March 26, 2010

Solomonov seminar
Institut Jožefa Štefana

**Persistence-based Clustering**

Primož Škraba
Geometrica INRIA-Saclay

joint work with
Frédéric Chazal, Steve Y. Oudot, Leonidas J. Guibas
Clustering

- Input samples/point cloud
• Input samples/point cloud

• "Important" segments/clusters
Clustering

- Input samples/point cloud
- "Important" segments/clusters
  ill-posed problem
Clustering

- Input samples/point cloud
- "Important" segments/clusters
  ill-posed problem
- Extensive previous work
  - $k$-means
  - spectral clustering
  - mode-seeking (mean-shift)
• Input samples/point cloud

• ”Important” segments/clusters
  ill-posed problem

• Extensive previous work
  - $k$-means
  - spectral clustering
  - mode-seeking (mean-shift)

• Our viewpoint:
  data points drawn at random from some unknown density distribution $f$
Definition of a Cluster

- Basins of attraction of the peaks of $f$
Definition of a Cluster

- Basins of attraction of the peaks of $f$
Definition of a Cluster

- Basins of attraction of the peaks of $f$
- Samples drawn from $f$
Definition of a Cluster

- Basins of attraction of the peaks of $f$
- Samples drawn from $f$

Can we approximate the basins of attraction from the samples?
Topological Interlude

• Studies how spaces are connected

• Classification of spaces
Topological Interlude

- Studies how spaces are connected
- Classification of spaces

- Computational topology
  - Qualitative, global description of large data sets
  - Only know local neighborhood information
Topological Interlude

- Studies how spaces are connected
- Classification of spaces

- Computational topology
  - Qualitative, global description of large data sets
  - Only know local neighborhood information

- Homology
Topological Interlude

• Studies how spaces are connected
• Classification of spaces

• Computational topology
  - Qualitative, global description of large data sets
  - Only know local neighborhood information

• Homology
Topological Interlude

- Studies how spaces are connected
- Classification of spaces

- Computational topology
  - Qualitative, global description of large data sets
  - Only know local neighborhood information

- Homology
Topological Interlude

• Studies how spaces are connected
• Classification of spaces

• Computational topology
  - Qualitative, global description of large data sets
  - Only know local neighborhood information

• Homology

  Connected components
  Holes
  Voids
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$. 

![Persistence-based Approach](image)
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}(\mathbb{R})$ as $\alpha$ spans $\mathbb{R}$. 

![Diagram showing the evolution of topology of super-level sets](attachment:diagram.png)
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$.
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$. 
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$. 

![Diagram showing the evolution of topology of super-level sets](image-url)
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$. 

![Graph showing the evolution of topology](image-url)
Filtrations

- Filtration: sequence of super-level sets $\hat{f}^{-1}(\left[\alpha, \infty\right))$ as $\alpha$ decreases.

$$X_\infty \subseteq \ldots \subseteq X_\alpha \subseteq \ldots \subseteq X_\beta \subseteq \ldots \subseteq X_0 \subseteq \ldots \subseteq X_{-\infty} \quad \forall \alpha \geq \beta$$
Filtrations

- Filtration: sequence of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ decreases.

\[ X_\infty \subseteq \ldots \subseteq X_\alpha \subseteq \ldots \subseteq X_\beta \subseteq \ldots \subseteq X_0 \subseteq \ldots \subseteq X_{-\infty} \quad \forall \alpha \geq \beta \]

- Indexed by a parameter
Filtrations

• Filtration: sequence of super-level sets \( \hat{f}^{-1}([\alpha, \infty)) \) as \( \alpha \) decreases.

\[
X_\infty \subseteq \ldots \subseteq X_\alpha \subseteq \ldots \subseteq X_\beta \subseteq \ldots \subseteq X_0 \subseteq \ldots \subseteq X_{-\infty} \quad \forall \alpha \geq \beta
\]

• Indexed by a parameter
Filtrations

• Filtration: sequence of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ decreases.

$X_{\infty} \subseteq \ldots \subseteq X_{\alpha} \subseteq \ldots \subseteq X_{\beta} \subseteq \ldots \subseteq X_0 \subseteq \ldots \subseteq X_{-\infty} \quad \forall \alpha \geq \beta$

• Indexed by a parameter
  - Birth and death time
Filtrations

- Filtration: sequence of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ decreases.

$$X_\infty \subseteq \ldots \subseteq X_\alpha \subseteq \ldots \subseteq X_\beta \subseteq \ldots \subseteq X_0 \subseteq \ldots \subseteq X_{-\infty} \quad \forall \alpha \geq \beta$$

- Indexed by a parameter
  - Birth and death time

- Rich mathematical structure

"Topological Persistence and Simplification,"
Edelsbrunner, Letscher, and Zomorodian, FOCS ’00

"Computing Persistent Homology,"
Carlsson and Zomorodian, SOCG ’04
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}(\alpha, \infty)$ as $\alpha$ spans $\mathbb{R}$. 

\begin{center}
\includegraphics[width=\textwidth]{diagram.png}
\end{center}
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$. 

![Diagram showing the evolution of topology of super-level sets](image)
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets \( f^{-1}([\alpha, \infty)) \) as \( \alpha \) spans \( \mathbb{R} \).
Persistence-based Approach
in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$.

- finite set of intervals (barcode) encode birth/death of homological features.
Persistence-based Approach in a nutshell...

- evolution of topology of super-level sets $\hat{f}^{-1}([\alpha, \infty))$ as $\alpha$ spans $\mathbb{R}$.

- finite set of intervals (barcode) encode birth/death of homological features.

- barcode of $\hat{f}$ is close to barcode of $f$ provided that $\|\hat{f} - f\|_\infty$ is small.

[Cohen-Steiner, Edelsbrunner, Harer ’05]
Topological Definition of a Cluster

- Basins of attraction of “significant” peaks of $f$
- Samples drawn from $f$
Topological Definition of a Cluster

- Basins of attraction of “significant” peaks of $f$
- Samples drawn from $f$

Clusters: Prominent peaks correspond to persistent connected components of the super-level set filtration of $f$
Approximating Persistence
Approximating Persistence
Approximating Persistence
Approximating Persistence
Approximating Persistence
Approximating Persistence
Approximating Persistence

\[ d_B^\infty(D, D') = \inf_{\Phi: D \to D'} \text{multibijection} \left( \sup_{p \in D} d^\infty(p, \Phi(p)) \right) \]
Approximating Persistence

- Stable under function perturbations
- Stable under pairwise distances

Bottleneck distance

\[ d_B^\infty(D, D') = \inf_{\Phi:D \to D'} \text{ multibijection} \left( \sup_{p \in D} d^\infty(p, \Phi(p)) \right) \]
Better-known Barcodes

- Dendograms
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms

- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms

- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration
Better-known Barcodes

- Dendograms
- Single-linkage clustering
  - Distance filtration

The lengths are stable but the connections are not.
How do we compute clusters from a PD?
Computing Clusters

How do we compute clusters from a PD?

**Input:** Samples with estimated density $\hat{f}$
Computing Clusters

How do we compute clusters from a PD?

**Input:** Samples with estimated density $\hat{f}$

Two steps:
1. Mode-seeking step [Koontz et. al. '76]
Computing Clusters

How do we compute clusters from a PD?

**Input:** Samples with estimated density $\hat{f}$

Two steps:
1. Mode-seeking step [Koontz et. al. ’76]
How do we compute clusters from a PD?

**Input:** Samples with estimated density $\hat{f}$

Two steps:

1. Mode-seeking step [Koontz et. al. '76]
2. Merge clusters according to persistence
Algorithm

• Input: $f(x), R_\delta, \alpha$
Algorithm

- Input: \( f(x), \mathcal{R}_\delta, \alpha \)

1. Sort \( x \) according to \( f \)

2. For \( x \in L \)
   
   2a. For neighbors of \( x \) in \( \mathcal{R}_\delta \)
   
   If no higher neighbors \( \Rightarrow \) new cluster
   
   else assign \( x \) to \( \nabla f \)

2b. For adjacent clusters \( y \) to \( x \)
   
   if \( |f(y) - f(x)| \leq \alpha \)

   merge into oldest adjacent cluster
Putting it together

- Estimate density
- Run algorithm with $\alpha = \infty$
  - Standard persistence algorithm
- Use persistence diagram to choose threshold
- Re-run algorithm
Putting it together

- Estimate density
- Run algorithm with $\alpha = \infty$
  - Standard persistence algorithm
- Use persistence diagram to choose threshold
- Re-run algorithm
Choice of Graphs

- Notion of a neighborhood
Choice of Graphs

- Notion of a neighborhood
- Many choices
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
    - Simple to compute
    - Requires only pairwise distances
    - Can be built in any metric space
    - Provable reconstruction
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
  - k-NN graph
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
  - k-NN graph
Choice of Graphs

• Notion of a neighborhood

• Many choices
  - Vietoris-Rips graph
  - k-NN graph

  - Simple to compute
  - Requires only pairwise distances
  - Can be built in any metric space
  - Sparse
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
  - k-NN graph
  - Delaunay graph
Choice of Graphs

- Notion of a neighborhood
- Many choices
  - Vietoris-Rips graph
  - k-NN graph
  - Delaunay graph
Choice of Graphs

- Notion of a neighborhood

- Many choices
  - Vietoris-Rips graph
  - k-NN graph
  - Delaunay graph

- Parameter-free

- Sparse

- Fast computation in low-dimensional Euclidean space
Density Estimation

- A field of study in itself
Density Estimation

- A field of study in itself

- Gaussian kernel
  - Bandwidth parameter $h$

$$\hat{f}(p) = \frac{1}{|N_p|} \sum_{q \in N_p} e^{-\frac{d(p-q_i)^2}{2h}}$$
Density Estimation

- A field of study in itself

- Gaussian kernel
  - Bandwidth parameter $h$
    \[
    \hat{f}(p) = \frac{1}{|\mathcal{N}_p|} \sum_{q \in \mathcal{N}_p} e^{-\frac{d(p-q_i)^2}{2h}}
    \]

- Distance to a measure
  - Number of neighbors
    \[
    \hat{f}(p) = \left( \frac{1}{k} \sum_{i=0}^{k-1} d(p, q_i)^2 \right)^{\frac{1}{2}}
    \]
Density Estimation

- A field of study in itself

- Gaussian kernel
  - Bandwidth parameter $h$

$$
\hat{f}(p) = \frac{1}{|\mathcal{N}_p|} \sum_{q \in \mathcal{N}_p} e^{-\frac{d(p-q)^2}{2h}}
$$

- Distance to a measure
  - Number of neighbors

$$
\hat{f}(p) = \left( \frac{1}{k} \sum_{i=0}^{k-1} d(p,q_i)^2 \right)^{\frac{1}{2}}
$$

Require $\|f - \hat{f}\|_\infty \leq \epsilon$
Assumptions: $\mathbb{X}$ Riemannian manifold, $f: \mathbb{X} \to \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

With an estimate of pairwise distances, we can build a Rips graph
Assumptions: \( \mathbb{X} \) Riemannian manifold, \( f : \mathbb{X} \to \mathbb{R} \) \( c \)-Lipschitz, 
\( L \) geodesic \( \varepsilon \)-cover of \( \mathbb{X} \), for some unknown \( \varepsilon > 0 \).

With an estimate of pairwise distances, we can build a Rips graph.

Guarantee:

The super-level set filtration and the (image)-Rips filtration 
\( \forall \delta \geq \varepsilon \) are \( 2c\delta \)-interleaved

\[ \downarrow \]

[Chazal, Cohen-Steiner, Glisse, Guibas, Oudot ’09]

their PD’s are \( 2c\delta \)-close.
Assumptions: \( \mathbb{X} \) Riemannian manifold, \( f : \mathbb{X} \rightarrow \mathbb{R} \) \( c \)-Lipschitz, 
\( L \) geodesic \( \varepsilon \)-cover of \( \mathbb{X} \), for some unknown \( \varepsilon > 0 \).

With an estimate of pairwise distances, we can build a Rips graph

Guarantee:

The super-level set filtration and the (image)-Rips filtration 
\( \forall \delta \geq \varepsilon \) are \( 2c\delta \)-interleaved

\[ \downarrow \]

[Chazal, Cohen-Steiner, Glisse, Guibas, Oudot '09]

their PD’s are \( 2c\delta \)-close.

Under sufficient sampling, we get a good approximation of the PD
Sampling

- Whole space is **not** uniformly sampled
  Approximation depends on $c\delta$
Sampling

- Whole space is **not** uniformly sampled
  Approximation depends on $c\delta$
Sampling

- Whole space is **not** uniformly sampled.
  Approximation depends on $c\delta$

- Approximation result holds in well-sampled regions w.h.p.
Sampling

- Whole space is **not** uniformly sampled
  Approximation depends on $c\delta$

- Approximation result holds in well-sampled regions w.h.p.

- Rips parameter $\delta =$ spatial scale
  - Trade-off

  Small $\delta =$ good approximation
  Large $\delta =$ holds over a larger part of the space
Sampling

- Whole space is **not** uniformly sampled
  - Approximation depends on $c\delta$

- Approximation result holds in well-sampled regions w.h.p.

- Rips parameter $\delta = \text{spatial scale}$
  - Trade-off
    - Small $\delta$ = good approximation
    - Large $\delta$ = holds over a larger part of the space
Sampling

- Whole space is **not** uniformly sampled
  Approximation depends on $c\delta$

- Approximation result holds in well-sampled regions w.h.p.

- Rips parameter $\delta = \text{spatial scale}$
  - Trade-off
    Small $\delta$ = good approximation
    Large $\delta$ = holds over a larger part of the space
Well-Separated Diagrams

- Define a *signal-to-noise* ratio
Number of Clusters

- Well-separated diagram has a stable number of clusters
  - Geometric proof
Feedback & Interpreting Diagrams

• If peaks are prominent enough, we will get the “right” number of clusters

• Theoretically,
  - The number of clusters is stable
  - The more samples, the smaller the noise
If peaks are prominent enough, we will get the “right” number of clusters.

Theoretically,

- The number of clusters is stable
- The more samples, the smaller the noise
Feedback & Interpreting Diagrams

- If peaks are prominent enough, we will get the “right” number of clusters

- Theoretically,
  - The number of clusters is stable
  - The more samples, the smaller the noise

- Practically,
  - Gives a sense of stability of the number of clusters
  - Choice of threshold transparent w.r.t. number of clusters
  - Can help with the choice of other parameters
Spatial Stability

- Can we say anything about the clusters themselves?
Spatial Stability

- Can we say anything about the clusters themselves?
  
  **Idea:** Prominent clusters have a minimum size under $c$-Lipschitz assumption
Spatial Stability

- Can we say anything about the clusters themselves?
  **Idea:** Prominent clusters have a minimum size under $c$-Lipschitz assumption

- Under small perturbations, prominent peak must be part of the “same” cluster
Spatial Stability

• Can we say anything about the clusters themselves?
  **Idea:** Prominent clusters have a minimum size under $c$-Lipschitz assumption

• Under small perturbations, prominent peak must be part of the “same” cluster
Can we say anything about the clusters themselves?

**Idea:** Prominent clusters have a minimum size under $c$-Lipschitz assumption

- Under small pertubations, prominent peak must be part of the “same” cluster
Spatial Stability

- Can we say anything about the clusters themselves?
  **Idea:** Prominent clusters have a minimum size under $c$-Lipschitz assumption

- Under small perturbations, prominent peak must be part of the “same” cluster
• Unstable part can be arbitrarily large
Spatial Stability

- Unstable part can be arbitrarily large
Results

- 3 datasets
  - 2 spirals in $\mathbb{R}^2$
  - 4 rings in $\mathbb{R}^3$
  - Alanine-dipeptide
• 2-dimensional example
Spirals
Spirals
Spirals
4 Rings

- Interlocking rings in $\mathbb{R}^3$
- 12K points total
4 Rings
4 Rings
Poorly Sampled 4 Rings
Biological Dataset

- Protein conformations
Biological Dataset

- Protein conformations

- Simulations
  - Short trajectories (ps)
  - Interest is in longer time scale (ms)
Biological Dataset

- Protein conformations

- Simulations
  - Short trajectories (ps)
  - Interest is in longer time scale (ms)

- Goal: find stable states
Biological Dataset

- Protein conformations

- Simulations
  - Short trajectories (ps)
  - Interest is in longer time scale (ms)

- Goal: find stable states

- Alanine-dipeptide
  - 200K conformations in $\mathbb{R}^{22}$
  - Non-Euclidean metric (RMSD over best rigid transforms)
Alanine-dipeptide
Alanine-dipeptide
Alanine-dipeptide

<table>
<thead>
<tr>
<th>Rank</th>
<th>(Unnormalized) Prominence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>∞</td>
</tr>
<tr>
<td>2</td>
<td>5677</td>
</tr>
<tr>
<td>3</td>
<td>3828</td>
</tr>
<tr>
<td>4</td>
<td>1335</td>
</tr>
<tr>
<td>5</td>
<td>850</td>
</tr>
<tr>
<td>6</td>
<td>316</td>
</tr>
<tr>
<td>7</td>
<td>258</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>22</td>
</tr>
</tbody>
</table>
Take-Home Message

- Practical clustering algorithm (efficient in space and time)
  - works in any metric space

- Generic algorithm
  - choice of graph
  - choice of estimator
Take-Home Message

• Practical clustering algorithm (efficient in space and time)
  - works in any metric space

• Generic algorithm
  - choice of graph
  - choice of estimator

• Persistence diagram is the structure of the function
Take-Home Message

- Practical clustering algorithm (efficient in space and time)
  - works in any metric space
- Generic algorithm
  - choice of graph
  - choice of estimator
- Persistence diagram is the structure of the function
- Theoretical guarantees
  - Number of clusters
  - Spatial stability
- Many interesting questions....
Current Work

- Soft-clustering
Current Work

- Soft-clustering

- Higher-dimensional features
Current Work

- Soft-clustering
- Higher-dimensional features
- Other interesting functions
Thank You

- Code available: http://geometrica.saclay.inria.fr/data/ToMATo/

- Questions?