July 9, 2009

TGDA Workshop

Persistence based Clustering

Primož Skraba

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

- Input samples
Clustering

• Input samples

• "Important" segments/clusters
Clustering

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

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

- Input samples

- "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 “significant” peaks of $f$
Definition of a Cluster

• Basins of attraction of “significant” peaks of $f$
Definition of a Cluster

- Basins of attraction of “significant” peaks of $f$
Outline

- Background: scalar field analysis
- Algorithm
- Number of clusters
- Results (Interpretation of persistence diagrams)
- Spatial stability
- Conclusions
Scalar Field Analysis*

Setting: $X$ topological space, $f : X \to \mathbb{R}$

Input: A finite sampling $L$ of $X$, the values of $f$ at the sample points

*[Chazal, Guibas, Oudot, Skraba '09]
Scalar Field Analysis*

**Setting:** $\mathbb{X}$ topological space, $f : \mathbb{X} \to \mathbb{R}$

**Input:** A finite sampling $L$ of $\mathbb{X}$, the values of $f$ at the sample points

*[Chazal, Guibas, Oudot, Skraba '09]*
Scalar Field Analysis*

**Setting:** \( \mathbb{X} \) topological space, \( f : \mathbb{X} \rightarrow \mathbb{R} \)

**Input:** A finite sampling \( L \) of \( \mathbb{X} \), the values of \( f \) at the sample points

**Goal:** Analyze landscape of graph(\( f \)):  
- *prominent* peaks/valleys  
- basins of attraction

*[Chazal, Guibas, Oudot, Skraba '09]*
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 \( \hat{f}^{-1}([\alpha, \infty)) \) 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 $\mathbb{R}$.
- 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]
Persistence-Based Approach

**Assumptions:** \( \mathbb{X} \) triangulated space, \( f : \mathbb{X} \to \mathbb{R} \) Lipschitz continuous

→ build PL approximation \( \hat{f} \) of \( f \)

→ apply persistence algo. to \( \pm \hat{f} \) [Edelsbrunner, Letscher, Zomorodian '00]

\[ \beta_0 \]

(6 prominent peaks)
Persistence-Based Approach

Assumptions: \( \mathbb{X} \) triangulated space, \( f : \mathbb{X} \rightarrow \mathbb{R} \) Lipschitz continuous

\( \rightarrow \) build PL approximation \( \hat{f} \) of \( f \)

\( \rightarrow \) apply persistence algo. to \( \pm \hat{f} \) [Edelsbrunner, Letscher, Zomorodian ’00]

\( \beta_0 \) (6 prominent peaks) \( \beta_1 \) (ring-shaped basin of attraction)
Approximation of Super-Level Sets

**Assumptions:** $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \rightarrow \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$. 
Approximation of Super-Level Sets

**Assumptions:** $\mathbb{X}$ Riemannian manifold, $f : \mathbb{X} \to \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

- Access to $L$ **not** $\mathbb{X}$
Approximation of Super-Level Sets

**Assumptions:** $X$ Riemannian manifold, $f: X \rightarrow \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $X$, for some unknown $\varepsilon > 0$.

- Access to $L$ **not** $X$

$$
\begin{align*}
F^\alpha &:= f^{-1}([\alpha, \infty)) \\
L_\alpha &:= L \cap F^\alpha \\
L^\varepsilon_\alpha &:= \bigcup_{p \in L_\alpha} B_X(p, \varepsilon)
\end{align*}
$$

\begin{array}{|l|}
\hline
\forall \alpha \in \mathbb{R}, \quad L^\varepsilon_{\alpha + c\varepsilon} \subseteq F^\alpha \subseteq L^\varepsilon_{\alpha - c\varepsilon} \\
\hline
\end{array}
Approximation of Super-Level Sets

**Assumptions:** $\mathbb{X}$ Riemannian manifold, $f: \mathbb{X} \to \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $\mathbb{X}$, for some unknown $\varepsilon > 0$.

- Access to $L$ **not** $\mathbb{X}$

\[
\begin{align*}
F^\alpha &:= f^{-1}([\alpha, \infty)) \\
L_\alpha &:= L \cap F^\alpha \\
L_\alpha^\varepsilon &:= \bigcup_{p \in L_\alpha} B_\mathbb{X}(p, \varepsilon)
\end{align*}
\]

$\forall \alpha \in \mathbb{R}, L_\alpha^\varepsilon + c\varepsilon \subseteq F^\alpha \subseteq L_\alpha^\varepsilon - c\varepsilon$
Approximation of Super-Level Sets

**Assumptions:** \( X \) Riemannian manifold, \( f : X \to \mathbb{R} \) \( c \)-Lipschitz, 
\( L \) geodesic \( \varepsilon \)-cover of \( X \), for some unknown \( \varepsilon > 0 \).

- Access to \( L \) **not** \( X \)

\[
\begin{align*}
F^\alpha &:= f^{-1}([\alpha, \infty)) \\
L_\alpha &:= L \cap F^\alpha \\
L_\alpha^\varepsilon &:= \bigcup_{p \in L_\alpha} B_X(p, \varepsilon)
\end{align*}
\]

\( \forall \alpha \in \mathbb{R}, \ L_\alpha^\varepsilon + c\varepsilon \subseteq F^\alpha \subseteq L_\alpha^\varepsilon - c\varepsilon \)
Approximation of Super-Level Sets

**Assumptions:** \( \mathbb{X} \) Riemannian manifold, \( f : \mathbb{X} \rightarrow \mathbb{R} \) \( c \)-Lipschitz, 
\( L \) geodesic \( \varepsilon \)-cover of \( \mathbb{X} \), for some unknown \( \varepsilon > 0 \).

- Access to \( L \) **not** \( \mathbb{X} \)

\[
\begin{align*}
F^\alpha &: = f^{-1}([\alpha, \infty)) \\
L_\alpha &: = L \cap F^\alpha \\
L^\varepsilon_\alpha &: = \bigcup_{p \in L_\alpha} B_\mathbb{X}(p, \varepsilon)
\end{align*}
\]

\( \forall \alpha \in \mathbb{R}, \ L^\varepsilon_{\alpha+c\varepsilon} \subseteq F^\alpha \subseteq L^\varepsilon_{\alpha-c\varepsilon} \)

the filtrations \( \{ F^\alpha \}_{\alpha \in \mathbb{R}} \) and \( \{ L^\varepsilon_\alpha \}_{\alpha \in \mathbb{R}} \) are \( c\varepsilon \)-interleaved

\( \Downarrow \)

their barcodes are \( c\varepsilon \)-close.

[Chazal, Cohen-Steiner, Glisse, Guibas, Oudot '09]
Approximation of Super-Level Sets

**Assumptions:** \( \mathbb{X} \) Riemannian manifold, \( f : \mathbb{X} \to \mathbb{R} \) \( c \)-Lipschitz, \( L \) geodesic \( \varepsilon \)-cover of \( \mathbb{X} \), for some unknown \( \varepsilon > 0 \).

**Guarantee:**

\[ \forall \delta \geq \varepsilon, \{ F_\alpha \}_{\alpha \in \mathbb{R}} \text{ and } \{ R^\delta (L_\alpha) \hookrightarrow R^{2\delta} (L_\alpha) \}_{\alpha \in \mathbb{R}} \text{ are } 2c\delta \text{-interleaved} \]

\[ \Downarrow \]

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

their barcodes are \( 2c\delta \)-close.
Approximation of Super-Level Sets

**Assumptions:** \( \mathcal{X} \) Riemannian manifold, \( f : \mathcal{X} \to \mathbb{R} \) \( c \)-Lipschitz, \( L \) geodesic \( \epsilon \)-cover of \( \mathcal{X} \), for some unknown \( \epsilon > 0 \).

(6 prominent peaks) (ring-shaped basin of attraction)
Approximation of Super-Level Sets

**Assumptions:** \( \mathbb{X} \) Riemannian manifold, \( f : \mathbb{X} \rightarrow \mathbb{R} \) \( c \)-Lipschitz, \( L \) geodesic \( \varepsilon \)-cover of \( \mathbb{X} \), for some unknown \( \varepsilon > 0 \).
Approximation of Super-Level Sets

**Assumptions:** $X$ Riemannian manifold, $f : X \rightarrow \mathbb{R}$ $c$-Lipschitz, $L$ geodesic $\varepsilon$-cover of $X$, for some unknown $\varepsilon > 0$. 
Homological Features and Clusters

- Samples drawn from $f$
- Estimate $\hat{f}$ from samples
Homological Features and Clusters

- Samples drawn from $f$
- Estimate $\hat{f}$ from samples

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

How do we compute clusters from a barcode?
Computing Clusters

How do we compute clusters from a barcode?

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

How do we compute clusters from a barcode?

**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 barcode?

**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 barcode?

**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: $f(x), \mathcal{R}_\delta, \alpha$
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
Theoretical Guarantees

- Applying the result from scalar field work
Theoretical Guarantees

- Applying the result from scalar field work
  Approximation depends on $c\delta$
Theoretical Guarantees

- Applying the result from scalar field work
  Approximation depends on $c\delta$
  Whole space is **not** uniformly sampled
Theoretical Guarantees

- Applying the result from scalar field work
  Approximation depends on \( c\delta \)
  Whole space is \textbf{not} uniformly sampled

- Approximation result holds in well-sampled regions \( \text{w.h.p.} \)
Theoretical Guarantees

- Applying the result from scalar field work
  Approximation depends on $c\delta$
  Whole space is not uniformly sampled

- Approximation result holds in well-sampled regions w.h.p.
- More points $\Rightarrow$ more of the space
Number of Clusters

- Define a *signal-to-noise* ratio

**Definition:** Given two values $d_2 > d_1 \geq 0$, the persistence diagram $D_0f$ is called $(d_1, d_2)$-separated if every point of $D_0f$ lies either in the region $D_1$ above the diagonal line $y = x - d_1$ or in the region $D_2$ below the diagonal $y = x - d_2$ and to the right of the vertical line $x = d_2$. 
Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.
Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.
Approximation

- Assume enough points that up to $c\delta$ is well-sampled w.h.p.
Feedback and Interpreting Diagrams

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

• Practically,
  - Gives a sense of stability of the number of clusters
  - Choice of threshold transparent w.r.t. number of clusters

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

- Synthetic dataset
- Image segmentation
- Alanine-dipeptide conformations
4 Rings

- Interlocking rings in $\mathbb{R}^3$
- 600k (100k + 500k) points total
4 Rings
Image Segmentation

- Each pixel is assigned color coordinates in LUV space
Landscape
Koala
Incorporating Spatial Information

- Neighborhood graph: proximity in LUV space and image
Alanine-dipeptide Conformations

- Clustering in 22-dim space
- 192k points
Alanine-dipeptide Conformations

![Graph showing the relationship between metastability and number of clusters.]

<table>
<thead>
<tr>
<th>Rank</th>
<th>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>
Spatial stability

- Number of clusters are correct
- Can we say anything about the clusters themselves?
  1. Each prominent cluster has a stable part
  2. Unstable part can be very large
Spatial stability

- Number of clusters are correct
- Can we say anything about the clusters themselves?

1. Each prominent cluster has a stable part
2. Unstable part can be very large
Spatial stability

- Number of clusters are correct
- Can we say anything about the clusters themselves?
  1. Each prominent cluster has a stable part
  2. Unstable part can be very large
Stable Part

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

- Under small pertubations, prominent peak part of the “same” cluster

- Soft clustering
  1. Run the algorithm multiple times, with small pertubations
  2. Find one-to-one correspondance between clusters
  3. Find stable and unstable parts
Conclusions

• Practical clustering algorithm (efficient in space and time)

• General framework
  - Use your favorite density estimator
  - Choice of neighborhood graph

• Easily-interpreted feedback
  - No “black box” effect

• Theoretical guarantees
  - Number of clusters
  - Spatial stability

• Soft-clustering

• Higher-dimensional features