Stability of a Multi-Parameter Persistent Homology Approach to Functional and Structural Data

Presenter: Haibin Hang[†]

Joint work with W. Kim[‡], F. Memoli[‡] and W. Mio[†]

[†] Department of Mathematics, Florida State University
[‡] Department of Mathematics, Ohio State University

Functional Data and Structural Data

Functional data: data sets where each data object $\mathbb{X}_f = (X, d_X, f)$ is a metric space $\mathbb{X} = (X, d_X)$ and a continuous function $f\colon X \to \mathbf{R}$.

Structural data: data sets where each data object $\mathbb{X}_\alpha = (X, d_X, \alpha)$ is a metric space $\mathbb{X} = (X, d_X)$ and a Borel probability measure α.

We simply call $\mathbb{X}_f = (X, d_X, f)$ a functional triple and call $\mathbb{X}_\alpha = (X, d_X, \alpha)$ a metric measure space or mm-space.

These types of data are prevalent in biomedical imaging, network data analysis, material science etc.

Functional Data and Structural Data

Functional data: data sets where each data object $\mathbb{X}_f = (X, d_X, f)$ is a metric space $\mathbb{X} = (X, d_X)$ and a continuous function $f\colon X \to \mathbf{R}$.

Structural data: data sets where each data object $\mathbb{X}_\alpha = (X, d_X, \alpha)$ is a metric space $\mathbb{X} = (X, d_X)$ and a Borel probability measure α.

We simply call $\mathbb{X}_f = (X, d_X, f)$ a functional triple and call $\mathbb{X}_\alpha = (X, d_X, \alpha)$ a metric measure space or mm-space.

These types of data are prevalent in biomedical imaging, network data analysis, material science etc.

Functional Data and Structural Data

Functional data: data sets where each data object $\mathbb{X}_f = (X, d_X, f)$ is a metric space $\mathbb{X} = (X, d_X)$ and a continuous function $f\colon X \to \mathbf{R}$.

Structural data: data sets where each data object $\mathbb{X}_\alpha = (X, d_X, \alpha)$ is a metric space $\mathbb{X} = (X, d_X)$ and a Borel probability measure α.

We simply call $\mathbb{X}_f = (X, d_X, f)$ a functional triple and call $\mathbb{X}_\alpha = (X, d_X, \alpha)$ a metric measure space or mm-space.

These types of data are prevalent in biomedical imaging, network data analysis, material science etc.

Functional Data and Structural Data

Functional data: data sets where each data object $\mathbb{X}_f = (X, d_X, f)$ is a metric space $\mathbb{X} = (X, d_X)$ and a continuous function $f\colon X \to \mathbf{R}$.

Structural data: data sets where each data object $\mathbb{X}_\alpha = (X, d_X, \alpha)$ is a metric space $\mathbb{X} = (X, d_X)$ and a Borel probability measure α.

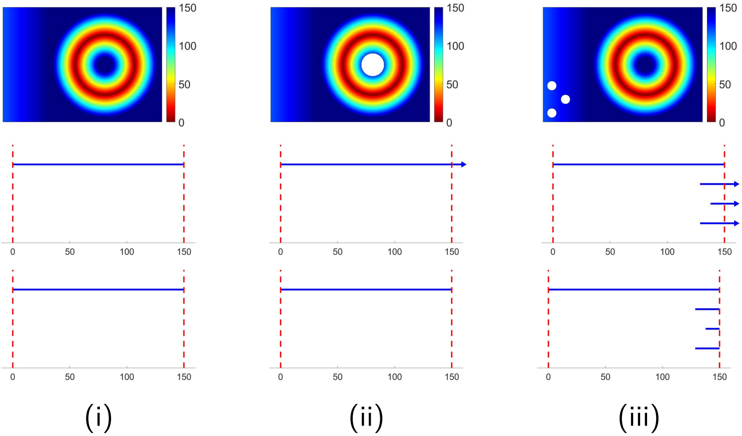
We simply call $\mathbb{X}_f = (X, d_X, f)$ a functional triple and call $\mathbb{X}_\alpha = (X, d_X, \alpha)$ a metric measure space or mm-space.

These types of data are prevalent in biomedical imaging, network data analysis, material science etc.

Motivation

Eg: Similar signals on different domains. In our discussion,

- function $\Longleftrightarrow$ signal;
- **high** function value $\Longleftrightarrow$ **weak** signals!



Goals:

- To construct a metric to compare two signals and downplay the regions with weak signals;
- Use persistent homology (or persistent diagram) to reduce the metric by giving a computable lower bound;

Example: Let f and g have the same domain X, a commonly used metric is $d(f,g) := \|f-g\|_\infty$ and the stability theorem

$$d_{\mathrm{I}}(\mathrm{H}_*(f), \mathrm{H}_*(g)) \leq \|f-g\|_\infty$$

leads to a computable reduction of d.

- The persistent homology level reduction of the metric should give a similar effect of "barcode trimming" in the example.

Goals:

- To construct a metric to compare two signals and downplay the regions with weak signals;
- Use persistent homology (or persistent diagram) to reduce the metric by giving a computable lower bound;

Example: Let f and g have the same domain X, a commonly used metric is $d(f, g) := \|f - g\|_\infty$ and the stability theorem

$$d_{\mathrm{I}}(\mathrm{H}_*(f), \mathrm{H}_*(g)) \leq \|f - g\|_\infty$$

leads to a computable reduction of d.

- The persistent homology level reduction of the metric should give a similar effect of "barcode trimming" in the example.

Goals:

- To construct a metric to compare two signals and downplay the regions with weak signals;
- Use persistent homology (or persistent diagram) to reduce the metric by giving a computable lower bound;

Example: Let f and g have the same domain X, a commonly used metric is $d(f,g) := \|f - g\|_\infty$ and the stability theorem

$$d_{\mathrm{I}}(\mathrm{H}_*(f), \mathrm{H}_*(g)) \leq \|f - g\|_\infty$$

leads to a computable reduction of d.

- The persistent homology level reduction of the metric should give a similar effect of "barcode trimming" in the example.

Goals:

- To construct a metric to compare two signals and downplay the regions with weak signals;
- Use persistent homology (or persistent diagram) to reduce the metric by giving a computable lower bound;

Example: Let f and g have the same domain X, a commonly used metric is $d(f,g) := \|f - g\|_\infty$ and the stability theorem

$$d_{\mathrm{I}}(\mathrm{H}_*(f), \mathrm{H}_*(g)) \leq \|f - g\|_\infty$$

leads to a computable reduction of d.

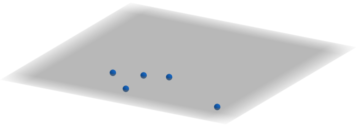
- The persistent homology level reduction of the metric should give a similar effect of "barcode trimming" in the example.

The Metric Which Downplays the Impact of Weak Signals

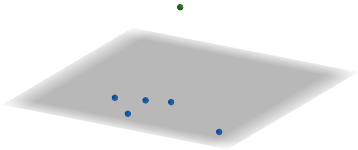
Definition (Cone of functional data)

Given a functional triple $\mathbb{X}_f = (X, d_X, f)$, the new triple $\mathcal{C}\mathbb{X}_f = (\overline{X}, r_{\overline{X}}, f^*)$ is constructed in the following way: for $x, x' \in X$,

1. $\overline{X} = X \sqcup \{*\}$;
2. $r_{\overline{X}}(x, x') = d_X(x, x')$ and $r_{\overline{X}}(x, *) = r_{\overline{X}}(*, x) = 0$;
3. $f^*(x) = f(x)$ and $f^*(*) = \sup_{x \in X} f(x)$.



(i)X

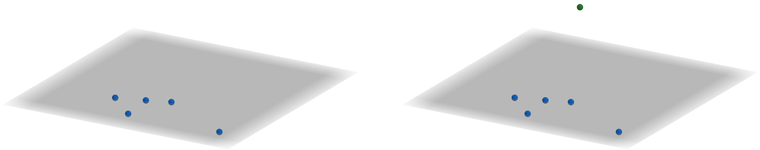


(ii)$\overline{X}$

Definition (Cone of functional data)

Given a functional triple $\mathbb{X}_f = (X, d_X, f)$, the new triple $\mathcal{C}\mathbb{X}_f = (\overline{X}, r_{\overline{X}}, f^*)$ is constructed in the following way: for $x, x' \in X$,

1. $\overline{X} = X \sqcup \{*\}$;
2. $r_{\overline{X}}(x, x') = d_X(x, x')$ and $r_{\overline{X}}(x, *) = r_{\overline{X}}(*, x) = 0$;
3. $f^*(x) = f(x)$ and $f^*(*) = \sup_{x \in X} f(x)$.



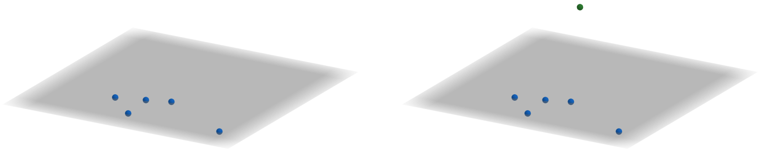
(i)X (ii)$\overline{X}$

The Metric Which Downplays the Impact of Weak Signals

Definition (Cone of functional data)

Given a functional triple $\mathbb{X}_f = (X, d_X, f)$, the new triple $\mathcal{C}\mathbb{X}_f = (\overline{X}, r_{\overline{X}}, f^*)$ is constructed in the following way: for $x, x' \in X$,

1. $\overline{X} = X \sqcup \{*\}$;
2. $r_{\overline{X}}(x, x') = d_X(x, x')$ and $r_{\overline{X}}(x, *) = r_{\overline{X}}(*, x) = 0$;
3. $f^*(x) = f(x)$ and $f^*(*) = \sup_{x \in X} f(x)$.



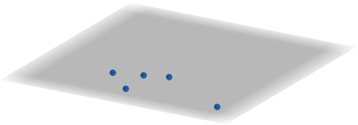
(i)X (ii)$\overline{X}$

The Metric Which Downplays the Impact of Weak Signals

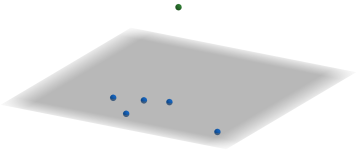
Definition (Cone of functional data)

Given a functional triple $\mathbb{X}_f = (X, d_X, f)$, the new triple $\mathcal{C}\mathbb{X}_f = (\overline{X}, r_{\overline{X}}, f^*)$ is constructed in the following way: for $x, x' \in X$,

1. $\overline{X} = X \sqcup \{*\}$;
2. $r_{\overline{X}}(x, x') = d_X(x, x')$ and $r_{\overline{X}}(x, *) = r_{\overline{X}}(*, x) = 0$;
3. $f^*(x) = f(x)$ and $f^*(*) = \sup_{x \in X} f(x)$.



(i)X

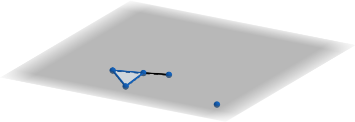


(ii)$\overline{X}$

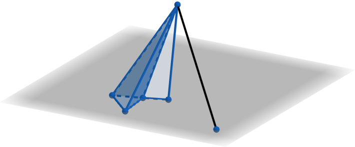
Given any set Z with non-negative symmetric function $r : Z \times Z \to \mathbb{R}$, we define **Rips complex** at scale $s \geq 0$

$$\mathcal{R}_s(Z) := \{\sigma \subset Z | \sigma \text{ is finite set and } \mathrm{diam}(\sigma) \leq s\},$$

here $\mathrm{diam}(\sigma) := \max_{z_1, z_2 \in \sigma} r(z_1, z_2)$.



(i)$\mathcal{R}_s(X)$

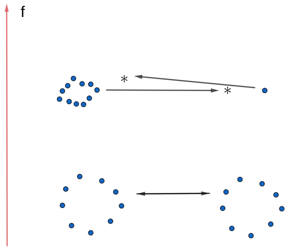


(ii)$\mathcal{R}_s(\overline{X})$

Definition (ε-matching pair)

A pair of maps $\phi : \overline{X} \to \overline{Y}$ and $\psi : \overline{Y} \to \overline{X}$ which preserve the cone points and satisfy the following is said to be an *ε-matching pair* between $\mathcal{C}\mathbb{X}_f$ and $\mathcal{C}\mathbb{Y}_g$: for all $x, x' \in X$ and all $y, y' \in Y$,

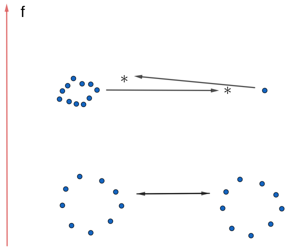
1. $g^* \circ \phi(x) < f^*(x) + \varepsilon$, $f^* \circ \psi(y) < g^*(y) + \varepsilon$;
2. $r_{\overline{Y}}(\phi(x), \phi(x')) < r_{\overline{X}}(x, x') + \varepsilon$, $r_{\overline{X}}(\psi(y), \psi(y')) < r_{\overline{Y}}(y, y') + \varepsilon$;
3. $r_{\overline{X}}(x, \psi \circ \phi(x)) < 2\varepsilon$, $r_{\overline{Y}}(y, \phi \circ \psi(y)) < 2\varepsilon$.



Definition (ε-matching pair)

A pair of maps $\phi : \overline{X} \to \overline{Y}$ and $\psi : \overline{Y} \to \overline{X}$ which preserve the cone points and satisfy the following is said to be an *ε-matching pair* between $\mathcal{C}\mathbb{X}_f$ and $\mathcal{C}\mathbb{Y}_g$: for all $x, x' \in X$ and all $y, y' \in Y$,

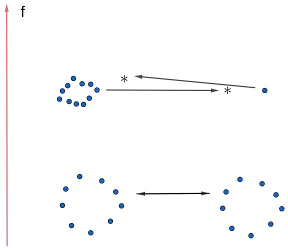
1. $g^* \circ \phi(x) < f^*(x) + \varepsilon$, $f^* \circ \psi(y) < g^*(y) + \varepsilon$;
2. $r_{\overline{Y}}(\phi(x), \phi(x')) < r_{\overline{X}}(x, x') + \varepsilon$, $r_{\overline{X}}(\psi(y), \psi(y')) < r_{\overline{Y}}(y, y') + \varepsilon$;
3. $r_{\overline{X}}(x, \psi \circ \phi(x)) < 2\varepsilon$, $r_{\overline{Y}}(y, \phi \circ \psi(y)) < 2\varepsilon$.



Definition (ε-matching pair)

A pair of maps $\phi : \overline{X} \to \overline{Y}$ and $\psi : \overline{Y} \to \overline{X}$ which preserve the cone points and satisfy the following is said to be an *ε-matching pair* between $\mathcal{C}\mathbb{X}_f$ and $\mathcal{C}\mathbb{Y}_g$: for all $x, x' \in X$ and all $y, y' \in Y$,

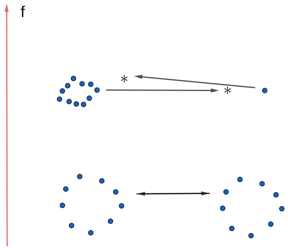
1. $g^* \circ \phi(x) < f^*(x) + \varepsilon$, $f^* \circ \psi(y) < g^*(y) + \varepsilon$;
2. $r_{\overline{Y}}(\phi(x), \phi(x')) < r_{\overline{X}}(x, x') + \varepsilon$, $r_{\overline{X}}(\psi(y), \psi(y')) < r_{\overline{Y}}(y, y') + \varepsilon$;
3. $r_{\overline{X}}(x, \psi \circ \phi(x)) < 2\varepsilon$, $r_{\overline{Y}}(y, \phi \circ \psi(y)) < 2\varepsilon$.



Definition (ε-matching pair)

A pair of maps $\phi : \overline{X} \to \overline{Y}$ and $\psi : \overline{Y} \to \overline{X}$ which preserve the cone points and satisfy the following is said to be an ε-*matching pair* between $\mathcal{C}\mathbb{X}_f$ and $\mathcal{C}\mathbb{Y}_g$: for all $x, x' \in X$ and all $y, y' \in Y$,

1. $g^* \circ \phi(x) < f^*(x) + \varepsilon$, $f^* \circ \psi(y) < g^*(y) + \varepsilon$;
2. $r_{\overline{Y}}(\phi(x), \phi(x')) < r_{\overline{X}}(x, x') + \varepsilon$, $r_{\overline{X}}(\psi(y), \psi(y')) < r_{\overline{Y}}(y, y') + \varepsilon$;
3. $r_{\overline{X}}(x, \psi \circ \phi(x)) < 2\varepsilon$, $r_{\overline{Y}}(y, \phi \circ \psi(y)) < 2\varepsilon$.



Definition

$$\rho(\mathbb{X}_f, \mathbb{Y}_g) := \inf\{\varepsilon > 0 | \mathcal{C}\mathbb{X}_f \text{ and } \mathcal{C}\mathbb{Y}_g \text{ are } \varepsilon\text{-matched}\}.$$

Remark

Since ρ does not satisfy the triangle inequality, we consider the metric generated by it.

Definition (matching distance)

$d_{\mathrm{M}}(\mathbb{X}_f, \mathbb{Y}_g) := \inf \sum_{i=1}^{m} \rho(\mathbb{Z}_{i-1}, \mathbb{Z}_i)$, Here the infimum is taken over all sequences $\{\mathbb{Z}_i\}_{i=0}^{m}$ of functional data with $\mathbb{Z}_0 = \mathbb{X}_f$ and $\mathbb{Z}_m = \mathbb{Y}_g$.

For $\mathcal{C}\mathbb{X}_f = (\overline{X}, r_{\overline{X}}, f^*)$, we take sub-level set and Rips complex **consecutively**:

$$\overline{X}_a := \{x \in \overline{X} | f^*(x) \leq a\}$$

$$\mathcal{R}_s(\overline{X}_a) := \{\sigma \subset \overline{X}_a | \mathrm{diam}(\sigma) \leq s\}$$

and formalize a **bi-parameter** filtration

$$\mathcal{R}(\mathcal{C}\mathbb{X}_f) := \{\mathcal{R}_s(\overline{X}_a) \hookrightarrow \mathcal{R}_t(\overline{X}_b)\}_{0 \leq s \leq t, a \leq b}$$

$$\mathrm{H}_*\mathcal{R}(\mathcal{C}\mathbb{X}_f) := \{\mathrm{H}_*(\mathcal{R}_s(\overline{X}_a)) \to \mathrm{H}_*(\mathcal{R}_t(\overline{X}_b))\}_{0 \leq s \leq t, a \leq b}$$

Theorem (Stability)

For any pair of triples $\mathbb{X}_f, \mathbb{Y}_g$ we have

$$d_{\mathrm{I}}\big(\mathrm{H}_*\mathcal{R}(\mathcal{C}\mathbb{X}_f), \mathrm{H}_*\mathcal{R}(\mathcal{C}\mathbb{Y}_g)\big) \leq d_{\mathrm{M}}(\mathbb{X}_f, \mathbb{Y}_g).$$

Strategy to Structural Data

Our way to deal with structural data is to transform it into functional data via a centrality function.

Definition

Given mm-space $\mathbb{X}_\alpha = (X, d_X, \alpha)$ and $p \geq 1$, the p-centrality function is defined as

$$s_{\alpha,p}(x) = \left(\int_X d_X^p(x, x') d\alpha(x') \right)^{1/p}$$

for $\forall x \in X$.

Now we have an operator $\mathcal{S}_p$ from structural data to functional data s.t.

$$(X, d_X, \alpha) \mapsto (X, d_X, s_{\alpha,p}).$$

We use the pull-back of d_M as a metric on structural data

$$d_{M,p}(\mathbb{X}_\alpha, \mathbb{Y}_\beta) := d_M(\mathcal{S}_p\mathbb{X}_\alpha, \mathcal{S}_p\mathbb{Y}_\beta)$$

Remark

The centrality function usually take small values at the center of a distribution. Hence $d_{M,p}$ stress regions with high density.

Then the stability theorem becomes

Theorem

For any pair of triples $\mathbb{X}_\alpha, \mathbb{Y}_\beta$ we have

$$d_{\mathrm{I}}(\mathrm{H}_*\mathcal{R}\mathcal{C}\mathcal{S}_p\mathbb{X}_\alpha, \mathrm{H}_*\mathcal{R}\mathcal{C}\mathcal{S}_p\mathbb{Y}_\beta) \leq d_{M,p}(\mathbb{X}_\alpha, \mathbb{Y}_\beta)$$

Some Consistency Results

Corollary

Given two Borel measures α and β on metric space $\mathbb{X} = (X, d_X)$ we have

$$d_{\mathrm{I}}(\mathrm{H}_*\mathcal{RCS}_p\mathbb{X}_\alpha, \mathrm{H}_*\mathcal{RCS}_p\mathbb{X}_\beta) \leq W_p(\alpha, \beta).$$

For simplicity, we call $\mathrm{H}_*\mathcal{RCS}_p\mathbb{X}_\alpha$ the 2D-persistent homology of $\mathbb{X}_\alpha$.

Let $\{x_i\}$ be a sequence of independent identically distributed random variables under α and

- $\alpha_n = 1/n \sum_{i=1}^{n} \delta_{x_i}$;
- $A_n = \{x_1, x_2, \cdots, x_n\}$.

Recall that W_p metrize weak convergence and α_n converges to α weakly a.s., we get:

Corollary

The 2D-persistent homology of (X, d_X, α_n) converges to the 2D-persistent homology of (X, d_X, α) under d_{I} almost surely.

Furthermore, we can discretize the domain X:

Corollary

Given mm-space (X, d_X, α) with $\mathrm{supp}[\alpha] = X$ compact, then the 2D-persistent homology of (A_n, d_X, α_n) converges to the 2D-persistent homology of (X, d_X, α) under d_{I} almost surely.

Thanks for Your Attention!