パーシステントホモロジーの機械学習

草野　元紀　(Genki Kusano)*

本研究の概要

本講演では点集合のトポロジーを考える. ただ単に点集合のホモロジー群を考えても面白くはないが, パーシステントホモロジー [ZC05] というものを考えると点集合の特徴的なトポロジーを記述する事ができる. パーシステントホモロジーの材料科学での応用 [HNH$^+$16] では, シリカ (SiO_2) という物質の原子配置を点集合として考える. シリカは固体液体の他に, ガラス状態という特殊な状態になる事がありその構造解析が材料科学の応用での目標になる. 固体状態であれば結晶構造になるが, 温度を上げて液体やガラス状態になると, 熱揺らぎの影響も受けて図 1 のように一見無秩序な配置に見える.

図 1: 液体状態（左）とガラス状態（右）のシリカの原子配置.

材料科学者の目には液体は確かに無秩序だが, ガラスは何かしらの規則・幾何構造があるのではという考察がなされていた. しかし, その幾何構造は名前のつくほど知られたものでもなければ, どのように特徴付けて良いか分からずシリカの液体ガラス相転移点付近での構造解析は未解決問題であった. [HNH$^+$16] では, 原子配置からパーシステントホモロジーを計算し, それをパーシステント図として可視化する事でガラスの幾何構造を解析している. 図 2 を見ると, 液体とガラスのパーシステント図は違うため, 液体とガラスの幾何構造は違うと結論付けられる.

図 2: 液体状態（左）とガラス状態（右）のシリカから作られるパーシステント図.

筆者の研究は, “（データの幾何形状を表現した）パーシステント図を見てその違いを判別する” という恣意的な作業を “**パーシステント図に対する統計解析・機械学習**[1]**の手法を確立して, その手法を用いて定量的に違いを評価する**” に変える事である. ここでは, パーシステント図に対する統計解析・機械学習手法としてカーネル法を用いる. 一般に, データが非ベクトルデータの場合はその平均などの統計量を直接求める事がで

本研究は研究は JSPS 科研費 JP17J02401 の助成を受けたものである.

* 〒 980-8578 宮城県仙台市青葉区荒巻字青葉 6-3 東北大学大学院理学研究科数学専攻博士課程 2 年
e-mail: genki.kusano.r5@dc.tohoku.ac.jp

[1] 統計と機械学習の違いに関する私的意見としては, 統計は既にあるデータからその特徴を説明することを目的にし, 機械学習は手元にあるデータから未来のデータを予測することを目的にしている.

きない. 一方で, 任意のデータがベクトルの構造を持っているとも限らず, パーシステント図もベクトルとして扱うことはできない. カーネル法は任意のデータをベクトルに変換する手法であり, 数学的にはデータを無限次元ヒルベルト空間の元に変換するものの, 計算機上での実装が容易にできる点で機械学習でよく用いられる. 筆者はパーシステント図に対するカーネル法として, 福水健次氏（統数研）, 平岡裕章氏（東北大）との共同研究でPersistence weighted Gaussian kernel (PWGK, [KFH16, KFH17]) を提案して, 数理的性質の研究とデータ解析を行っている. 材料科学への応用では液体ガラス転移点問題を主成分分析や変化点検出により特徴付け, タンパク質の構造解析ではサポートベクターマシンによる二値分類を行った. 本講演ではパーシステントホモロジー, パーシステント図のカーネル法の構成法, 数値計算結果を紹介する.

本稿の残りは, 以下のキーワードの解説に当てる.

1章 パーシステントホモロジー・パーシステント図.

2章 Persistence weighted Gaussian kernel.

カーネル法に関しては [福10] を参照されたい.

1. パーシステントホモロジー

距離空間 (M, d_M) の有限点集合を X, X を中心とした半径 a の球の和集合を $B(X;a) := \bigcup_{i=1}^{n} B(\boldsymbol{x}_i; a)$ とする. ただし, $B(\boldsymbol{x}; a) = \{\boldsymbol{y} \in M \mid d_M(\boldsymbol{x}, \boldsymbol{y}) \leq a\}$ とする. 球の和集合を半径パラメーター a で集めた集合 $\mathbb{B}(X) := \{B(X;a)\}_{a \geq 0}$ をここでは X のフィルトレーションという. $a \leq b$ ならば包含関係 $B(X;a) \subset B(X;b)$ があるため, ホモロジー群間の射 $u_a^b : H_q(B(X;a)) \to H_q(B(X;b))$ を包含写像から誘導する. このとき, ホモロジー群の系列

$$H_q(\mathbb{B}(X)) : \cdots \to \ H_q(B(X;a)) \ \xrightarrow{u_a^b} \ H_q(B(X;b)) \ \to \cdots \ (a \leq b)$$

を X の q 次元パーシステントホモロジーという. パーシステントホモロジーは体係数多項式環や $\mathbf{A}_n$ 型箙の表現などの言葉で解釈する事ができ, 分解定理 [ZC05, CdS10] により適切な区間表現 $\mathbb{I}[b_i, d_i]$ を通じて

$$H_q(\mathbb{B}(X)) \cong \bigoplus_{i \in I} \mathbb{I}[b_i, d_i] \ \ (b_i \leq d_i)$$

と分解される. これをユークリッド空間 $\mathbf{R}^2$ 内に表示した多重集合

$$D_q(X) := \{(b_i, d_i) \mid i \in I\}$$

を X の q 次元パーシステント図という. パーシステント図の元 (b_i, d_i) はホモロジーの生成元の発生時間（birth time）を b_i, 消滅時間（death time）を d_i と記録しているものと解釈できる. ある生成元 $\alpha \in H_q(B(X;a))$ の発生消滅の組みが (b, d) であるとき, $\alpha = \iota_b^a(\beta)$ なるゼロでない生成元 $\beta \in H_q(B(X;b))$ の $b \leq a$ の最小値が発生時間であり, $\iota_a^d(\alpha) \in H_q(B(X;d))$ が初めてゼロ元となる $d \geq a$ が消滅時間である. 生存時間 $d - b$ が長い (resp. 短い) ほど, 対応する生成元が特徴的 (resp. ノイズ) であると捉えられる.

パーシステント図 D, E 間の距離としてはbottleneck距離

$$d_{\mathrm{W}_\infty}(D, E) = \inf_{\gamma} \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty$$

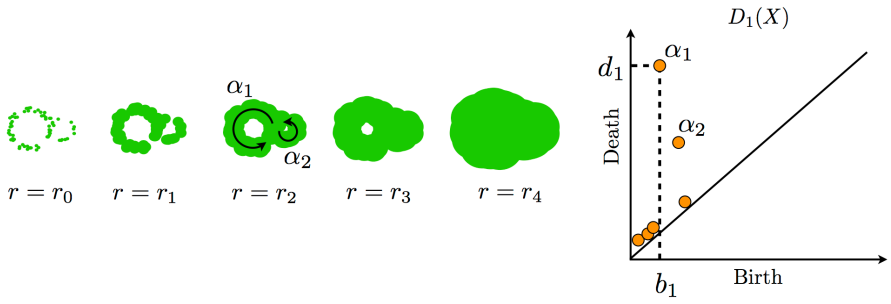

図 3: 球モデルのフィルトレーション (左) と対応するパーシステント図 (右)

が知られている． ただし, Δ は多重度が無限大の対角線集合 $\{(a,a) \mid a \in \mathbf{R}\}$ であり $\gamma : D \cup \Delta \to E \cup \Delta$ は全単射 [2]である． パーシステント図の集合を $\mathcal{D} := \{D : \text{persistence diagram} \mid d_{\mathrm{W}_\infty}(D,\Delta) < \infty\}$ とすると, $(\mathcal{D}, d_{\mathrm{W}_\infty})$ は距離空間になる. 次の, 写像 $X \mapsto D_q(X)$ がリプシッツ連続になる性質はパーシステント図の安定性として知られている:

命題 1 ([CdSO14])**.** 距離空間 (M, d_M) の有限部分集合 X, Y に対し,

$$d_{\mathrm{W}_\infty}(D_q(X), D_q(Y)) \leq d_{\mathrm{H}}(X,Y)$$

が成立する. ただし, d_H はハウスドルフ距離

$$d_{\mathrm{H}}(X,Y) = \max\left\{\sup_{\boldsymbol{x}\in X}\inf_{\boldsymbol{y}\in Y} d_M(\boldsymbol{x},\boldsymbol{y}), \sup_{\boldsymbol{y}\in Y}\inf_{\boldsymbol{x}\in X} d_M(\boldsymbol{x},\boldsymbol{y})\right\}$$

である.

例えば, X を真のデータ, Y を X の観測データとした時に, 多少ノイズが乗ったとしてもそのパーシステント図は大きく異ならないため, Y から X の位相的情報を推論する事ができる（図 4）.

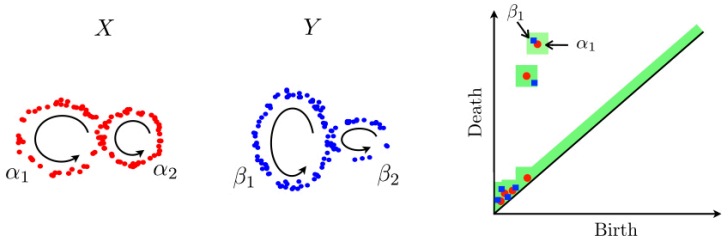

図 4: 二つの集合 X, Y (左) と対応するパーシステント図 (右). 緑の領域は $D_q(Y)$ の $\|\cdot\|_\infty$ での ε 近傍であり, $D_q(X)$ のすべての点が緑の領域に入っていることがわかる.

2. Persistence weighted Gaussian kernel

関数 $k : \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}$ を $\mathbf{R}^2$ 上の正定値カーネル [3]とする. Moore-Aronszajn の定理から正定値カーネルに付随する再生核ヒルベルト空間 $\mathcal{H}_k$ がただ一つ存在し, $\mathcal{H}_k$ は関

[2] 対角線集合の多重度を考慮することで $D \cup \Delta$ から $E \cup \Delta$ への全単射は常に存在する.

[3] 集合 Ω 上の正定値カーネル $k : \Omega \times \Omega \to \mathbf{R}$ とは (i) 対称関数 $k(x,y) = k(y,x)$ であり, (ii) 任意の $x_1, \cdots, x_n \in \Omega$ に対し行列 $(k(x_i,x_j))_{i,j=1,\cdots,n}$ が半正定値行列になるものとして定義される.

数族 $\{k(\cdot,x) \mid x \in \mathbf{R}^2\}$ を基底に持ち, その内積は $\langle k(\cdot,y), k(\cdot,x)\rangle_{\mathcal{H}_k} = k(x,y)$ として与えられる. 正定値カーネル k が可測で有界のとき, 関数 $w : \mathbf{R}^2 \to [0,\infty)$ に対して, Persistence weighted Gaussian kernel (PWGK, [KFH16]) によるパーシステント図 D のベクトル表現を次のように与える[4]:

$$V^{k,w}(D) := \sum_{x \in D} w(x) k(\cdot, x) \in \mathcal{H}_k$$

関数 w は重み関数と呼び, パーシステント図の各点はその位置（主に, 対角線集合からの距離）によって重要度が異なるため, w によってその重要度を調整する. 正定値カーネル k としてガウスカーネル $k_{\mathrm{G}}(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ を選択すると, $V^{k_{\mathrm{G}},w}(D)$ はパーシステント図の各点を正規分布に置き換えたものになる. 加えて, 次の安定性定理を得た.

定理 1 ([KFH17])**.** 正定値カーネル k は可測, 有界, 連続とし, 重み関数 w には次の定数 $B, L > 0$ が任意の $D, E \in \mathcal{D}$ に対し存在するとする:

$$\sup_{D \in \mathcal{D}} \sum_{x \in D} w(x) \leq B, \quad \sup_{\gamma : D \cup \Delta \to E \cup \Delta} \sum_{x \in D \cup \Delta} |w(x) - w(\gamma(x))| \leq L \sup_{x \in D \cup \Delta} \|x - \gamma(x)\|_\infty .$$

このとき, 写像 $D \mapsto V^{k,w}(D)$ はリプシッツ連続になる.

この定理は命題1と同様に, データの摂動が本統計解析手法に致命的な影響を与えないことを保証している.

参考文献

[CdS10] Gunnar Carlsson and Vin de Silva. Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405, 2010.

[CdSO14] Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

[HNH+16] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016.

[KFH16] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Persistence weighted gaussian kernel for topological data analysis. In *International Conference on Machine Learning*, pages 2004–2013, 2016.

[KFH17] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Kernel method for persistence diagrams via kernel embedding and weight factor. *arXiv preprint arXiv:1706.03472*, 2017.

[ZC05] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

[福10] 福水健次. **カーネル法入門—正定値カーネルによるデータ解析**. 朝倉書店, 2010.

[4] より正確には, パーシステント図の重み付き測度表示を Bochner 積分によるカーネル埋め込み (Kernel embedding) で定義される.