Kernel Methods for Topological Data Analysis

Kenji Fukumizu
The Institute of Statistical Mathematics (Tokyo, Japan)
Joint work with Genki Kusano and Yasuaki Hiraoka (Tohoku Univ.), supported by JST CREST

STM2016 at ISM. July 22, 2016
Topological Data Analysis

• TDA: a new method for extracting topological or geometrical information of data.

Key technology = Persistence homology
(Edelsbrunner et al 2002; Carlsson 2005)

Background
• Complex data:
  Data with complex structure must be analyzed.
• Progress of computational topology:
  Computing topological invariants becomes easy.
TDA: Various applications

Data of highly complex geometric structure

Often difficult to define good feature vectors / descriptors

Brain Science

Brain artery trees
  e.g. age effect
  (Bendich et al 2014)

Structure change of proteins
  eg. open / closed
  (Kovacev-Nikolic et al 2015)

Material Science

Liquid

Glass

Shape signature,
  natural image statistics
  (Freedman & Chen 2009)

Persistence homology provides a compact representation for such data.

Computer Vision

Non-crystal materials
  (Nakamura, Hiraoka, Hirata, Escolar, Nishiura. 
  Nanotechnology 26 (2015))

Biochemistry

etc...
Outline

• A brief introduction to persistence homology

• Statistical approach with kernels to topological data analysis

• Applications
  • Material science
  • Protein classification

• Summary
Topology

\[ \cong \]
**Topology:** two sets are equivalent if one is deformed to the other without tearing or attaching.

**Topological invariants:** any equivalent sets take the same value.

<table>
<thead>
<tr>
<th>Connected components</th>
<th>Ring</th>
<th>Cavity</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram 1" /></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><img src="image2.png" alt="Diagram 2" /></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><img src="image3.png" alt="Diagram 3" /></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><img src="image4.png" alt="Diagram 4" /></td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Algebraic Topology

- Algebraic treatment of topological spaces

Compute various topological invariances e.g. Euler number

Classify topological spaces with topological invariances.
• Homology group: independent “holes”

\[ H_k(X): k\text{-th homology group} \text{ of topological space } X \]  
\[ (k = 0, 1, 2, \ldots) \]

\( k \)-dimensional holes

\( H_0(X) \): connected components
\( H_1(X) \): rings
\( H_2(X) \): cavities

...
Topology of statistical data?

True structure

$\varepsilon$-balls (e.g. manifold learning)

Noisy finite sample

Small $\varepsilon \rightarrow$ disconnected object

Stable extraction of topology is NOT easy!

Large $\varepsilon \rightarrow$ small ring is not visible
Persistence Homology

• All $\varepsilon$ considered
  
  $X = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$,  \quad X_\varepsilon := \bigcup_{i=1}^m B_\varepsilon(x_i)$

Two rings （generators of 1 dim homology）
persist in a long interval.
• Persistence homology (formal definition)
   Filtration of topological spaces $\mathcal{X}: X_1 \subset X_2 \subset \cdots \subset X_L$

   $PH_k(\mathcal{X}): H_k(X_1) \to H_k(X_2) \to \cdots \to H_k(X_L) \cong \bigoplus_{i=1}^{m_k} I[b_i, d_i]$  \hspace{1cm} \text{Irreducible decomposition}

   $I[b, d] \cong 0 \to \cdots \to 0 \to K \to \cdots \to K \to 0 \to \cdots \to 0 \quad K: \text{field}$

   The lifetime (birth, death) of each generator is rigorously defined, and can be computed numerically.
• Two popular (equivalent) expressions of PH

**Barcode**

Bar from the birth to death of each generator

**Persistence diagram (PD)**

Plots of the birth (b) and death (d) of each generator of PH in a 2D graph \((d \geq b)\).

**Handy descriptors or features of complex geometric objects**

Barcodes and PD are considered for each dimension.
Beyond topology

• PH contains geometrical information more than topology
Statistical approach with kernels to topological data analysis
Statistical approach to TDA

• **Conventional TDA**

**Data**  
**Computation of PH**  
**Visualization (PD)**  
**Analysis by experts**

- e.g. Molecular dynamics simulation
- Software
  - CGAL / PHAT

PHAT: Persistent Homology Algorithm Toolbox  [https://bitbucket.org/phat-code/phat](https://bitbucket.org/phat-code/phat)
• Statistical approach to TDA
  (Kusano, Fukumizu, Hiraoka ICML 2016; Reininghaus et al CVPR 2015; Kwitt et al NIPS2015; Fasy et al 2014)

Many data sets

Many PD’s

Computation of PH

But how?

Statistical analysis of PD’s

Features / Descriptors
Kernel representation of PD

• Vectorization of PD by positive definite kernel
  
  • PD = Discrete measure \( \mu_D := \sum_{z \in PD} \delta_z \)
  
  • Kernel embedding of PD’s into RKHS
    \[ \mathcal{E}_k: \mu_D \mapsto \int k(\cdot, x) d\mu_D(x) = \sum_i k(\cdot, x_i) \in H_k, \]

  • For some kernels (e.g., Gaussian, Laplace), \( \mathcal{E}_k \) is injective.

  • By vectorization,
    • a number of methods for data analysis can be applied,
      SVM, regression, PCA, CCA, etc.
    • tractable computation is possible with kernel trick.

\( k \): positive definite kernel
\( H_k \): corresponding RKHS
Persistence Weighted Gaussian (PWG) Kernel

Generators close to the diagonal may be noise, and should be discounted.

\[ k_{PWG}(x, y) = w(x)w(y) \exp \left( -\frac{||y-x||^2}{2\sigma^2} \right) \]

\[ w(x) = w_{c,p}(x) := \arctan(C \text{Pers}(x)^p) \quad (C, p > 0) \]

\[ \text{Pers}(x) := d - b \text{ for } x \in \{(b, d) \in \mathbb{R}^2 | d \geq b\} \]
• Stability with PWG kernel embedding
  • PWGK defines a distance on the persistence diagrams,

\[
d_k(D_1, D_2) := \| \varepsilon_k(D_1) - \varepsilon_k(D_2) \|_{H_k}, \quad D_1, D_2: \text{persistence diagrams}
\]

**Stability Theorem** (Kusano, Hiraoka, Fukumizu 2015)

\( M \): compact subset in \( \mathbb{R}^d \). \( S \subset M, T \subset \mathbb{R}^d \): finite sets.

If \( p > d + 1 \), then with PWG kernel \((p, C, \sigma)\),

\[
d_k(D_q(S), D_q(T)) \leq L \, d_H(S, T).
\]

\( L \): constant depending only on \( M, p, d, C, \sigma \)

\( D_q(S) \): \( q \) th persistence diagram of \( S \)

\( d_H \): Haussdorff distance

A small change of a set causes only a small change in PD

Lipschitz continuity

This stability is NOT known for Gaussian kernel.
2nd-level kernel (SVM for measures, Muandet, Fukumizu, Dinuzzo, Schölkopf 2012)

- RKHS-Gaussian kernel
  \[ K(\phi_1, \phi_2) = \exp \left( -\frac{\| \phi_1 - \phi_2 \|^2_{H_k}}{2\tau^2} \right) \]
  derives

- For Persistence diagrams \( D_i, D_j \):
  \[ K(D_i, D_j) = \exp \left( -\frac{\| \mathcal{E}_k(D_i) - \mathcal{E}_k(D_j) \|^2_{H_k}}{2\tau^2} \right) \]
Computational issue

The number of generators in a PD may be large ($\geq 10^3, 10^4$)

For $PD_i = \sum_{a=1}^{N_i} \delta_{x_a(i)} \cup \Delta$, $K(PD_i, PD_j) = \exp\left(-\frac{\|\varepsilon_k(PD_i) - \varepsilon_k(PD_j)\|_{H_k}^2}{2\tau^2}\right)$ requires computation

$$\|\varepsilon_k(PD_i) - \varepsilon_k(PD_j)\|_{H_k}^2 = \sum_{a=1}^{N_i} \sum_{b=1}^{N_i} k(x_a(i), x_b(i)) + \sum_{a=1}^{N_j} \sum_{b=1}^{N_j} k(x_a(j), x_b(j)) - 2 \sum_{a=1}^{N_i} \sum_{b=1}^{N_j} k(x_a(i), x_b(j)).$$

The number of $\exp\left(-\frac{\|x_a - x_b\|^2}{2\sigma^2}\right) = O(m^2N^2) \rightarrow$ computationally expensive for $N \approx 10^4$

$$N = \max\{N_i | i = 1, \ldots, n\}$$
• Approximation by random features (Rahimi & Recht 2008)

By Bochner’s theorem
\[
\exp\left(-\frac{\|x_a-x_b\|^2}{2\sigma^2}\right) = C \int e^{\sqrt{-1}\omega^T (x_a-x_b)} \left(\frac{\sigma^2}{2\pi}\right) e^{-\frac{\sigma^2\|\omega\|^2}{2}} d\omega
\]

(Fourier transform)

Approximation by sampling: \(\omega_1, \ldots, \omega_L\): i. i. d. \(\sim Q_\sigma\)
\[
\exp\left(-\frac{\|x_a-x_b\|^2}{2\sigma^2}\right) \approx C \frac{1}{L} \sum_{\ell=1}^L e^{\sqrt{-1}\omega_\ell^T x_a} e^{\sqrt{-1}\omega_\ell^T x_b}
\]

\[
\sum_{a=1}^{N_i} \sum_{b=1}^{N_j} k(x_a^{(i)}, x_b^{(j)}) \approx \frac{C}{L} \sum_{a=1}^{N_i} \sum_{b=1}^{N_j} \sum_{\ell=1}^L w(x_a^{(i)}) w(x_b^{(j)}) e^{\sqrt{-1}\omega_\ell^T x_a^{(i)}} e^{\sqrt{-1}\omega_\ell^T x_b^{(j)}}
\]

\[
= \frac{C}{L} \sum_{\ell=1}^L \sum_{a=1}^{N_i} w(x_a^{(i)}) e^{\sqrt{-1}\omega_\ell^T x_a^{(i)}} \sum_{b=1}^{N_j} w(x_b^{(j)}) e^{\sqrt{-1}\omega_\ell^T x_b^{(j)}}
\]

\(L\) dim.

Computational cost \(O(LN)\) \(\rightarrow\) 2nd level Gram matrix \(O(mLN + m^2L)\). c.f. \(O(m^2N^2)\)

Big reduction if \(L, n \ll N\)
Comparison: Persistence Scale Space Kernel

(Reininghaus et al 2015)

• PSS Kernel
  \[ k_R(x, y) = \frac{1}{8\pi t} \left\{ \exp\left( \frac{||x - y||^2}{8t} \right) - \exp\left( \frac{||x - \bar{y}||^2}{8t} \right) \right\} \]
  \[ \bar{y} = (d, b) \text{ for } y = (b, d). \]

\( \mathcal{E}_k(D) \) is considered.

• Comparison between PWGK and PSSK
  • PWGK can control the discount around the diagonal independently of the bandwidth parameter.
  • PSSK is not shift-invariant \( \rightarrow \) Random feature approximation is not applicable.
  • In Reininghaus et al 2015, 2nd level kernel is not considered.
Synthetic example: SVM classification

• Classification of PD’s by SVM
  • One big circle (random location and sample size) $S_1$ with or without small circle $S_0$.
  • $Y = \text{XOR}(Z_1, Z_2)$
    • $Z_1$: Does $S_0$ exists? Yes/No
    • $Z_2$: Is the generator of $S_1$ within ($(b(S_1)<1 && d(S_1))$? Yes/No
  • Noise is added, in fact.
  • 100 for training and 100 for testing

• Result (correct classification)
  • PWGK (proposed): 83.8%
  • PSSK (comparison): 46.5%
Applications
Application 1: Transition of Silica (SiO$_2$)

If cooled down rapidly from the liquid state, SiO$_2$ changes into the glass state (not to crystal).

Goal: identify the temperature of phase transition.

Data: Molecular Dynamics simulation for SiO$_2$. 3D arrangements of the atoms are used for computing PD at 80 temperatures. (Nakamura et al 2015; Hiraoka et al 2015)

Examples of PD's

Liquid

Glass (Amorphous)
Change point detection

• Data along a parameter $t$

$$X_t, t = 1, \ldots, T.$$  

Kernel Change Point Analysis with Fisher Discriminant score (Harchoui et al 2009):

For each $t$, two classes are defined by the data before and after $t$.
Fisher score on RKHS is used.

• For each $t$, compute

$$\hat{m}_{1:t} = \frac{1}{t} \sum_{i=1}^{t} \Phi(X_i) \quad \text{and} \quad \hat{m}_{t+1:T} = \frac{1}{T-t} \sum_{i=t+1}^{T} \Phi(X_i).$$

• Compute

$$\Delta_t := \left\| (V_{1:t} + V_{t+1:T} + \gamma I)^{-\frac{1}{2}} (\hat{m}_{1:t} - \hat{m}_{t+1:T}) \right\|_{H_k}^2.$$  

• Find $\max_t \Delta_t$.

• For the packing problem, $X_t = \mathcal{E}_k(D_{\phi_t}) (t = 1, \ldots, 80)$.  

• Detection of liquid-glass state transition
  • Approach in physics:
    Estimation using derivatives of enthalpy curve, but not so accurate.
  • Our approach: purely data-driven
    Persistence diagrams, and then change point detection by Kernel FDR.
  • Number of generators in a PD is 30000 at most → difficult to use PSSK directly
  • PWGK (proposed) is applied with random features.
Detected change point = 3100K
Enthalpy by physicist: [2000K, 3500K]
• 2-dim plot by Kernel PCA

Sharp change between the two phases.
(Colored by the result of change point detection. Colors are not used for KPCA).

The result indicates that the phase can be identified by the snap-shot, while this is still controversial among physicists.
Application 2: Protein classification

• Structure of proteins $\rightarrow$ Functions

• The geometrical structure can be represented by persistence homology

• Classification of proteins with PD’s. SVM is used.
• Data A: Protein-drug binding

• M2 channel in the influenza A virus: a target of medicine. Biding an inhibitor changes the structure

  Cang, Mu, Wu, Opron, Xia, Wei, Molecular Based Mathematical Biology (2015) Fig. 3

• Task: Determine from the structure if there is rimantadine (inhibitor) in the M2 channel.

• Data: 3D-structures from NMR
  • 15 data for each of binding / non-binding.
  • Random choice of 10 training samples for each class. The rest is used for testing. 100 random choices for CV.
• Data B: 2 states of hemoglobin

  • Task: classify of the 2 states Relaxed (R) / Taut (T)

  • Data: 3D-structures from X-ray diffraction
    • R: 9 data, T: 10 data

    • Choice of one data from each class for testing, and the rest used for training.
    • All combinations are used for CV.

Relaxed (R)  Taut (T)

Cang, Mu, Wu, Opron, Xia, Wei, *Molecular Based Mathematical Biology* (2015) Fig. 4
• Results
  • Comparison with Cang et al (2015), where PH is used with 13 dimensional hand-made Molecular Topological Fingerprint (MTF).
  • PWGK + SVM: only 1st PH is used.

MTF

<table>
<thead>
<tr>
<th>#</th>
<th>Dim</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2nd longest lifetime</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3rd longest lifetime</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>Total sum of lifetime</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>Average lifetime</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Birth point of the longest generator</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>Longest lifetime</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>Birth points of the shortest generator among lifetime ≥1.5Å</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>Ave. medium points of generators among lifetime ≥1.5Å</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>Number of generators in [4.5, 5.5]Å, divided by total #atoms.</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>Number of generators in [3.5, 4.5]Å and (5.5, 6.5]Å, divided by total #atoms.</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>Total sum of lifetimes</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>Average lifetime</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>The birth point of the first generator.</td>
</tr>
</tbody>
</table>

CV classification rates

<table>
<thead>
<tr>
<th></th>
<th>A. Protein-Drug</th>
<th>B. Hemoglobin</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWGK</td>
<td>100</td>
<td>88.90</td>
</tr>
<tr>
<td>MTF*</td>
<td>(nbd) 93.91 / (bd) 98.31</td>
<td>84.50</td>
</tr>
</tbody>
</table>

* Results of MTF are taken from Cang et al. Molecular Based Mathematical Biology (2015).
Conclusion

• Topological data analysis
  • Key technology = persistence homology
  • PH can introduce useful features / descriptors for complex geometrical structures.
  • PH contains information more than topology.

• Statistical approach to topological data analysis
  • Statistical data analysis on many persistence diagrams.
  • Kernel methods introduce systematic data analysis to TDA.
    • Vectorization of persistence diagrams by kernel embedding.
    • Persistence weighted Gaussian kernel $\rightarrow$ flexible kernel for noise.
References


