 2\varepsilon$ 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 $\mathbb{Y}$.

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

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.

Sampling conditions in Hausdorff metric.

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