IHES, Jan. 19, 2017

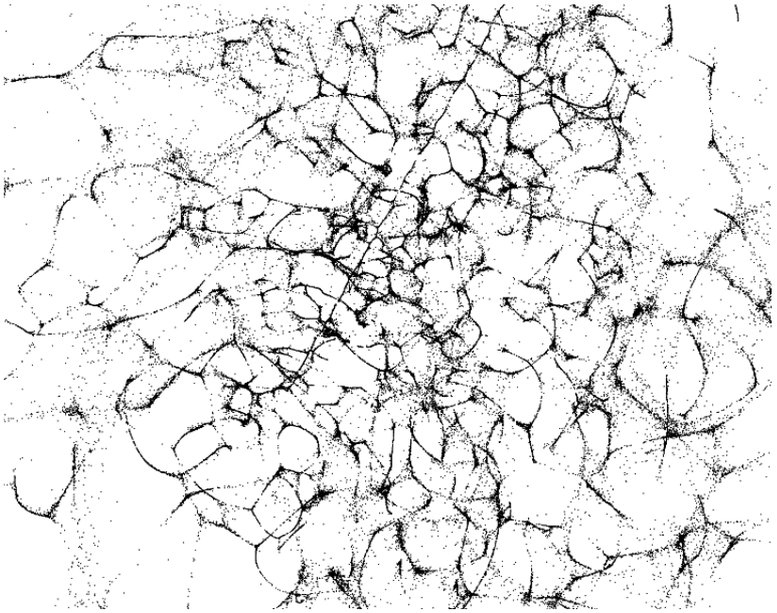
Statistics/Learning at Paris-Saclay

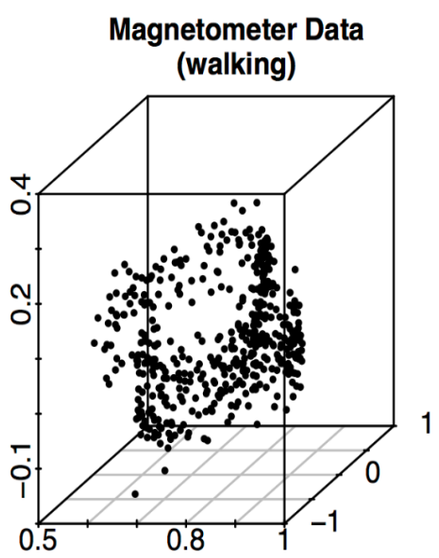
Persistent homology for data: stability properties and statistical aspects

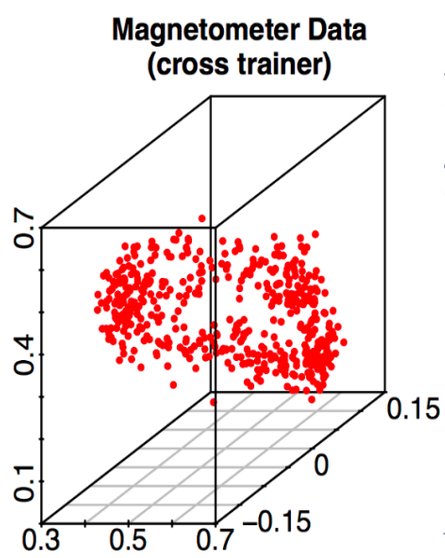
Frédéric Chazal
DataShape group
INRIA Saclay - Ile-de-France
frederic.chazal@inria.fr

Introduction

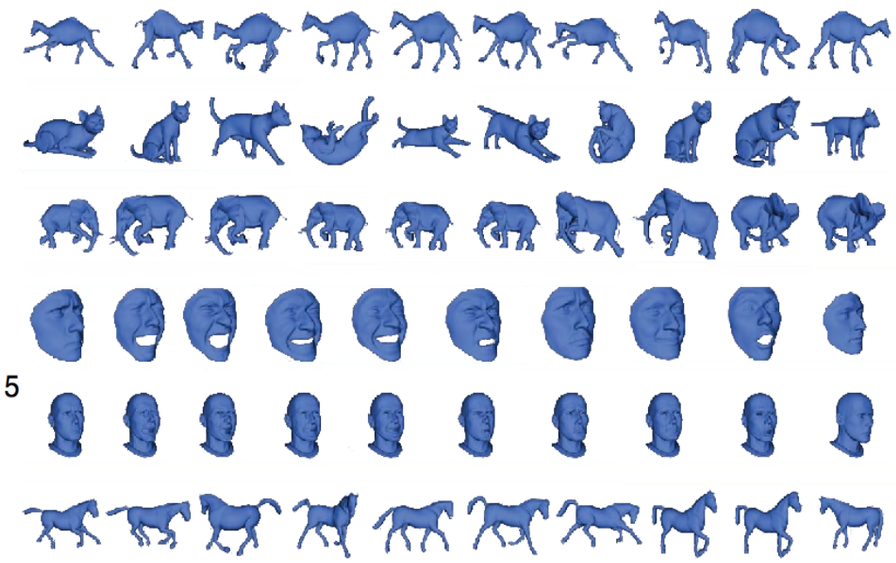
[distribution of galaxies]







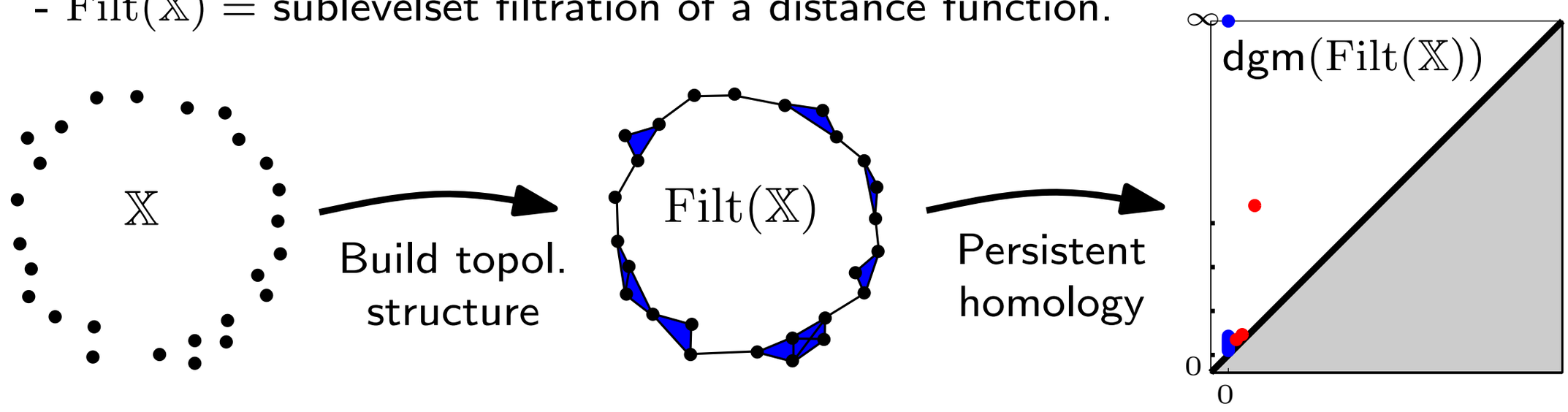
[Non-rigid shape database]



- Data often come as (sampling of) metric spaces or sets/spaces endowed with a similarity measure with, possibly complex, topological/geometric structure.
- Topological Data Analysis (TDA):
  - infer relevant topological and geometric features of these spaces.
  - take advantage of topol./geom. information for further processing of data (classification, recognition, learning, clustering,...).

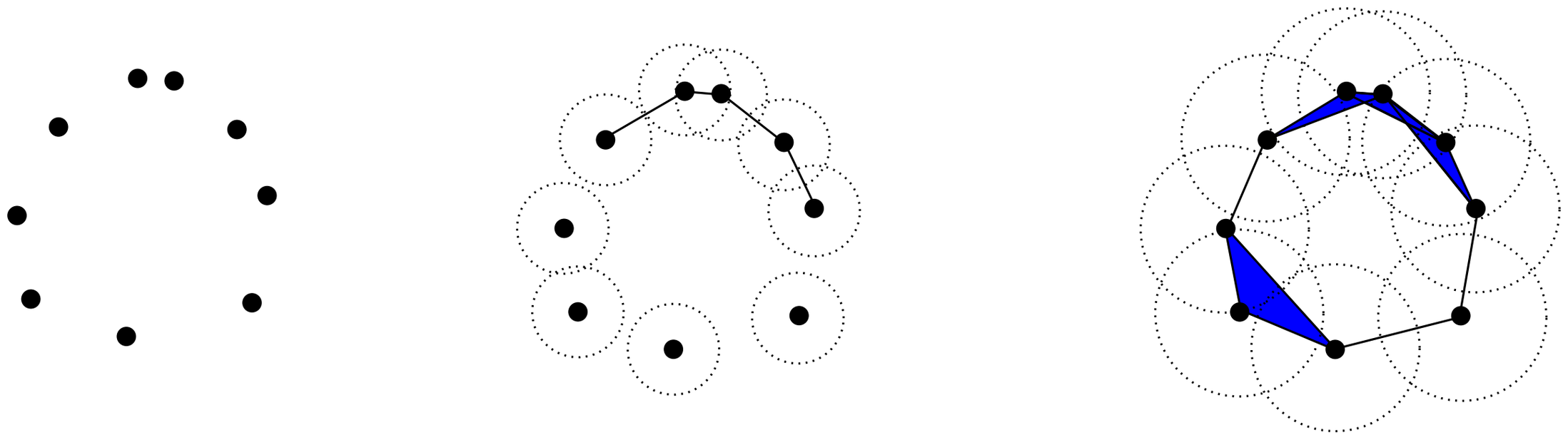
Persistent homology for (point cloud) data

Examples:
- $\mathrm{Filt}(\mathbb{X}) = \mathrm{Rips}_\alpha(\mathbb{X})$
- $\mathrm{Filt}(\mathbb{X}) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\mathbb{X})$
- $\mathrm{Filt}(\mathbb{X}) =$ sublevelset filtration of a distance function.



- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures (connections with stability properties).

Filtered simplicial complexes



A multiscale topological structure on top of the data:

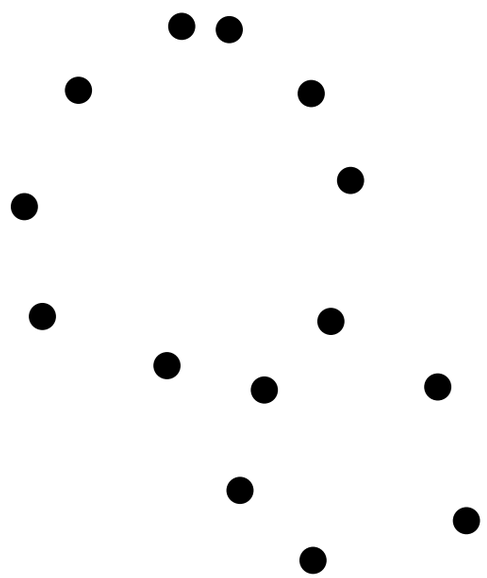
a filtered simplicial complex $\mathbb{S}$ built on top of a set $\mathbb{X}$ is a family $(\mathbb{S}_a \mid a \in \mathbf{R})$ of subcomplexes of some fixed simplicial complex $\overline{\mathbb{S}}$ with vertex set $\mathbb{X}$ s. t. $\mathbb{S}_a \subseteq \mathbb{S}_b$ for any $a \leq b$.

Example: Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space.

- The Vietoris-Rips filtration is the filtered simplicial complexe defined by: for $a \in \mathbf{R}$,

$$[x_0, x_1, \cdots, x_k] \in \mathrm{Rips}(X, a) \Leftrightarrow d_X(x_i, x_j) \leq a, \quad \text{for all } i, j.$$

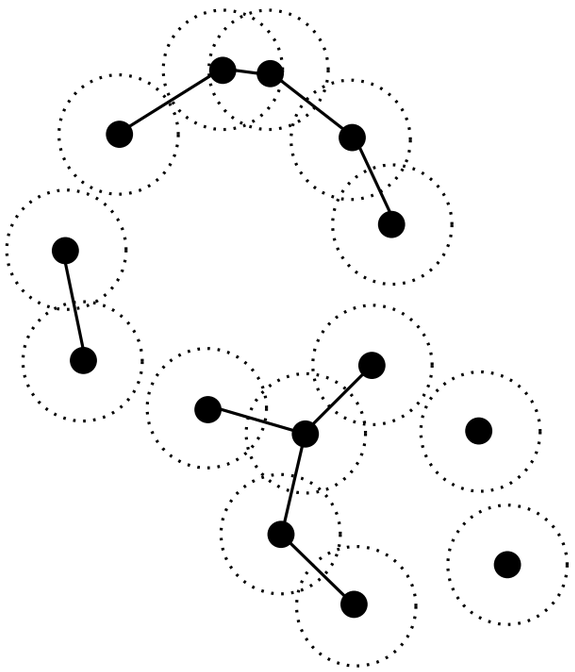
Persistent homology for (point cloud) data



$\widehat{\mathbb{X}}_m$: metric data set

- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

Persistent homology for (point cloud) data

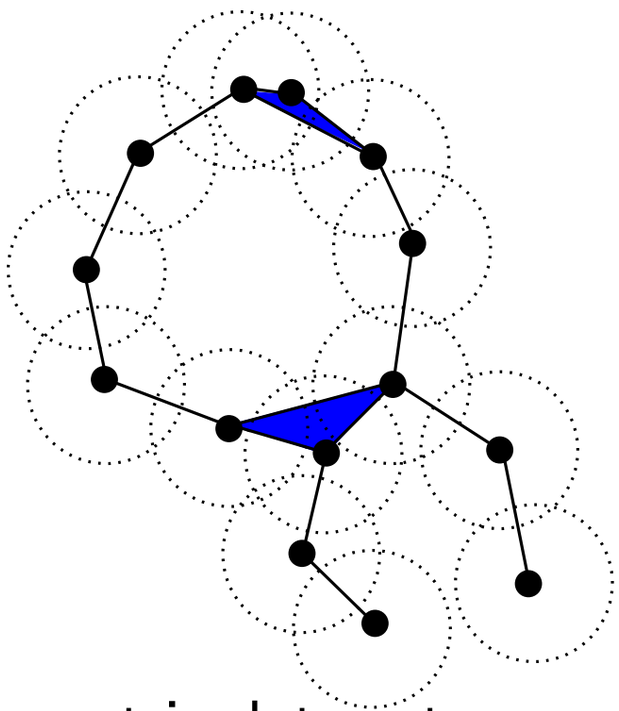


$\widehat{\mathbb{X}}_m$: metric data set

↳ $\mathrm{Filt}(\widehat{\mathbb{X}}_m)$: filtered simplicial complex

- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

Persistent homology for (point cloud) data

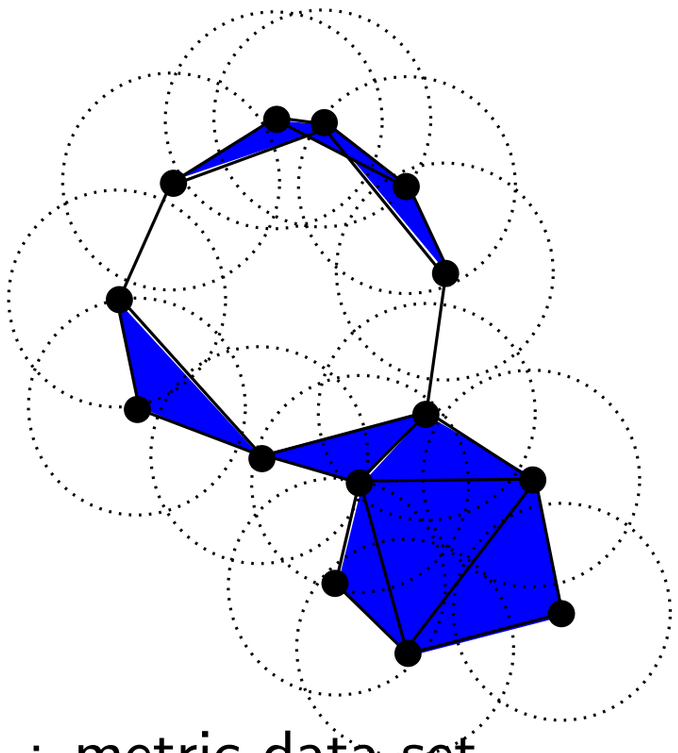


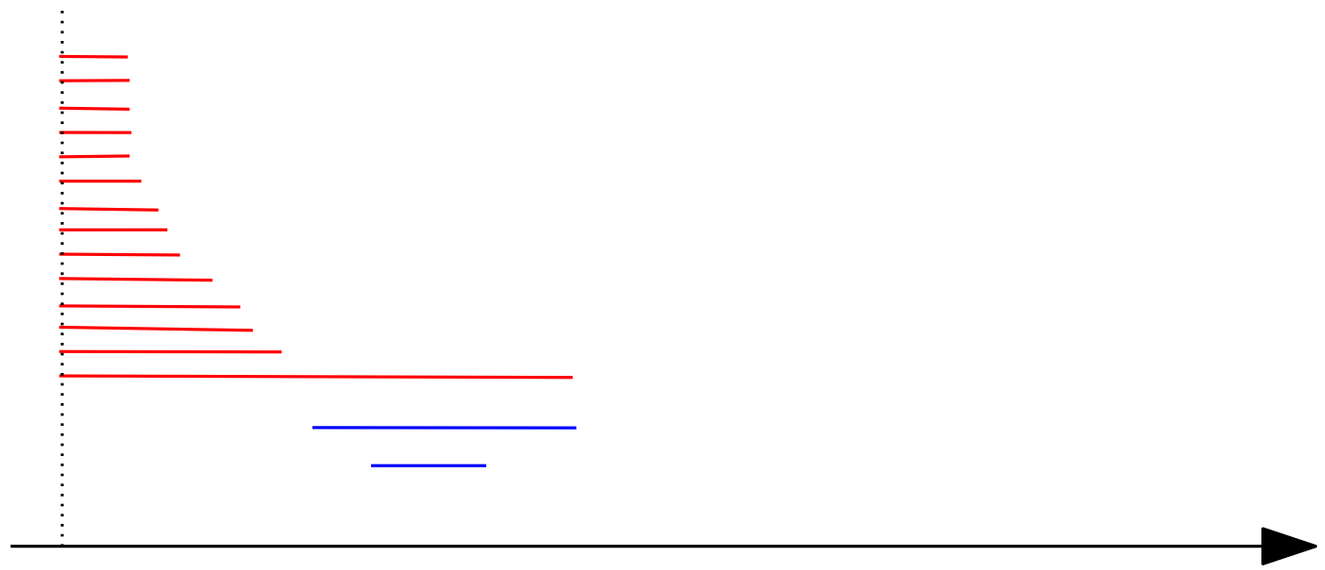
$\widehat{\mathbb{X}}_m$: metric data set

↳ $\mathrm{Filt}(\widehat{\mathbb{X}}_m)$: filtered simplicial complex

- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

Persistent homology for (point cloud) data



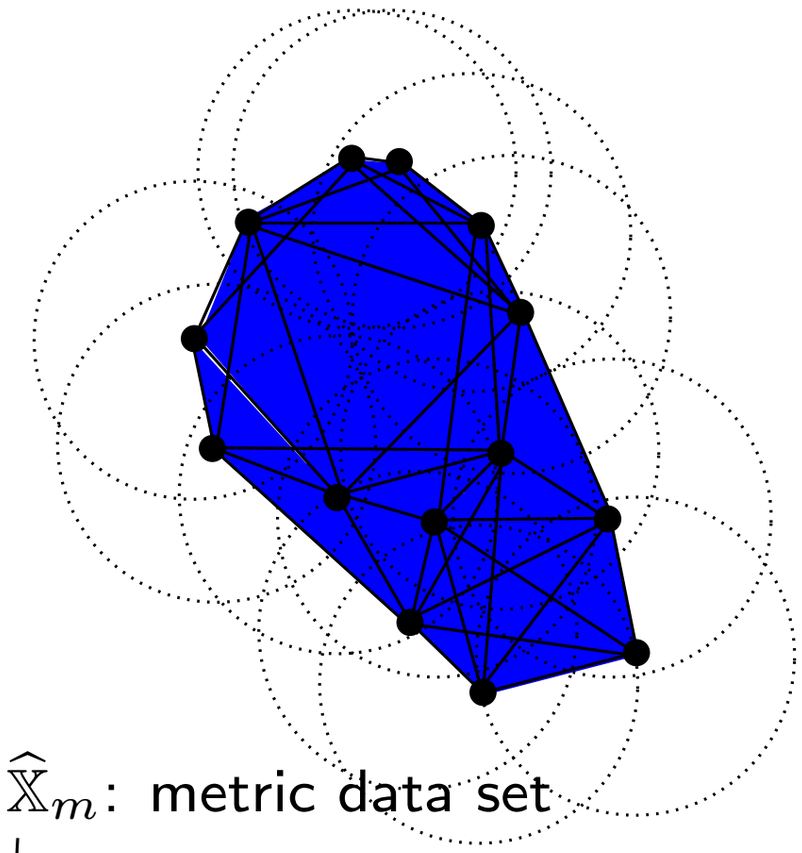


$\widehat{\mathbb{X}}_m$: metric data set

↳ $\mathrm{Filt}(\widehat{\mathbb{X}}_m)$: filtered simplicial complex

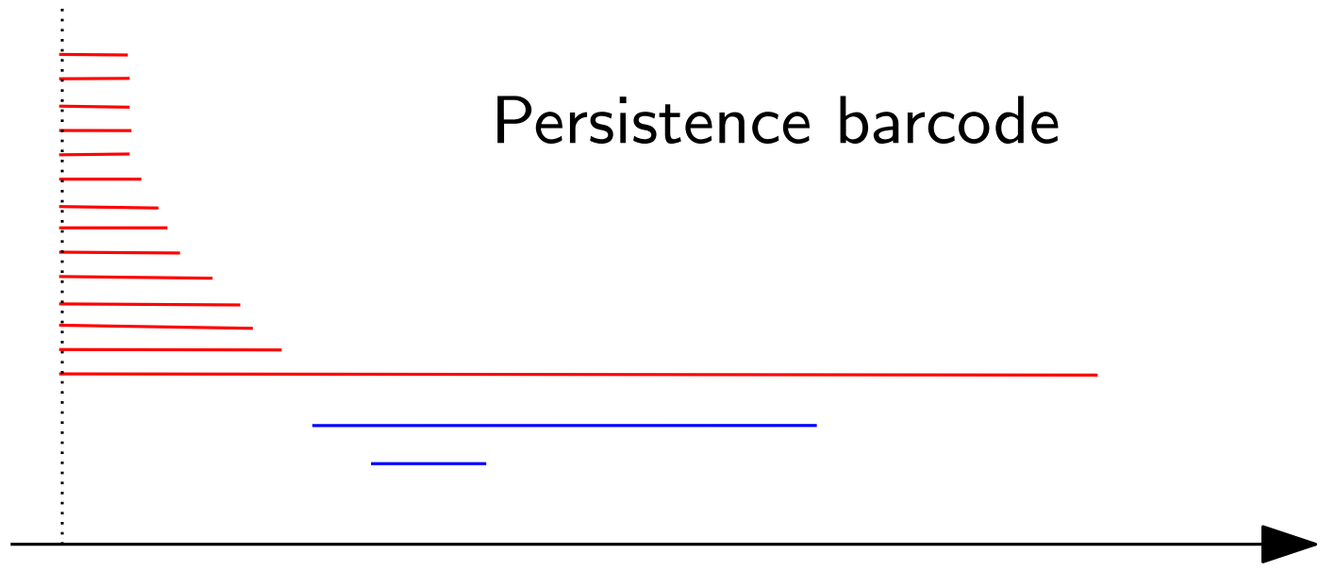
- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

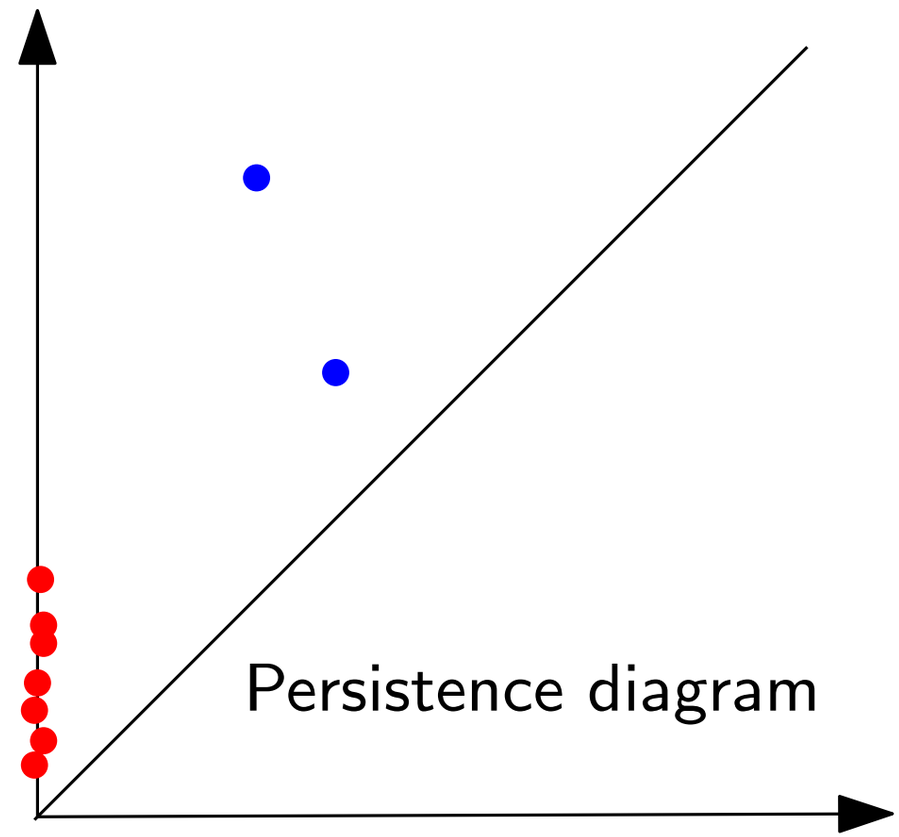
Persistent homology for (point cloud) data



$\widehat{\mathbb{X}}_m$: metric data set

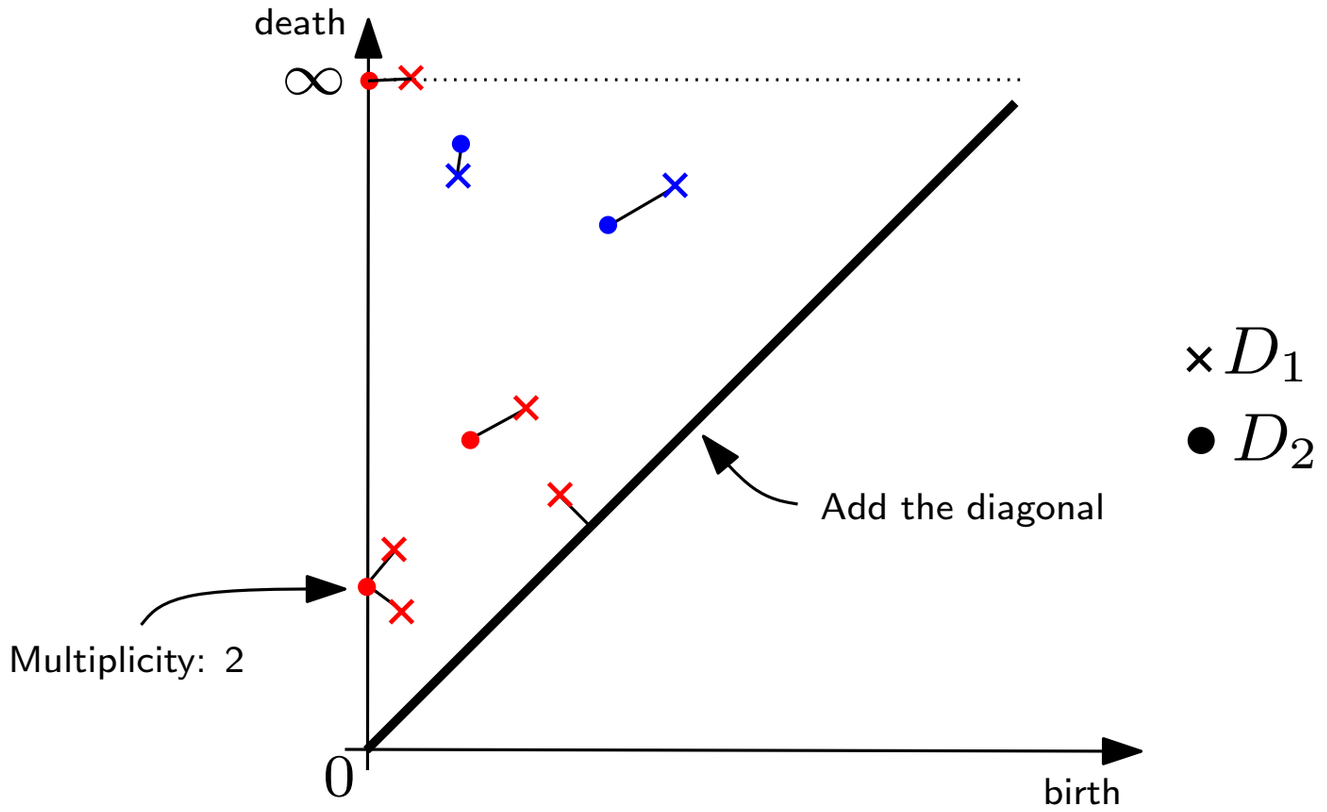
$\mathrm{Filt}(\widehat{\mathbb{X}}_m)$: filtered simplicial complex





- Build a geometric filtered simplicial complex on top of $\widehat{\mathbb{X}}_m \rightarrow$ multiscale topol. structure.
- Compute the persistent homology of the complex $\rightarrow$ multiscale topol. signature.
- Compare the signatures of "close" data sets $\rightarrow$ robustness and stability results.
- Statistical properties of signatures

Distance between persistence diagrams



The bottleneck distance between two diagrams D_1 and D_2 is

$$d_B(D_1, D_2) = \inf_{\gamma \in \Gamma} \sup_{p \in D_1} \|p - \gamma(p)\|_\infty$$

where Γ is the set of all the bijections between D_1 and D_2 and $\|p - q\|_\infty = \max(|x_p - x_q|, |y_p - y_q|)$.

$\rightarrow$ Persistence diagrams provide easy to compare topological signatures.

Stability properties

"Stability theorem": Close spaces/data sets have close persistence diagrams!

[C., de Silva, Oudot - Geom. Dedicata 2013].

If $\mathbb{X}$ and $\mathbb{Y}$ are pre-compact metric spaces, then

$$\mathrm{d_b}(\mathsf{dgm}(\mathrm{Rips}(\mathbb{X})), \mathsf{dgm}(\mathrm{Rips}(\mathbb{Y}))) \leq 2d_{GH}(\mathbb{X}, \mathbb{Y}).$$

Bottleneck distance

Gromov-Hausdorff distance

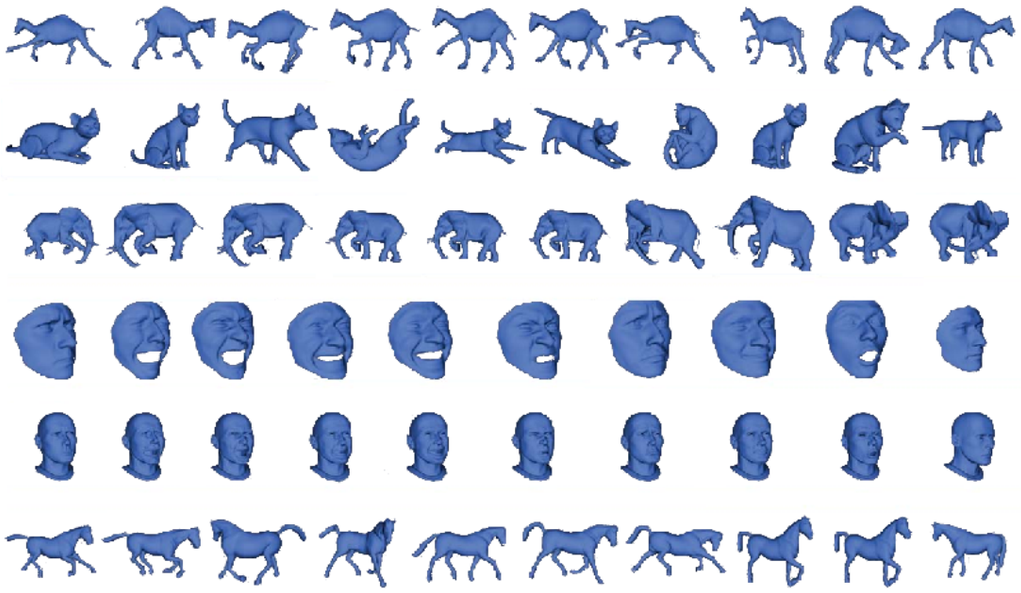
$$d_{GH}(\mathbb{X}, \mathbb{Y}) := \inf_{\mathbb{Z}, \gamma_1, \gamma_2} d_H(\gamma_1(\mathbb{X}), \gamma_2(\mathbb{X}))$$

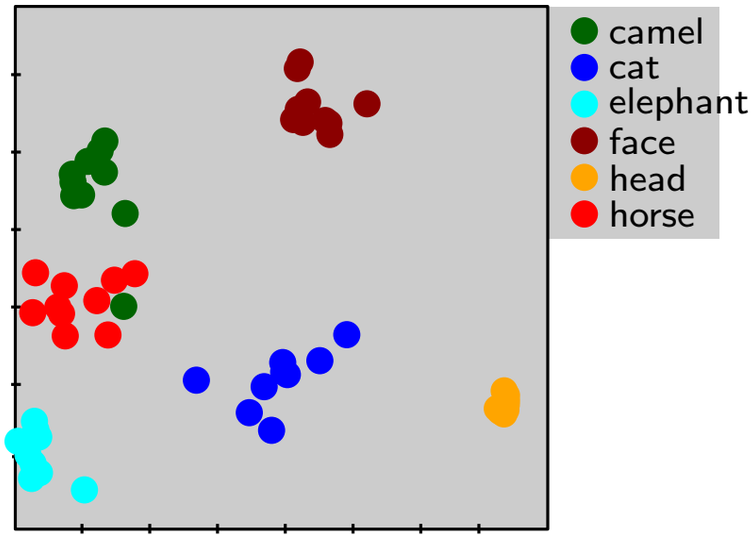
$\mathbb{Z}$ metric space, $\gamma_1 : \mathbb{X} \to \mathbb{Z}$ and $\gamma_2 : \mathbb{Y} \to \mathbb{Z}$ isometric embeddings.

Rem: This result also holds for other families of filtrations (particular case of a more general theorem).

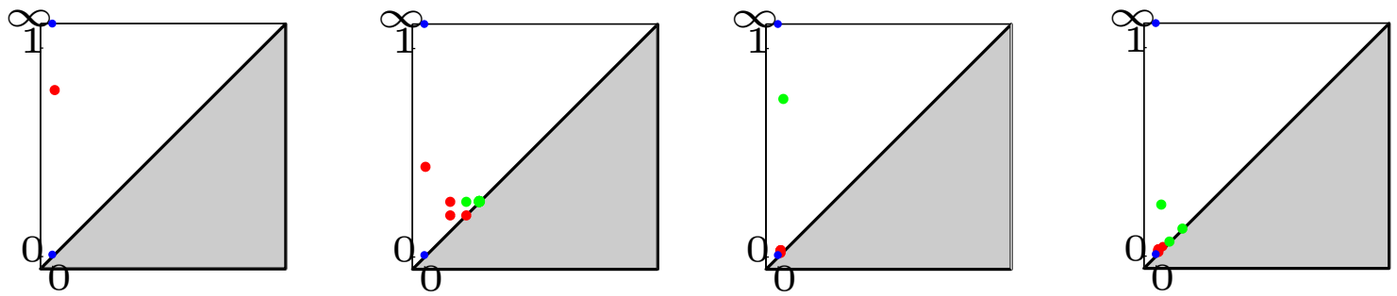
Illustration: non rigid shape classification

[C., Cohen-Steiner, Guibas, Mémoli, Oudot - SGP '09]





MDS using bottleneck distance.



- Non rigid shapes in a same class are almost isometric, but computing Gromov-Hausdorff distance between shapes is extremely expensive.
- Compare diagrams of sampled shapes instead of shapes themselves.

Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Examples:

- Let $\mathbb{S}$ be a filtered simplicial complex. If $V_a = \mathrm{H}(\mathbb{S}_a)$ and $v_a^b : \mathrm{H}(\mathbb{S}_a) \to \mathrm{H}(\mathbb{S}_b)$ is the linear map induced by the inclusion $\mathbb{S}_a \hookrightarrow \mathbb{S}_b$ then $(\mathrm{H}(\mathbb{S}_a) \mid a \in \mathbf{R})$ is a persistence module.
- Given a metric space $(\mathbb{X}, d_{\mathbb{X}})$, $\mathrm{H}(\mathrm{Rips}(\mathbb{X}))$ is a persistence module.
- If $f : X \to \mathbf{R}$ is a function, then the filtration defined by the sublevel sets of f, $\mathbb{F}_a = f^{-1}((-\infty, a])$, induces a persistence module at homology level.

Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Definition: A persistence module $\mathbb{V}$ is q-tame if for any $a < b$, v_a^b has a finite rank.

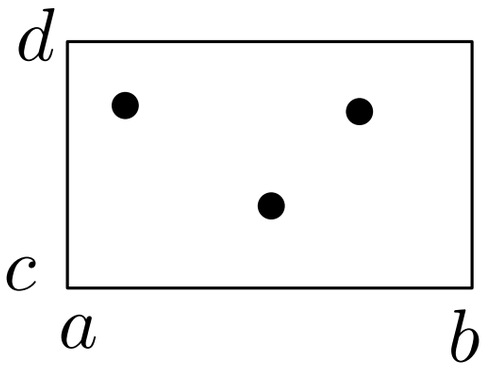
Theorem: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG'09], [C., de Silva, Glisse, Oudot 12]

q-tame persistence modules have well-defined persistence diagrams.

Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

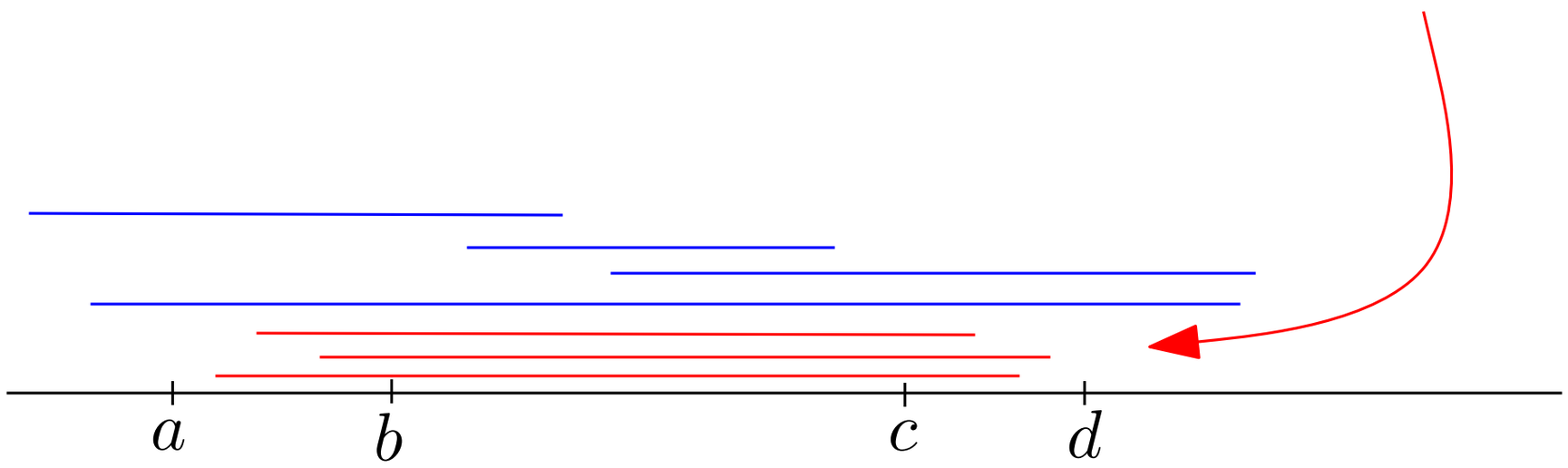
An idea about the definition of persistence diagrams:



Measures on rectangles:

Number of points in any rectangle $[a, b] \times [c, d]$ above the diagonal:

$$rk(v_b^c) - rk(v_b^d) + rk(v_a^d) - rk(v_a^c)$$



Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Definition: A persistence module $\mathbb{V}$ is q-tame if for any $a < b$, v_a^b has a finite rank.

Theorem: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG'09], [C., de Silva, Glisse, Oudot 12]

q-tame persistence modules have well-defined persistence diagrams.

Exercise: Let $\mathbb{X}$ be a precompact metric space. Then $\mathrm{H}(\mathrm{Rips}(\mathbb{X}))$ is q-tame.

Recall that a metric space $(\mathbb{X}, \rho)$ is precompact if for any $\epsilon > 0$ there exists a finite subset $F_\epsilon \subset \mathbb{X}$ such that $d_H(\mathbb{X}, F_\epsilon) < \epsilon$ (i.e. $\forall x \in X, \exists p \in F_\epsilon$ s.t. $\rho(x, p) < \epsilon$).

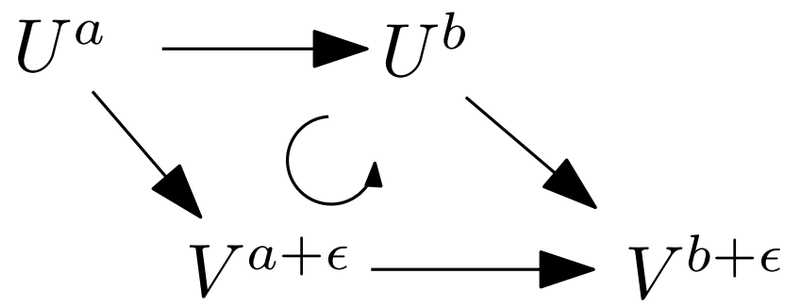
Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

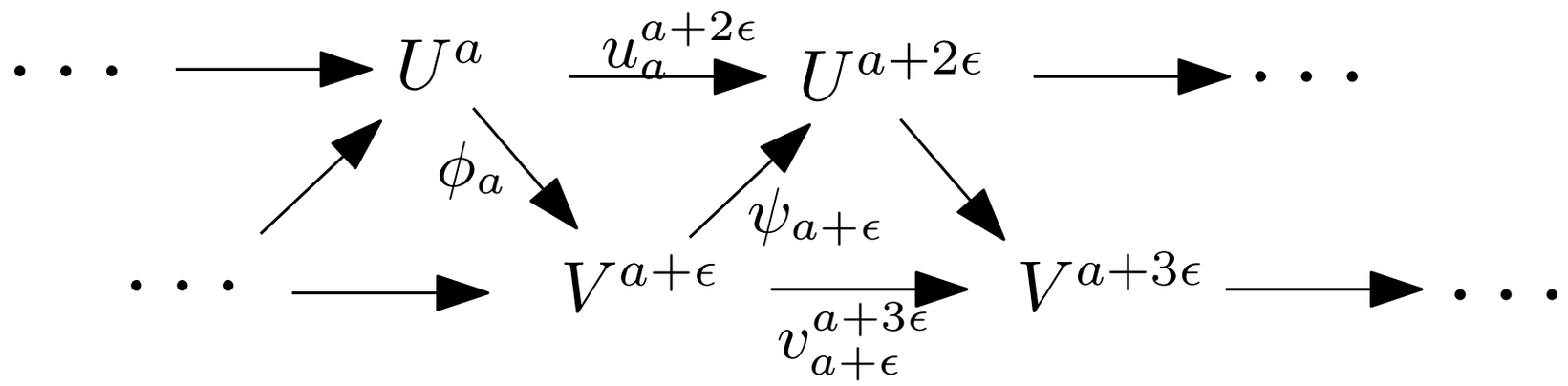
A homomorphism of degree ϵ between two persistence modules $\mathbb{U}$ and $\mathbb{V}$ is a collection Φ of linear maps

$$(\phi_a : U_a \to V_{a+\epsilon} \mid a \in \mathbf{R})$$

such that $v_{a+\epsilon}^{b+\epsilon} \circ \phi_a = \phi_b \circ u_a^b$ for all $a \leq b$.



An ε-interleaving between $\mathbb{U}$ and $\mathbb{V}$ is specified by two homomorphisms of degree ϵ $\Phi : \mathbb{U} \to \mathbb{V}$ and $\Psi : \mathbb{V} \to \mathbb{U}$ s.t. $\Phi \circ \Psi$ and $\Psi \circ \Phi$ are the "shifts" of degree 2ϵ between $\mathbb{U}$ and $\mathbb{V}$.



Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Stability Thm: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG '09], [C., de Silva, Glisse Oudot 12]

If $\mathbb{U}$ and $\mathbb{V}$ are q-tame and ϵ-interleaved for some $\epsilon \geq 0$ then

$$d_B(\mathsf{dgm}(\mathbb{U}), \mathsf{dgm}(\mathbb{V})) \leq \epsilon$$

Where do stability results come from?

Definition: A persistence module $\mathbb{V}$ is an indexed family of vector spaces $(V_a \mid a \in \mathbf{R})$ and a doubly-indexed family of linear maps $(v_a^b : V_a \to V_b \mid a \leq b)$ which satisfy the composition law $v_b^c \circ v_a^b = v_a^c$ whenever $a \leq b \leq c$, and where v_a^a is the identity map on V_a.

Stability Thm: [C., Cohen-Steiner, Glisse, Guibas, Oudot - SoCG '09], [C., de Silva, Glisse Oudot 12]

If $\mathbb{U}$ and $\mathbb{V}$ are q-tame and ϵ-interleaved for some $\epsilon \geq 0$ then

$$d_B(\mathsf{dgm}(\mathbb{U}), \mathsf{dgm}(\mathbb{V})) \leq \epsilon$$

Strategy: build filtrations that induce **q-tame** homology persistence modules and that turn out to be ϵ-**interleaved** when the considered spaces/functions are $O(\epsilon)$-*close*.

Some weaknesses

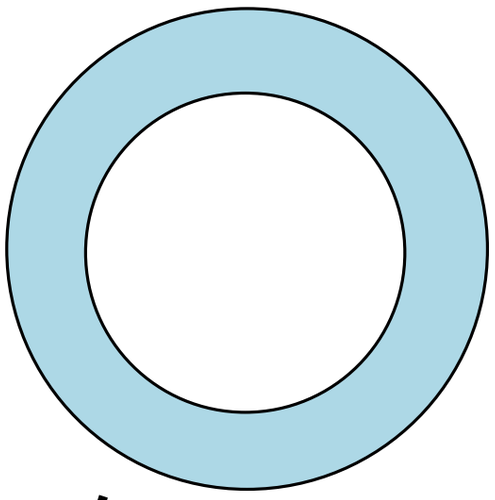
If $\mathbb{X}$ and $\mathbb{Y}$ are pre-compact metric spaces, then

$$\mathrm{d_b}(\mathsf{dgm}(\mathrm{Rips}(\mathbb{X})), \mathsf{dgm}(\mathrm{Rips}(\mathbb{Y}))) \leq d_{GH}(\mathbb{X}, \mathbb{Y}).$$

$\rightarrow$ Vietoris-Rips (or Cech, witness) filtrations quickly become prohibitively large as the size of the data increases ($\mathsf{O}(|\mathbb{X}|^d)$), making the computation of persistence practically almost impossible.

$\rightarrow$ Persistence diagrams of Rips-Vietoris (and Cěch, witness,..) filtrations and Gromov-Hausdorff distance are very sensitive to noise and outliers.

Statistical setting



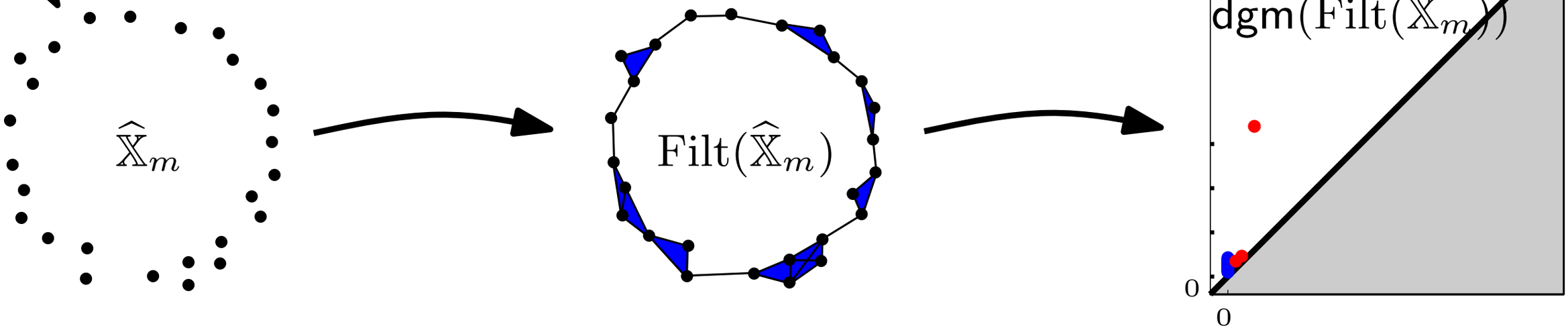
$(\mathbb{M}, \rho)$ metric space

μ a probability measure with compact support $\mathbb{X}_\mu$.

Examples:

- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \mathrm{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) =$ sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.

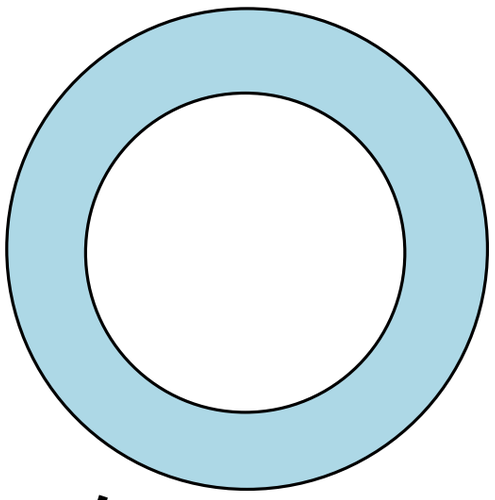
Sample m points according to μ.



Questions:

- Statistical properties of $\mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))$? $\mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m)) \to$? as $m \to +\infty$?

Statistical setting



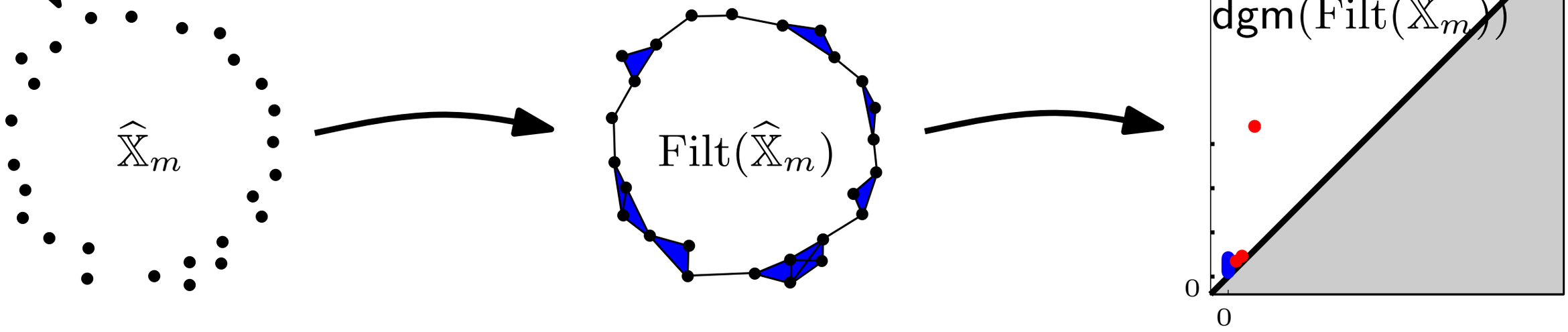
$(\mathbb{M}, \rho)$ metric space

μ a probability measure with compact support $\mathbb{X}_\mu$.

Examples:

- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \mathrm{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) =$ sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.

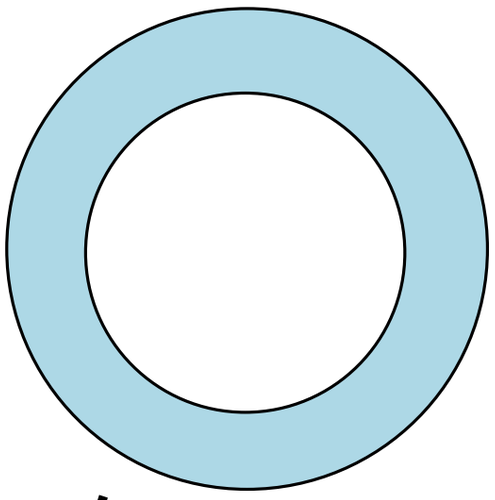
Sample m points according to μ.



Questions:

- Statistical properties of $\mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))$? $\mathrm{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m)) \to ?$ as $m \to +\infty$?
- Can we do more statistics with persistence diagrams?

Statistical setting



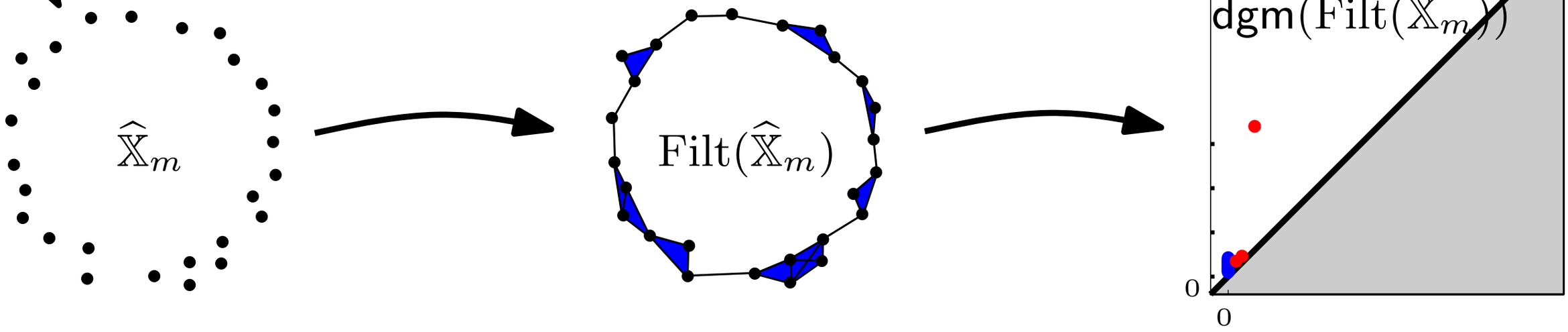
$(\mathbb{M}, \rho)$ metric space

μ a probability measure with compact support $\mathbb{X}_\mu$.

Examples:

- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \mathrm{Rips}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) = \check{\mathrm{C}}\mathrm{ech}_\alpha(\widehat{\mathbb{X}}_m)$
- $\mathrm{Filt}(\widehat{\mathbb{X}}_m) =$ sublevelset filtration of $\rho(., \mathbb{X}_\mu)$.

Sample m points according to μ.



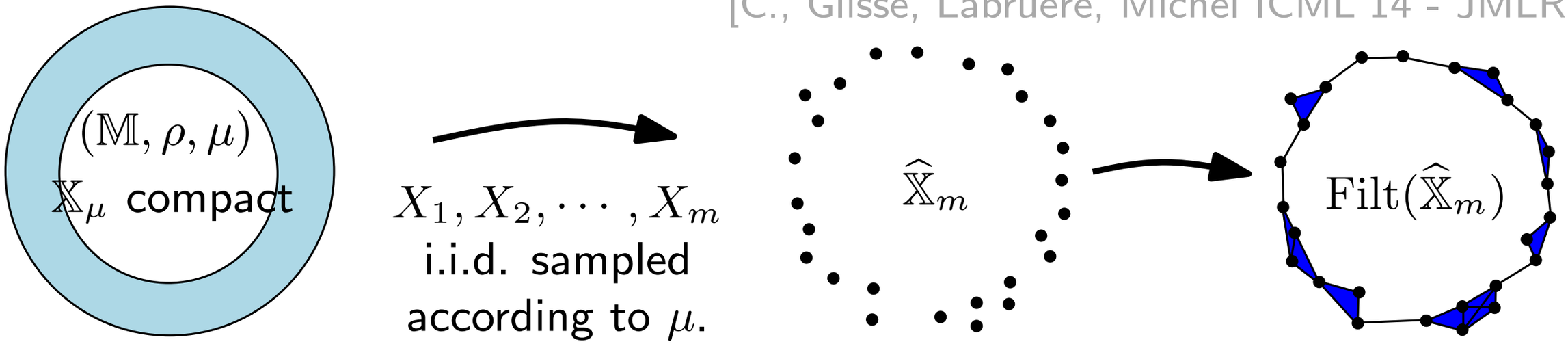
Stability thm: $\mathrm{d_b}(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))) \leq 2 d_{GH}(\mathbb{X}_\mu, \widehat{\mathbb{X}}_m)$

So, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))\right) > \varepsilon\right) \leq \mathbb{P}\left(d_{GH}(\mathbb{X}_\mu, \widehat{\mathbb{X}}_m) > \frac{\varepsilon}{2}\right)$$

Deviation inequality

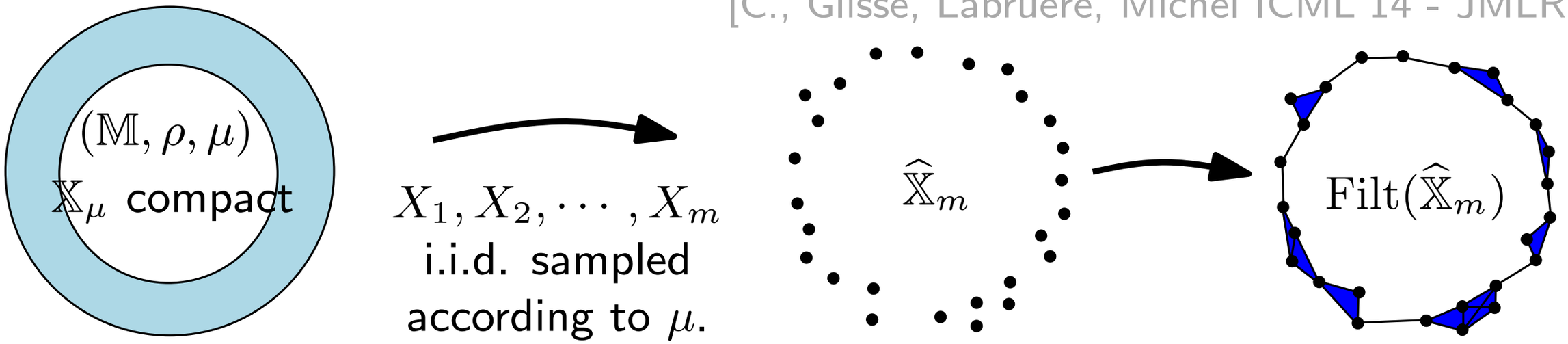
[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]



For $a, b > 0$, μ satisfies the (a, b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

Deviation inequality

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]



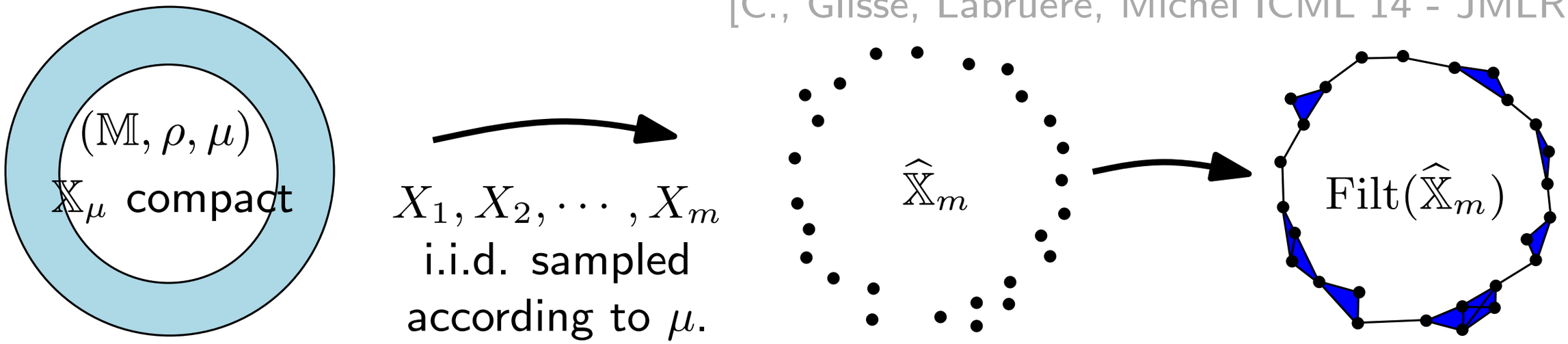
For $a, b > 0$, μ satisfies the (a, b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

Theorem: If μ satisfies the (a, b)-standard assumption, then for any $\varepsilon > 0$:

$$\mathbb{P}\left(\mathrm{d_b}\left(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}}_m))\right) > \varepsilon\right) \leq \min\left(\frac{8^b}{a\varepsilon^b} \exp(-ma\varepsilon^b), 1\right).$$

Deviation inequality

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]



For $a, b > 0$, μ satisfies the (a, b)-standard assumption if for any $x \in \mathbb{X}_\mu$ and any $r > 0$, we have $\mu(B(x, r)) \geq \min(ar^b, 1)$.

Sketch of proof:

1. Upperbound $\mathbb{P}\left(d_H(\mathbb{X}_\mu, \widehat{\mathbb{X}}_m) > \frac{\varepsilon}{2}\right)$.
2. (a, b) standard assumption $\Rightarrow$ an explicit upperbound for the covering number of $\mathbb{X}_\mu$ (by balls of radius $\varepsilon/2$).
3. Apply "union bound" argument.

$$C(\varepsilon) \leq P(\varepsilon/2) + \mu(B(x, \varepsilon/2)) \geq a(\varepsilon/2)^b$$

Minimax rate of convergence

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]

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

Minimax rate of convergence

[C., Glisse, Labruère, Michel ICML'14 - JMLR'15]

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

Theorem: Let $\mathcal{P}(a, b, \mathbb{M})$ be the set of (a, b)-standard proba measures on $\mathbb{M}$. Then:

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

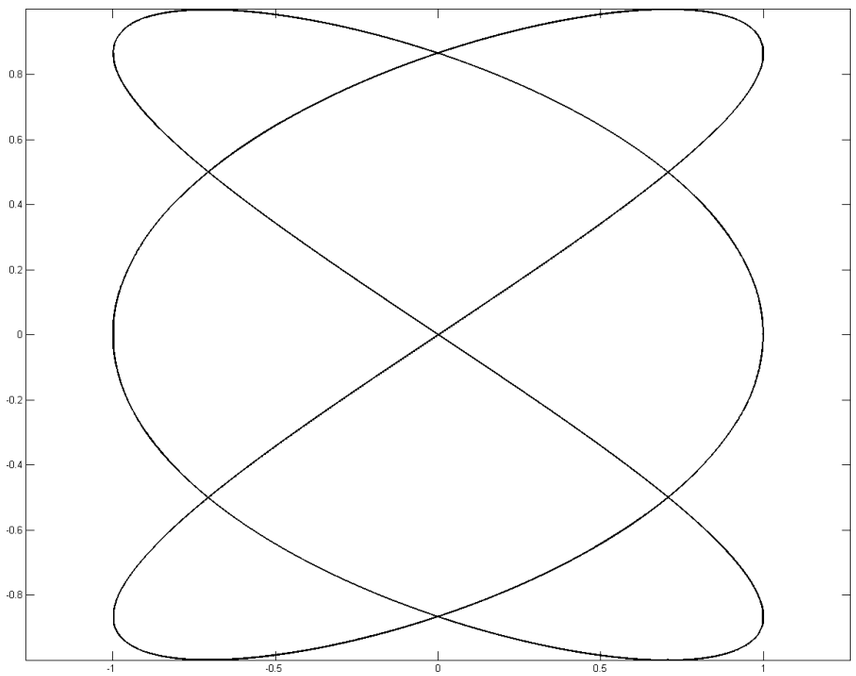
where the constant C only depends on a and b (not on $\mathbb{M}$!). Assume moreover that there exists a non isolated point x in $\mathbb{M}$ and let x_m be a sequence in $\mathbb{M} \setminus \{x\}$ such that $\rho(x, x_m) \leq (am)^{-1/b}$. Then for any estimator $\widehat{\mathsf{dgm}}_m$ of $\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu))$:

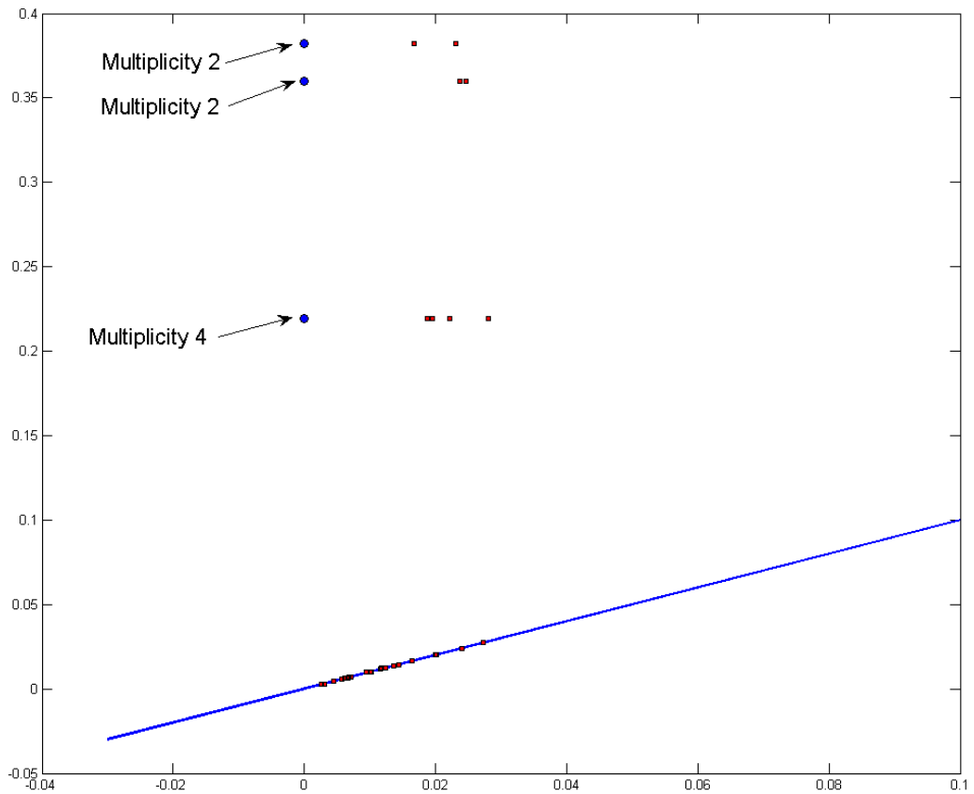
$$\liminf_{m \to \infty} \rho(x, x_m)^{-1} \sup_{\mu \in \mathcal{P}(a,b,\mathbb{M})} \mathbb{E}\left[\mathrm{d_b}(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \widehat{\mathsf{dgm}}_m)\right] \geq C'$$

where C' is an absolute constant.

Remark: we can obtain slightly better bounds if $\mathbb{X}_\mu$ is a submanifold of $\mathbb{R}^D$ - see [Genovese, Perone-Pacifico,Verdinelli, Wasserman 2011, 2012]

Numerical illustrations

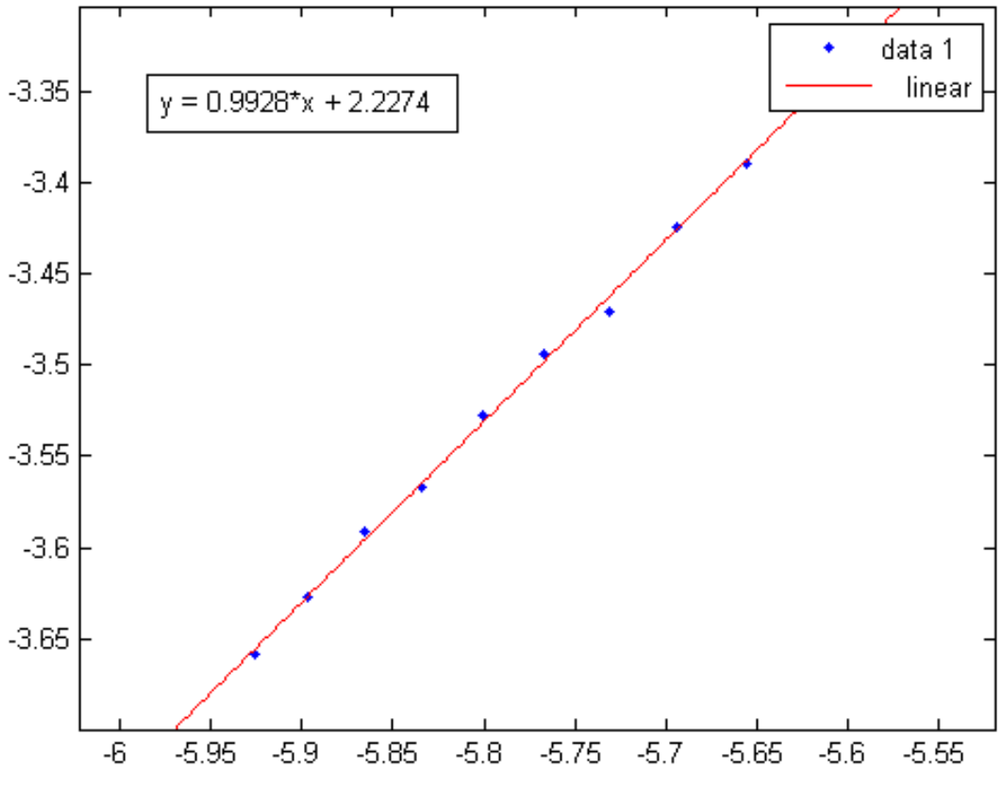




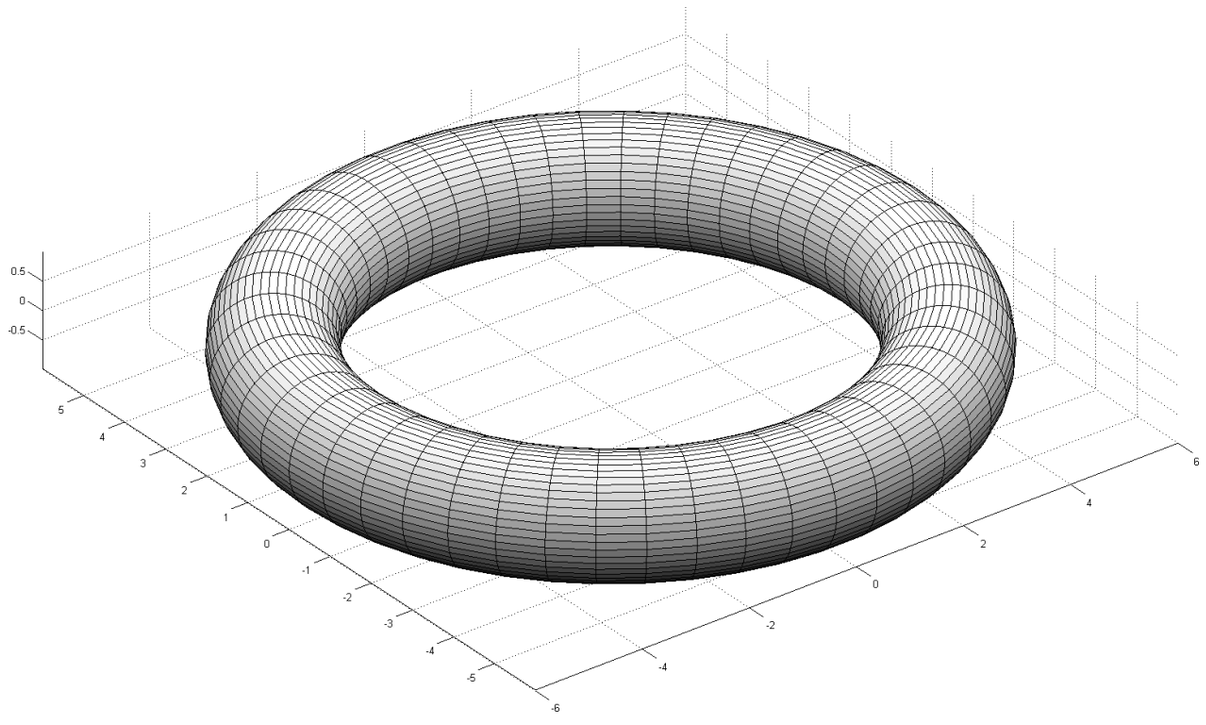
- μ: unif. measure on Lissajous curve $\mathbb{X}_\mu$.
- Filt: distance to $\mathbb{X}_\mu$ in $\mathbb{R}^2$.
- sample $k = 300$ sets of m points for $m = [2100 : 100 : 3000]$.
- compute

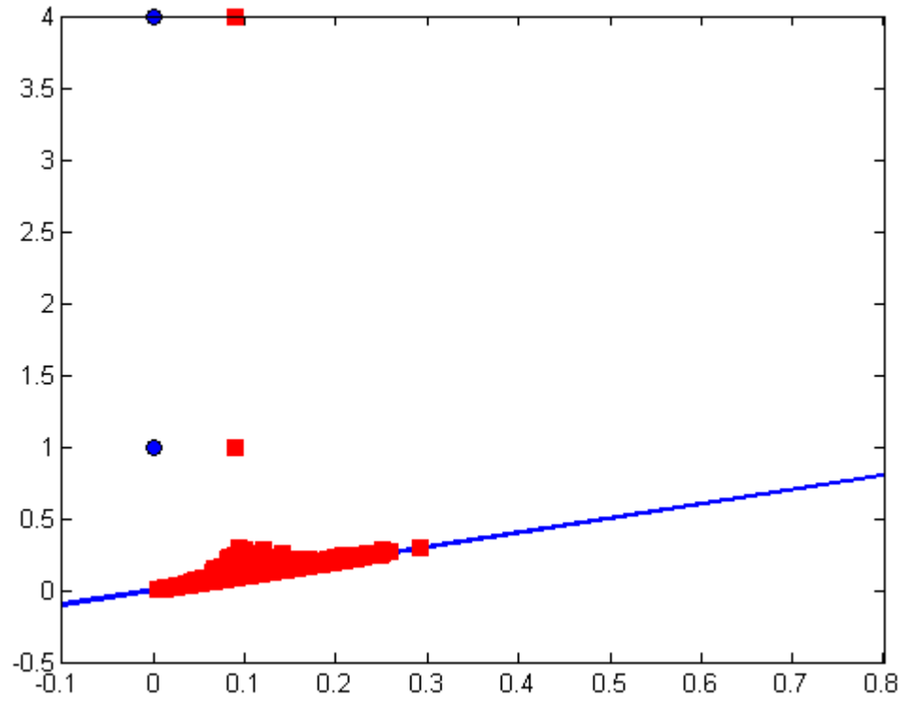
$$\widehat{\mathbb{E}}_m = \widehat{\mathbb{E}}[d_B(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}_n})))].$$

- plot $\log(\widehat{\mathbb{E}}_m)$ as a function of $\log(\log(m)/m)$.



Numerical illustrations

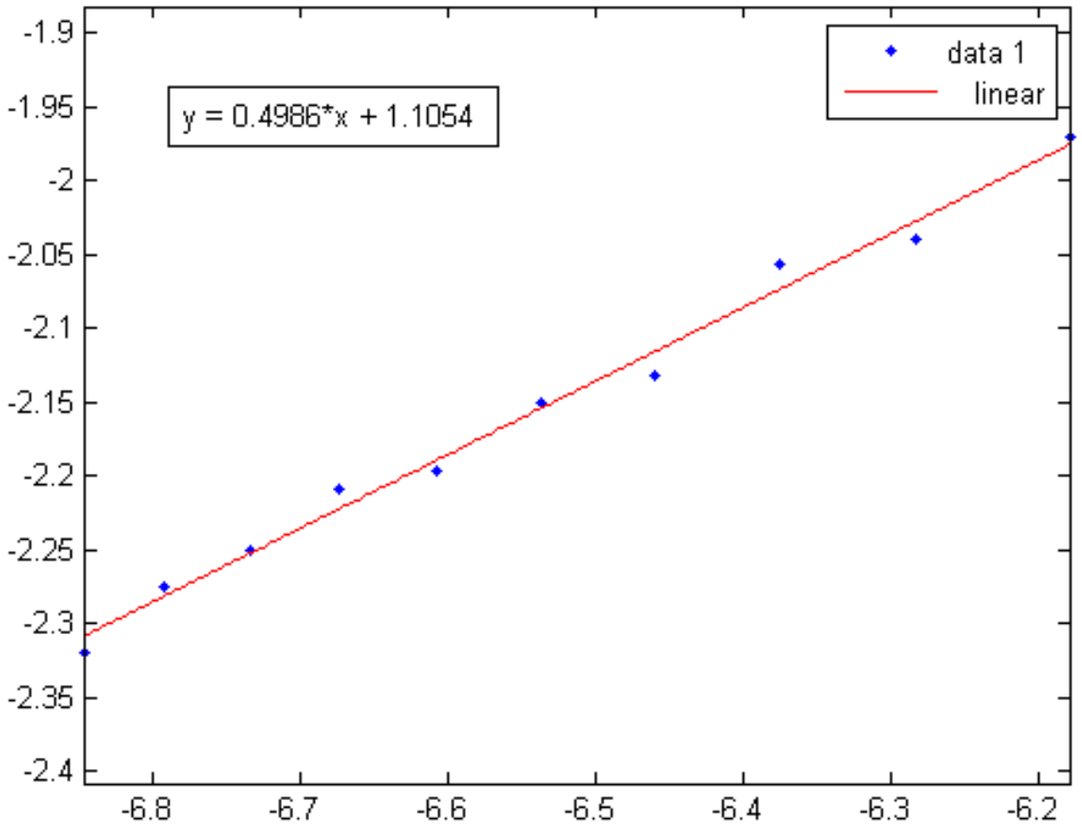




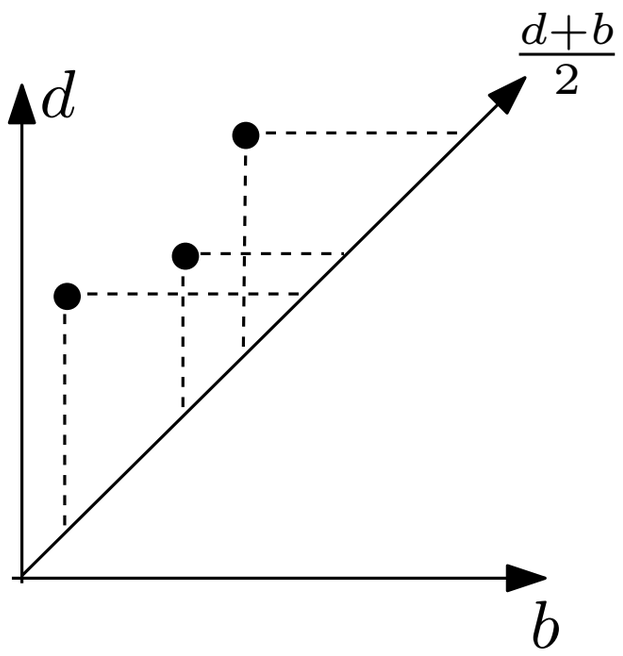
- μ: unif. measure on a torus $\mathbb{X}_\mu$.
- Filt: distance to $\mathbb{X}_\mu$ in $\mathbb{R}^3$.
- sample $k = 300$ sets of n points for $m = [12000 : 1000 : 21000]$.
- compute

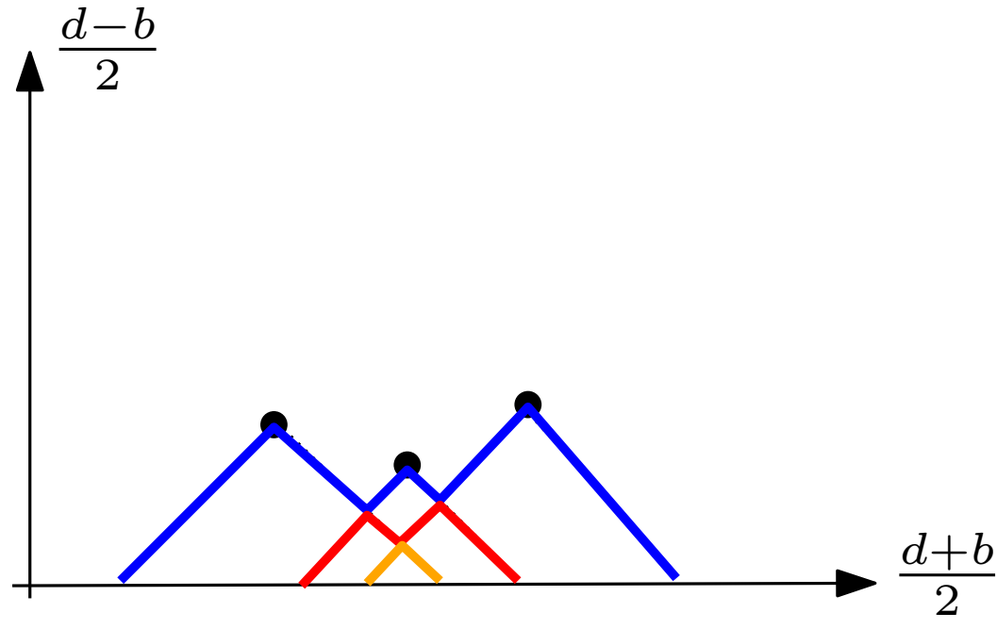
$$\widehat{\mathbb{E}}_m = \widehat{\mathbb{E}}[d_B(\mathsf{dgm}(\mathrm{Filt}(\mathbb{X}_\mu)), \mathsf{dgm}(\mathrm{Filt}(\widehat{\mathbb{X}_m})))].$$

- plot $\log(\widehat{\mathbb{E}}_m)$ as a function of $\log(\log(m)/m)$.



Persistence landscapes





$D = \{(\frac{d_i+b_i}{2}, \frac{d_i+b_i}{2})\} i \in I$

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

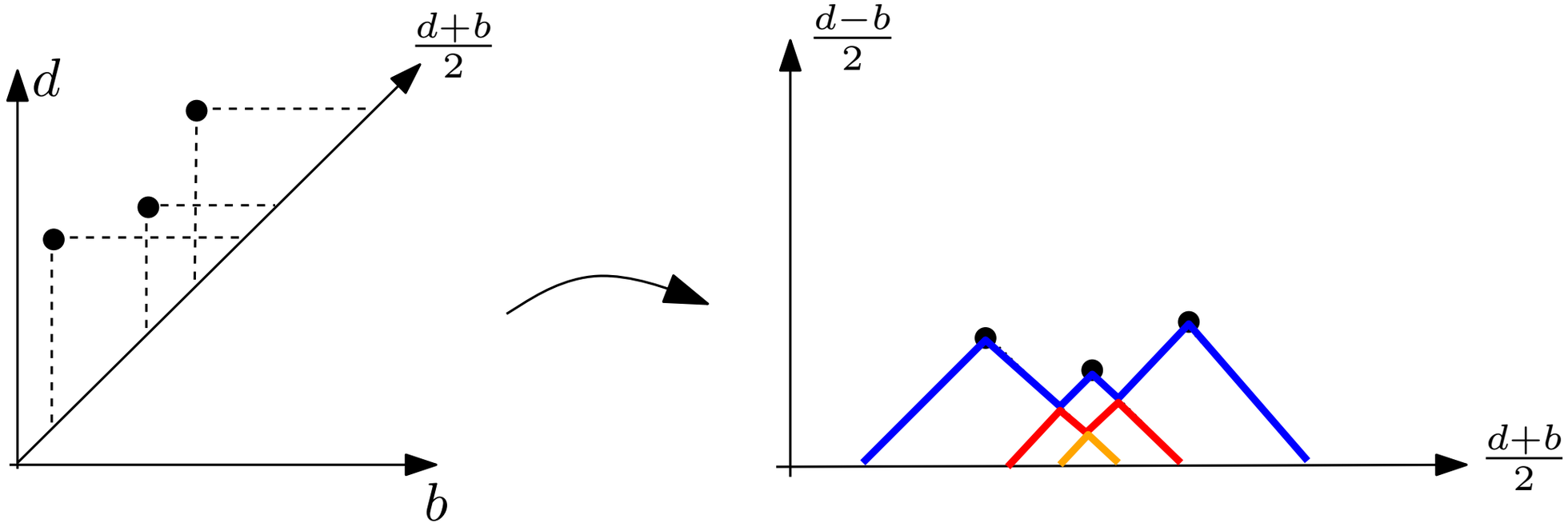
$$\Lambda_p(t) = \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & \text{otherwise.} \end{cases}$$

Persistence landscape [Bubenik 2012]:

$$\lambda_D(k,t) = \underset{p \in \mathsf{dgm}}{\mathsf{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

where kmax is the kth largest value in the set.

Persistence landscapes



Persistence landscape [Bubenik 2012]:

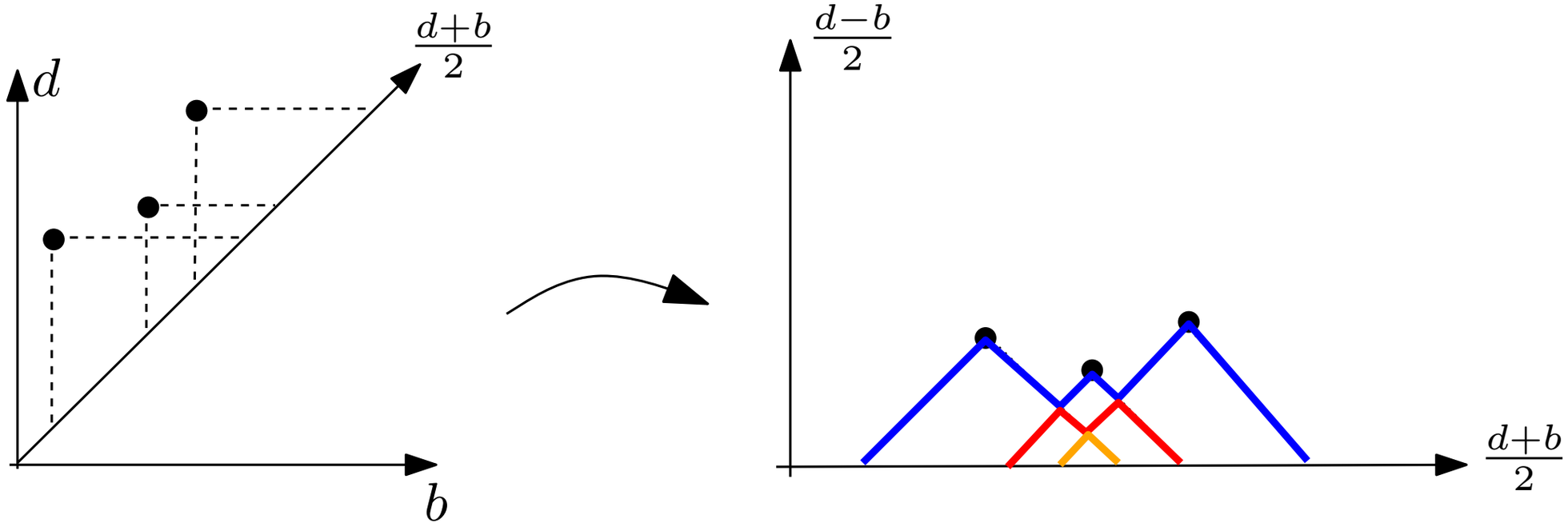
$$\lambda_D(k,t) = \underset{p\in \mathsf{dgm}}{\mathsf{kmax}}\ \Lambda_p(t), \quad t \in \mathbb{R}, k \in \mathbb{N},$$

Properties

- For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $0 \leq \lambda_D(k,t) \leq \lambda_D(k+1,t)$.
- For any $t \in \mathbb{R}$ and any $k \in \mathbb{N}$, $|\lambda_D(k,t) - \lambda_{D'}(k,t)| \leq d_B(D,D')$ where $d_B(D,D')$ denotes the bottleneck distance between D and D'.

stability properties of persistence landscapes

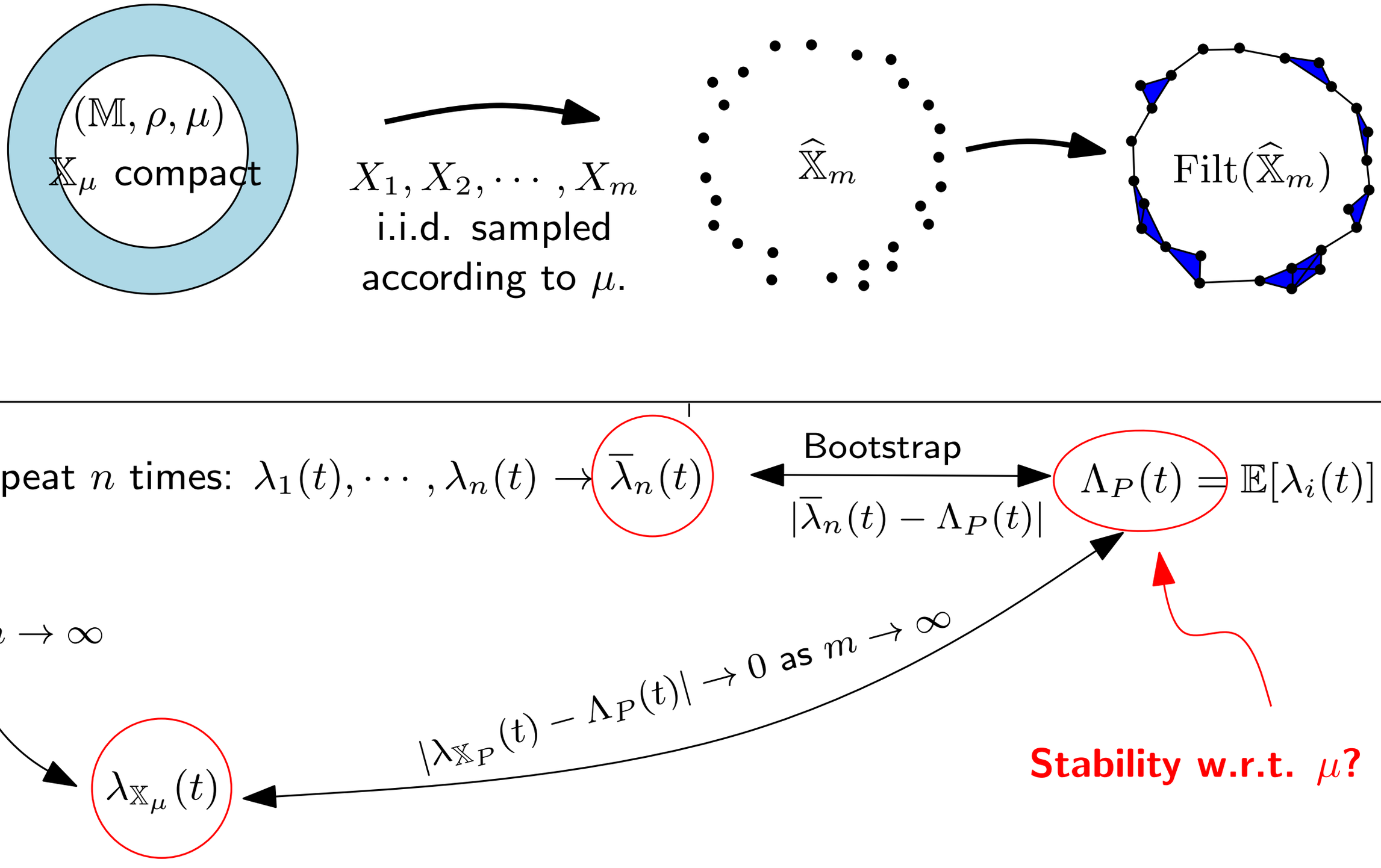
Persistence landscapes



- Persistence encoded as an element of a functional space (vector space!).
- Expectation of distribution of landscapes is well-defined and can be approximated from average of sampled landscapes.
- process point of view: convergence results and convergence rates $\rightarrow$ confidence intervals can be computed using bootstrap.

[C., Fasy, Lecci, Rinaldo, Wasserman SoCG 2014]

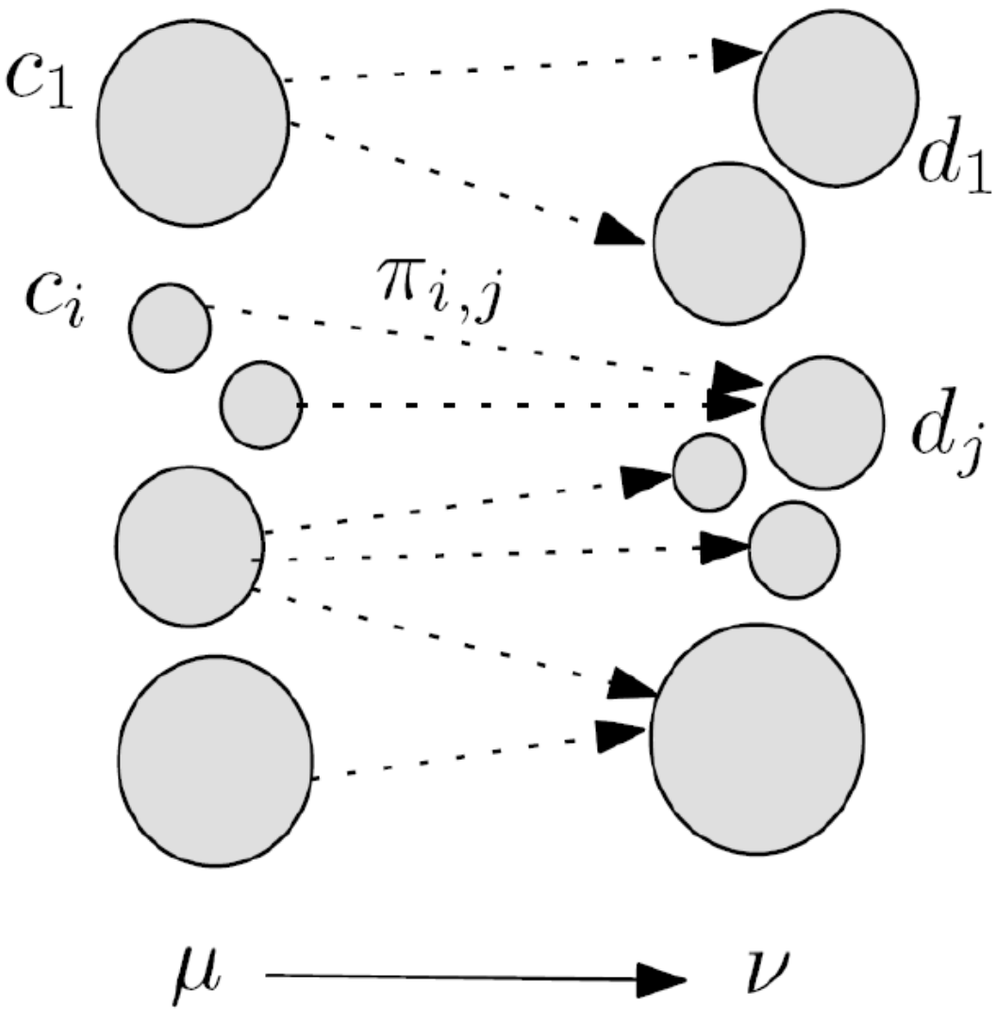
To summarize



Wasserstein distance

Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be probability measures on $\mathbb{M}$ with finite p-moments ($p \geq 1$).

"The" Wasserstein distance $W_p(\mu, \nu)$ quantifies the optimal cost of pushing μ onto ν, the cost of moving a small mass dx from x to y being $\rho(x, y)^p dx$.



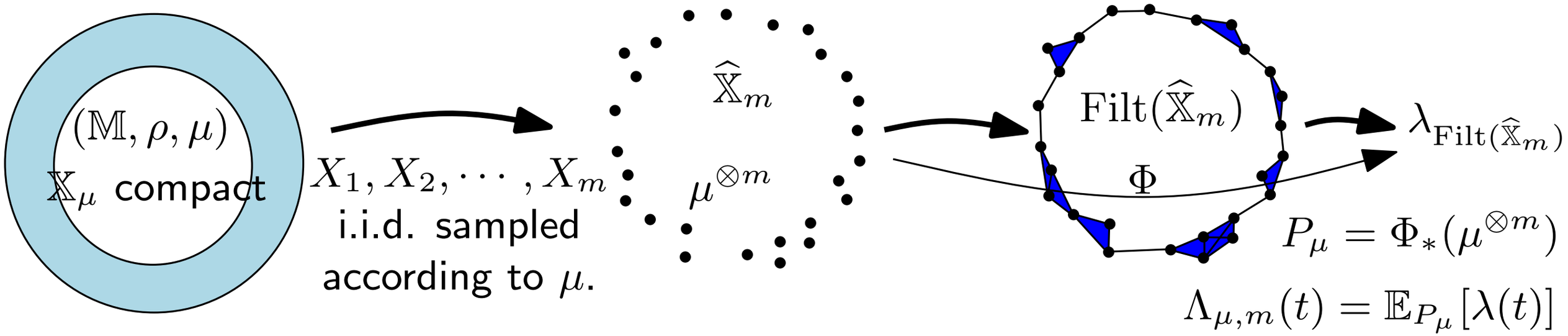
- Transport plan: Π a proba measure on $M \times M$ such that $\Pi(A \times \mathbb{R}^d) = \mu(A)$ and $\Pi(\mathbb{R}^d \times B) = \nu(B)$ for any borelian sets $A, B \subset M$.
- Cost of a transport plan:

$$C(\Pi) = \left(\int_{M \times M} \rho(x, y)^p d\Pi(x, y) \right)^{\frac{1}{p}}$$

- $W_p(\mu, \nu) = \inf_\Pi C(\Pi)$

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

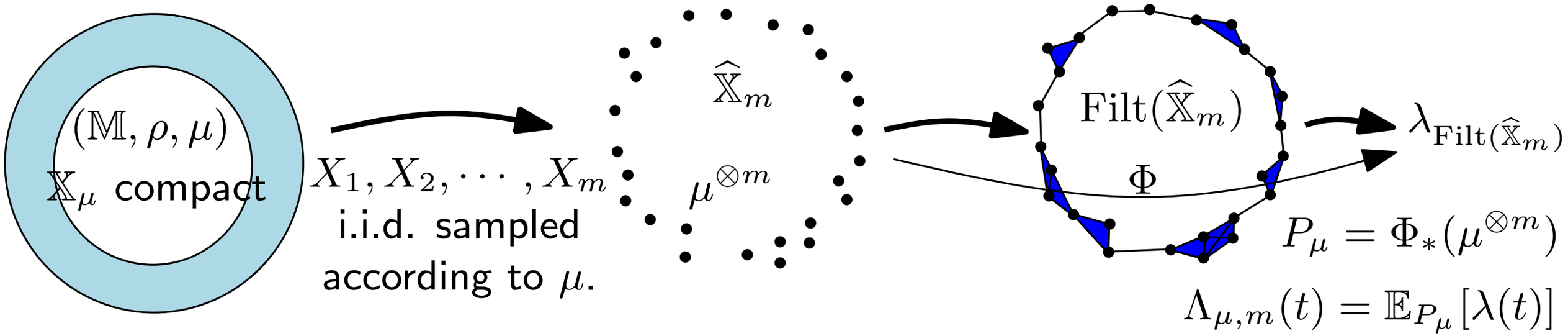
where W_p denotes the Wasserstein distance with cost function $\rho(x, y)^p$.

Remarks:

- similar results by Blumberg et al (2014) in the (Gromov-)Prokhorov metric (for distributions, not for expectations) ;
- Extended to point process setting y L. Decreusefond et al;
- $m^{\frac{1}{p}}$ cannot be replaced by a constant.

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

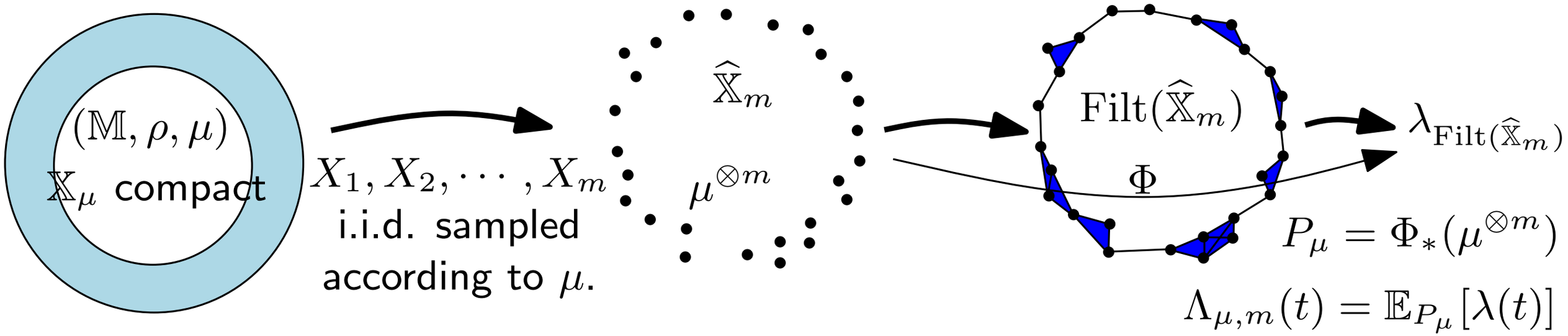
where W_p denotes the Wasserstein distance with cost function $\rho(x, y)^p$.

Consequences:

- Subsampling: efficient and easy to parallelize algorithm to infer topol. information from huge data sets.
- Robustness to outliers.
- R package TDA +Gudhi library: https://project.inria.fr/gudhi/software/

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]



Theorem: Let $(\mathbb{M}, \rho)$ be a metric space and let μ, ν be proba measures on $\mathbb{M}$ with compact supports. We have

$$\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq m^{\frac{1}{p}} W_p(\mu, \nu)$$

where W_p denotes the Wasserstein distance with cost function $\rho(x, y)^p$.

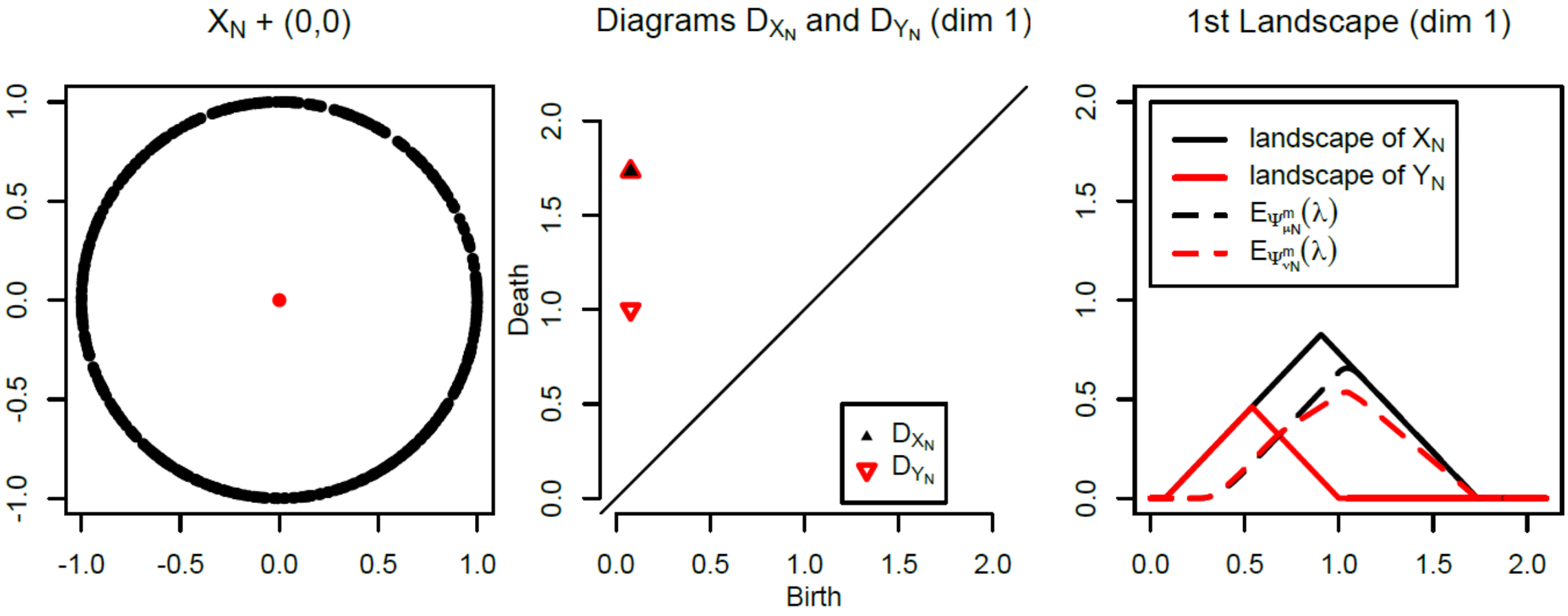
Proof:

1. $W_p(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}} W_p(\mu, \nu)$
2. $W_p(P_\mu, P_\nu) \leq W_p(\mu^{\otimes m}, \nu^{\otimes m})$ (stability of persistence!)
3. $\|\Lambda_{\mu,m} - \Lambda_{\nu,m}\|_\infty \leq W_p(P_\mu, P_\nu)$ (Jensen's inequality)

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

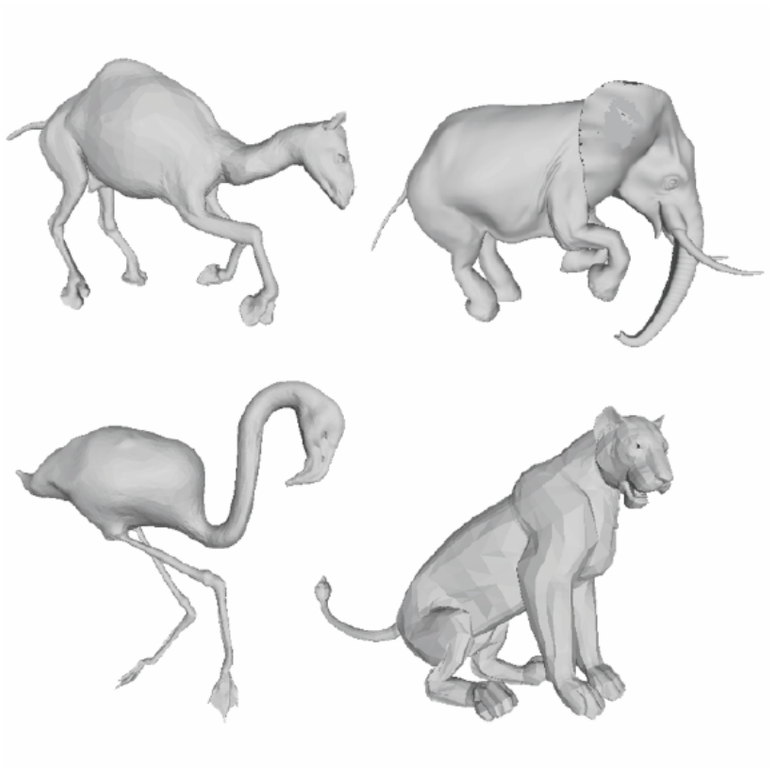
Example: Circle with one outlier.

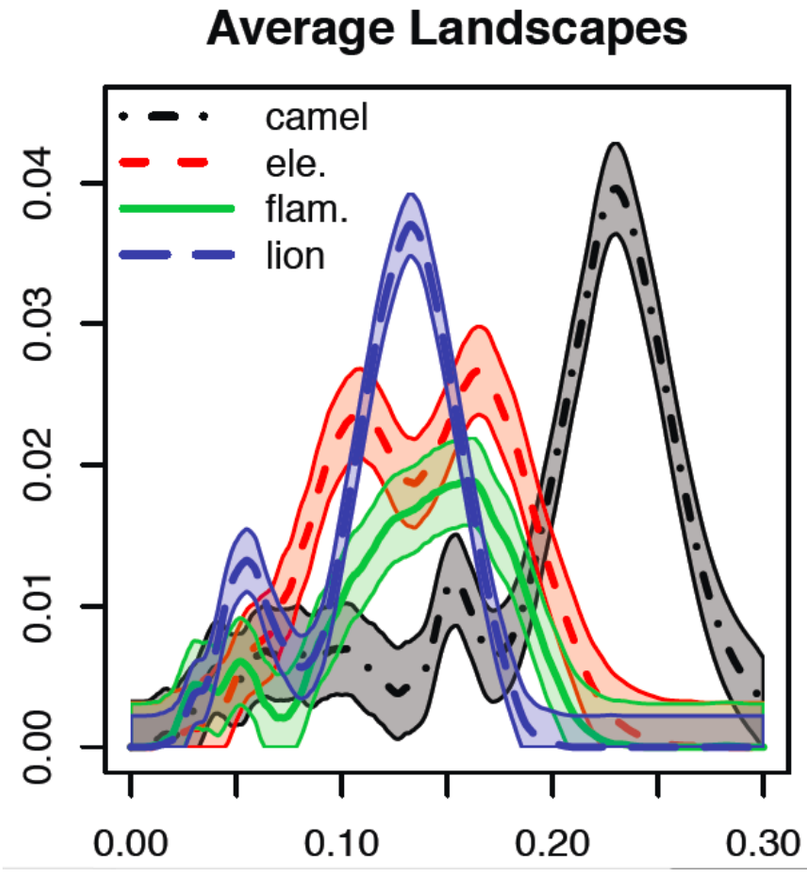


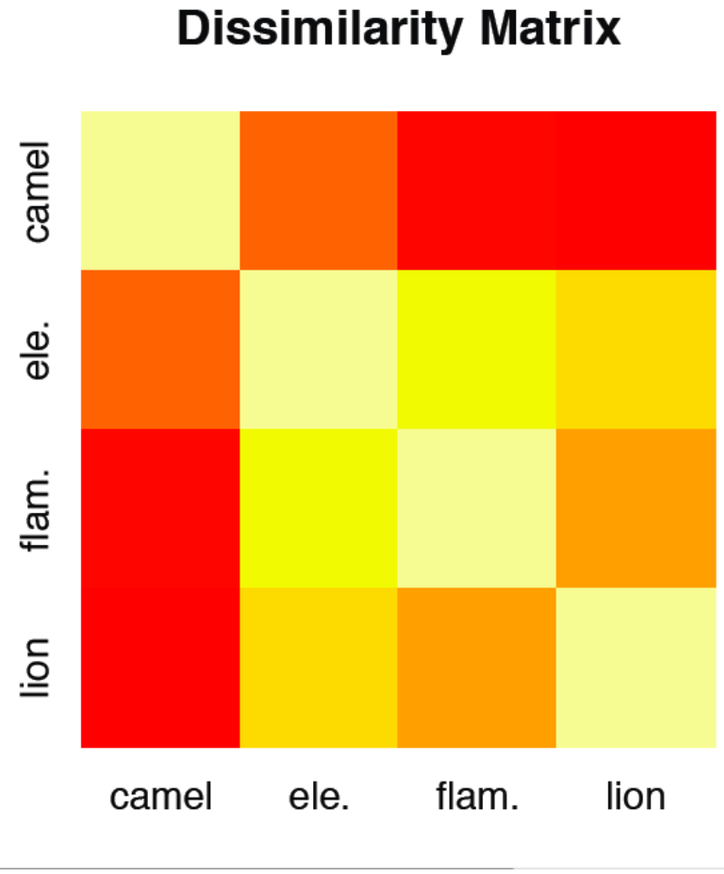
(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

Example: 3D shapes





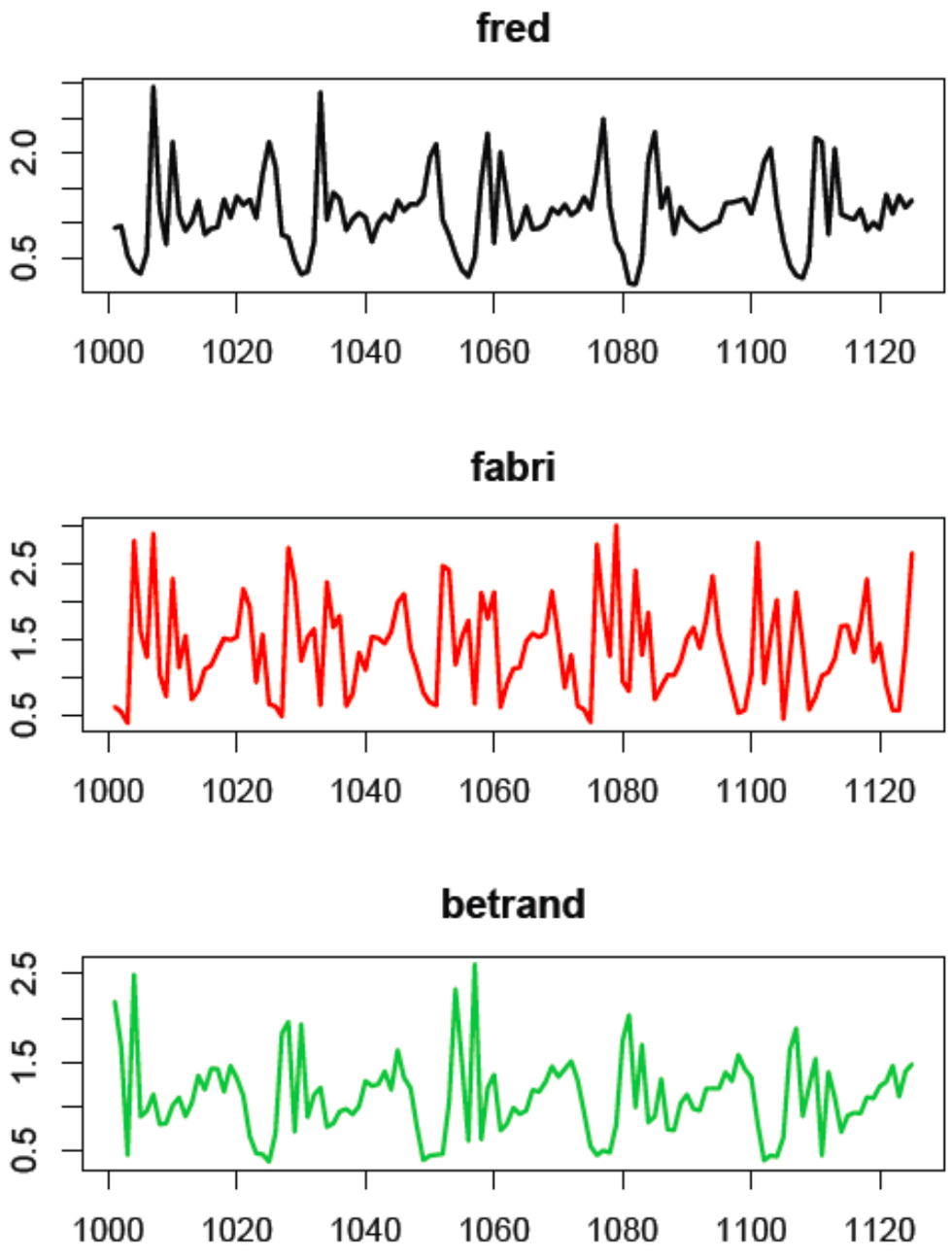


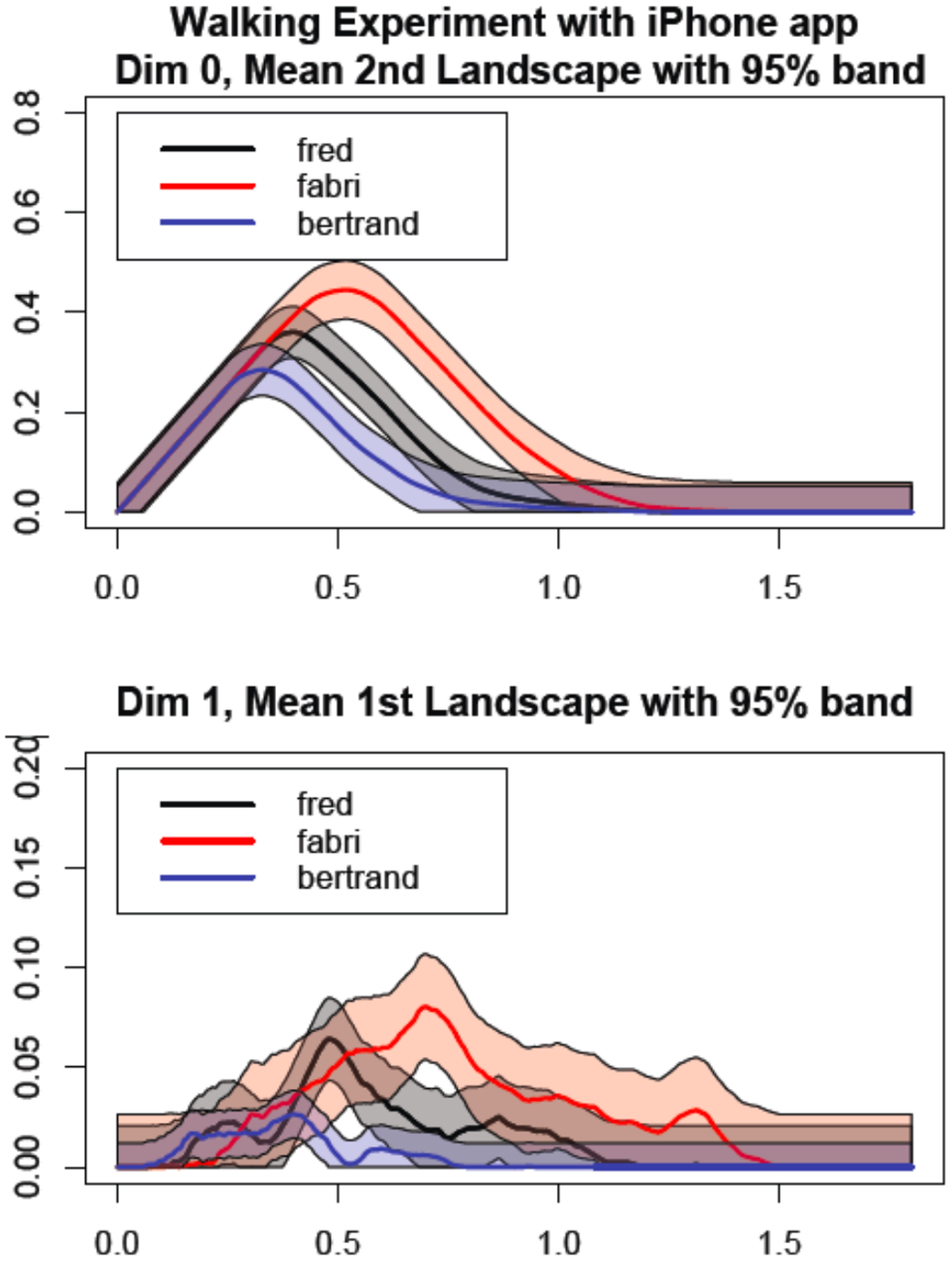
From $n = 100$ subsamples of size $m = 300$

(Sub)sampling and stability of expected landscapes

[C., Fasy, Lecci, Michel, Rinaldo, Wasserman ICML 2015]

(Toy) Example: Accelerometer data from smartphone.





- spatial time series (accelerometer data from the smarphone of users).
- no registration/calibration preprocessing step needed to compare!

Thank you for your attention!

Collaborators: V. de Silva, B. Fasy, D. Cohen-Steiner, M. Glisse, L. Guibas, C. Labruère, F. Lecci, F. Memoli, B. Michel, S. Oudot, A. Rinaldo, L. Wasserman

Software:

- The Gudhi library (C++): https://project.inria.fr/gudhi/software/
- R package TDA