Persistence Weighted Gaussian Kernel for Topological Data Analysis

福水 健次

統計数理研究所　数理・推論研究系／統計的機械学習研究センター

（兼）総合研究大学院大学, 物質・材料研究機構

ERATO感謝祭 SeasonIII

2016年8月9-10日 NII

- 発表内容

Persistence Weighted Gaussian Kernel for Topological Data Analysis

by Genki Kusano, Yasuaki Hiraoka, Kenji Fukumizu
International Conference on Machine Learning (ICML2016), June 2016

草野元紀（東北大,D1）

平岡裕章（東北大）

Acknowledgements:

- JST・CREST 現代の数理科学と連携するモデリング手法の構築（総括：坪井俊）
  「ソフトマター記述言語の創造に向けた位相的データ解析理論の構築」（平岡裕章・代表）
- JST・イノベハブ構築事業　物質・材料研究機構「情報統合型物質・材料開発イニシアティブ（Mi²i）」(拠点長：寺倉清之)

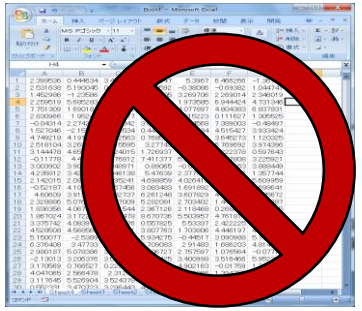
位相的データ解析 (TDA)

- TDA: データの位相的・幾何的情報を抽出するための新しい方法論

キーテクノロジー ＝ パーシステントホモロジー

（Edelsbrunner et al 2002; Carlsson 2005)

- 背景
  - データの質の変化.
    複雑な幾何的構造を持つデータを扱う必要

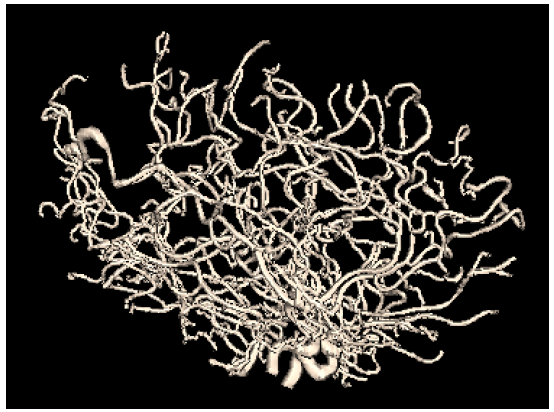


  - 計算トポロジーの発展.
    位相的不変量の計算機による計算が実行可能に.

TDA: さまざまな応用

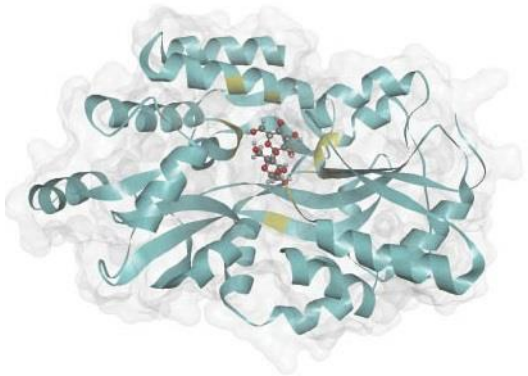
複雑な幾何構造を持つデータ → 特徴ベクトル／記述子の導入が困難

脳科学



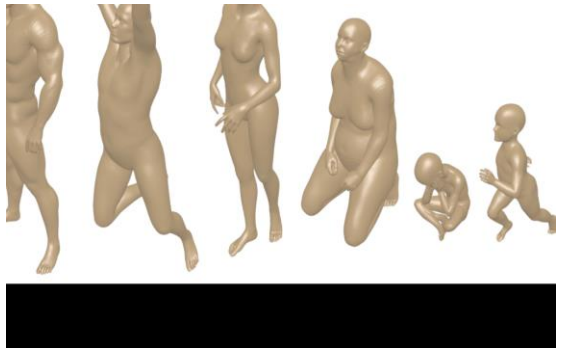
脳動脈の木構造
年齢／性別による構造の違い
(Bendich et al. *AoAS* 2016)

生化学



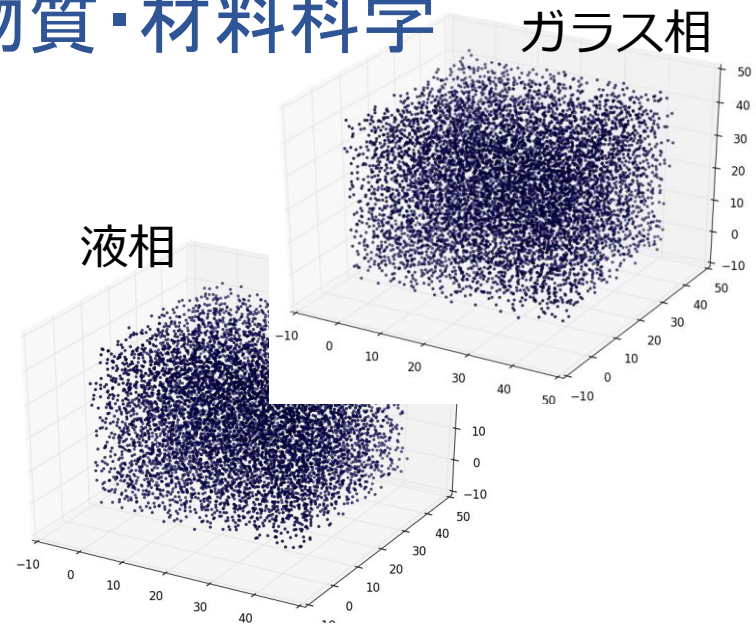
タンパク質の構造変化
構造による機能の違い
(Kovacev-Nikolic et al. *SAGMB* 2016)

コンピュータビジョン



形状の記述
(Reininghaus et al. *CVPR* 2015)

物質・材料科学



非晶質の記述
(Hiraoka et al. *PNAS 2016*)

講演の概要

- 位相的データ解析の概要
- パーシステント図のデータ解析
- 物質科学への応用

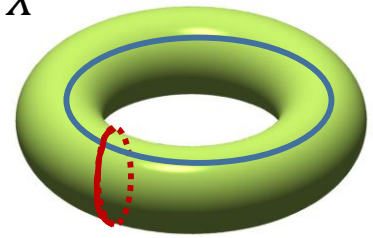
位相

$\simeq$

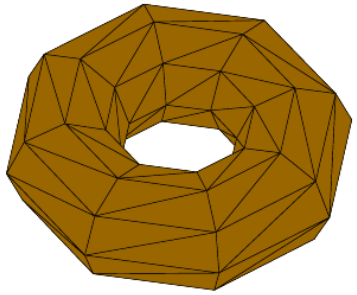
トポロジー　＝　切り貼りせずに連続的に移りあう図形は「同じ」とみなす

代数的位相幾何　＝　三角形(単体）による図形の記述 → 代数的扱い

X



≅



点，線，三角形の集まりとその
つながり具合を代数的に記述

代数的操作

位相的不変量

代表例：　ホモロジー群 $H(X)$
各次元の本質的に異なる
「穴」を捉える.

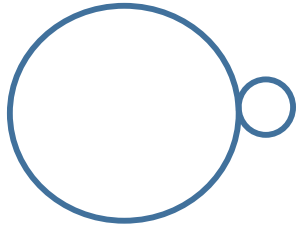
ホモロジー群の生成元
＝　本質的に異なる「穴」

「穴」の個数

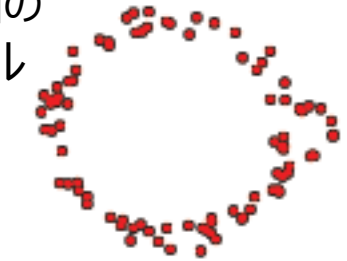
	$\dim H_0(X)$ 連結成分	$\dim H_1(X)$ ring	$\dim H_2(X)$ cavity
• ≅ ○ ≅ 	1	0	0
◎ ≅ □ ≅ 	1	1	0
 ≅ 	1	0	1

統計的データへのトポロジーの応用?

真の幾何的構造



ノイズのある
有限個の
サンプル



半径 ε の球（ε - ball）
の和集合を考えることに
より真の構造を捉えよう

(e.g.多様体学習)

小さい ε → 不連続な集合

適切なε（スケール）の設定は
難しい！

大きい ε → 小さいリングはつぶれてしまう

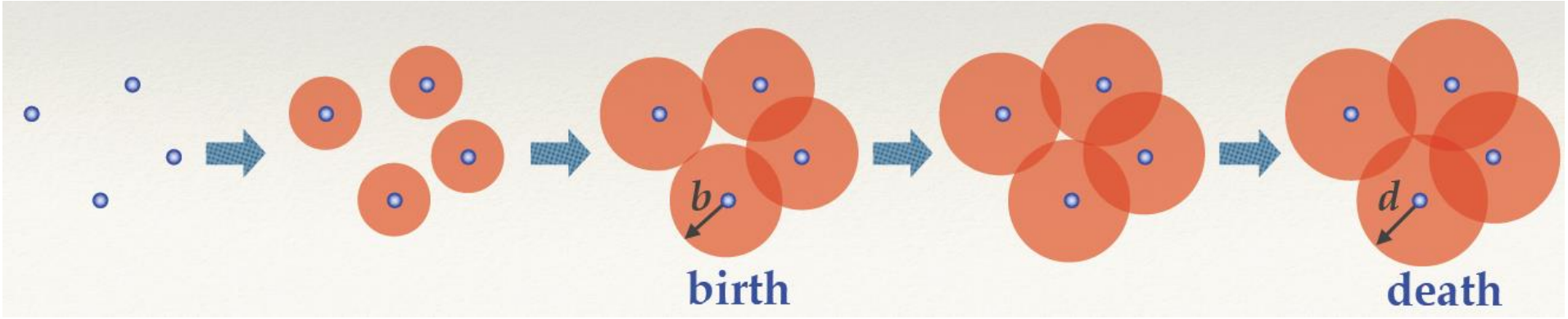
パーシステントホモロジー (PH, Edelsbrunner 2002)

- すべての ε を同時に考える.

$X = \{x_i\}_{i=1}^m \subset \mathbf{R}^d$, $X_\varepsilon := \cup_{i=1}^m B_\varepsilon(x_i)$ （ε 球の和集合）

異なるパラメータ $H_q(X_{\varepsilon_i}), H_q(X_{\varepsilon_j})$ $(\varepsilon_i < \varepsilon_j)$ に対し，ホモロジー生成元の関係づけが可能（発生，対応，消滅）→ 各生成元の発生・消滅時刻が定まる.（数学的な定理）

例： 1次元ホモロジー群 $H_1(X)$ の生成元の発生と消滅



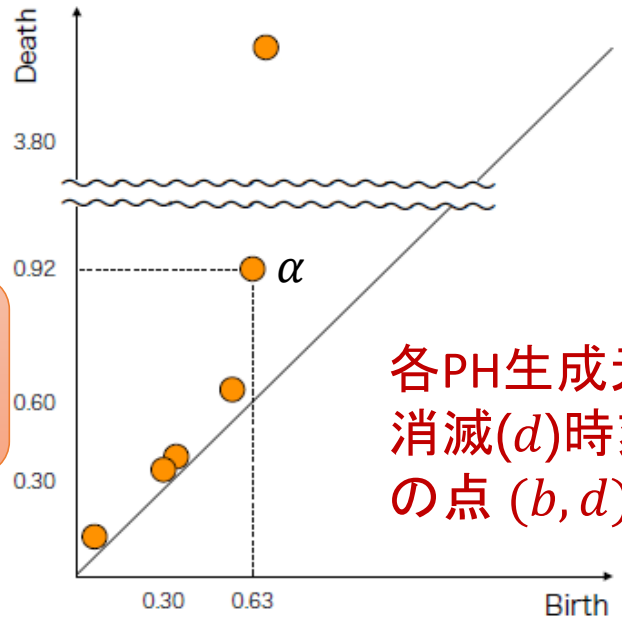
- PHの２つの表現法（等価な表現）

バーコード

各生成元のbirthから
deathまで線分を引く.

TDA: データの
記述子として使う

パーシステント図（Persistent diagram, PD）



各PH生成元の発生(b),
消滅(d)時刻を，2Dグラフ上
の点 (b, d) で表したもの.

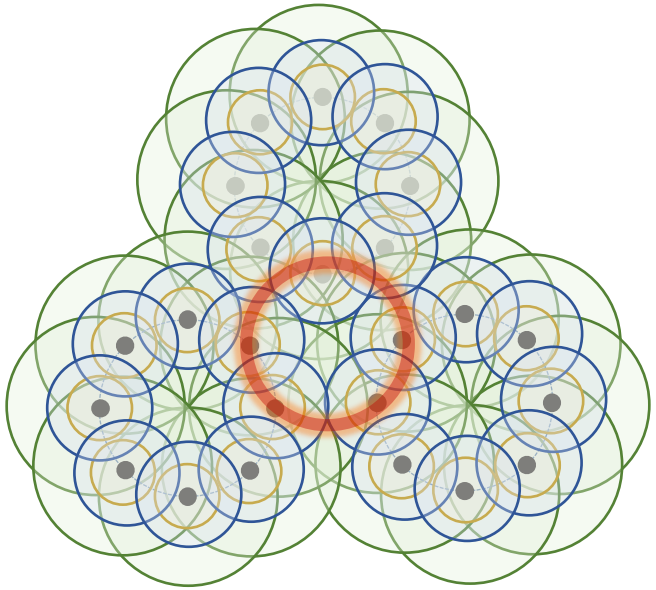
安定した構造の「穴」は
長いインターバルで存在するはず

各次元で，バーコード / パーシステント図を考える.

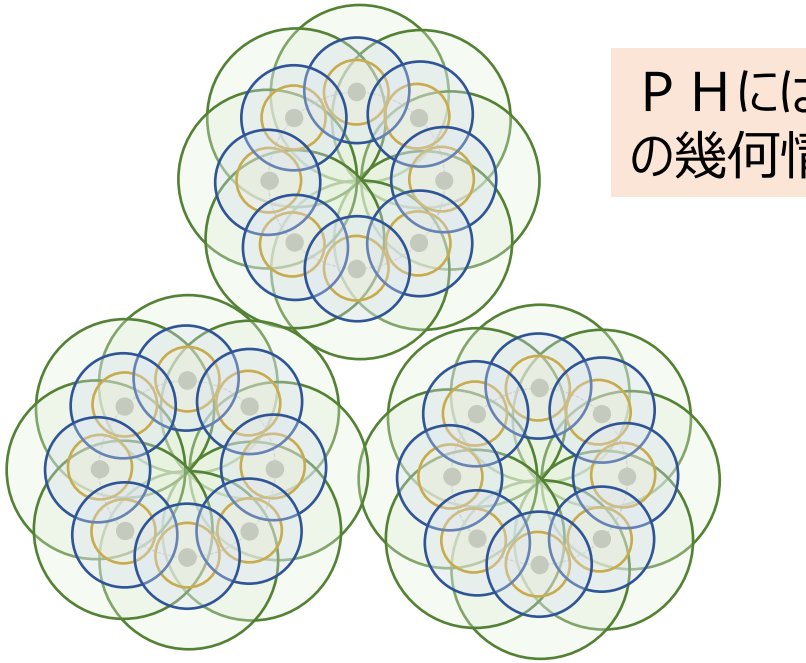
- PHの厳密な定義，計算法は本講演では扱わない (Carlsson 2009; 平岡2013)

位相を超えて

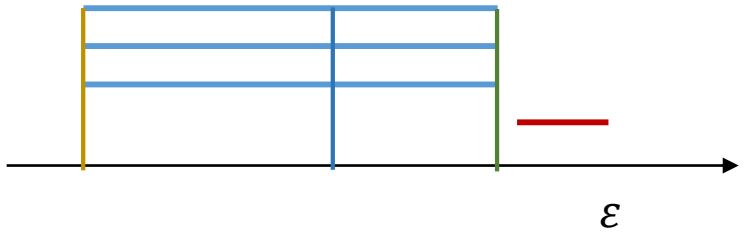
- パーシステントホモロジーは，「位相」以外の情報も含んでいる



1次元PH
のバーコード



ＰＨには位置情報など
の幾何情報も含まれる



パーシステント図の統計的データ解析

- 統計的な位相的データ解析のスキーム

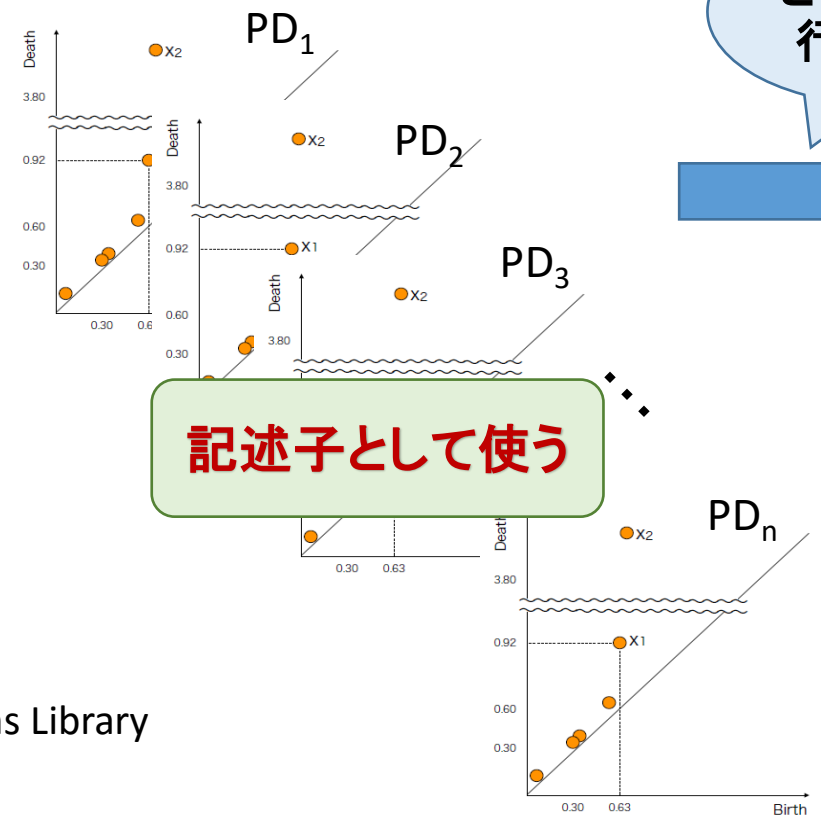
（Kusano, Hiraoka, F. 2015; Reininghaus et al CVPR2015; Kwitt et al NIPS2015; Fasy et al 2014）

多くのデータセット

PHの計算

既存のソフトウエア
e.g. CGAL / PHAT

多くのパーシステント図



記述子として使う

どのように行うか？

パーシステント図に対するデータ解析

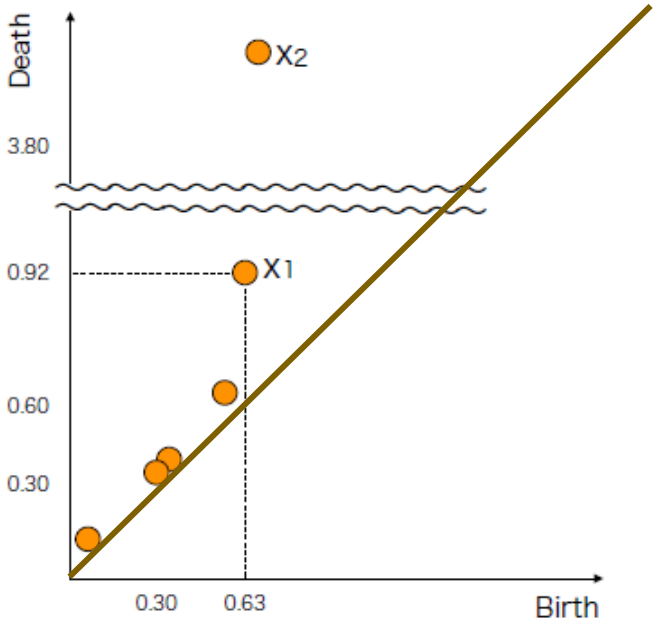
CGAL: The Computational Geometry Algorithms Library
http://www.cgal.org/
PHAT: Persistent Homology Algorithm Toolbox
https://bitbucket.org/phat-code/phat

パーシステント図のベクトル化

- 正定値カーネルによるPDのベクトル化
  - パーシステント図： $D = \{x_i\} \cup \Delta$, （点の重複度も考慮）
    $\{x_i\}$：生成・消滅時刻 $\{(b,d) \in \overline{\mathbf{R}}^2 | d > b\}$
  - PD = 離散測度と思える $\mu_D := \sum_{x_i \in D \setminus \Delta} \delta_{x_i}$
  - PDのカーネル埋め込み

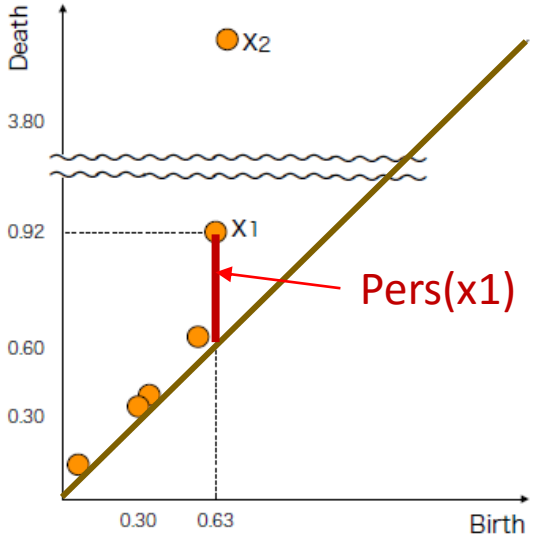
$$\mathcal{E}_k: \quad \mu_D \mapsto \int k(\cdot, x) d\mu_D(x) = \sum_i k(\cdot, x_i) \in H_k, \quad \text{ベクトル化}$$

e.g. ガウスカーネル $\sum_i \delta_{x_i} \mapsto \sum_i \exp\left(-\frac{\|y-x_i\|^2}{2\sigma^2}\right)$



ベクトルデータに対する多くのデータ解析手法が適用可能, PCA, CCA, etc
カーネル法なので，グラム行列計算に還元（カーネルトリック）．

Persistence Weighted Gaussian Kernel : PD用のカーネル (Kusano, Hiraoka, Fukumizu, 2015)



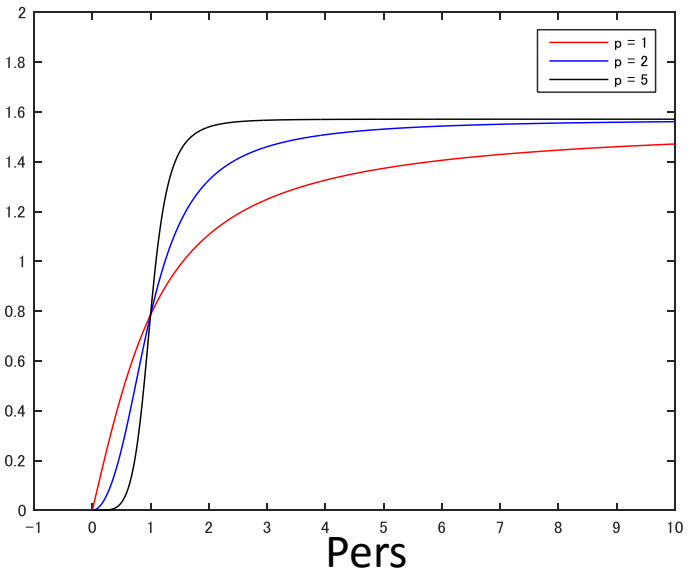
アイデア：対角線に近い生成元はノイズの可能性が高い
→ 重みを小さくする

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

重み関数 $w(x) = w_{C,p}(x) := \arctan(C\mathrm{Pers}(x)^p)$

$(C, p > 0)$

$\mathrm{Pers}(x) := d - b$ for $x \in \{(b,d) \in \mathbf{R}^2 | d \geq b\}$



- Stabilityを有することも示される（p ＞次元+1）

  点集合がHausdorff距離の意味で微小に動いたとき，PDのカーネル表現もRKHSノルムの意味で微小にしか動かない.
  （ガウスカーネルでは未解決）

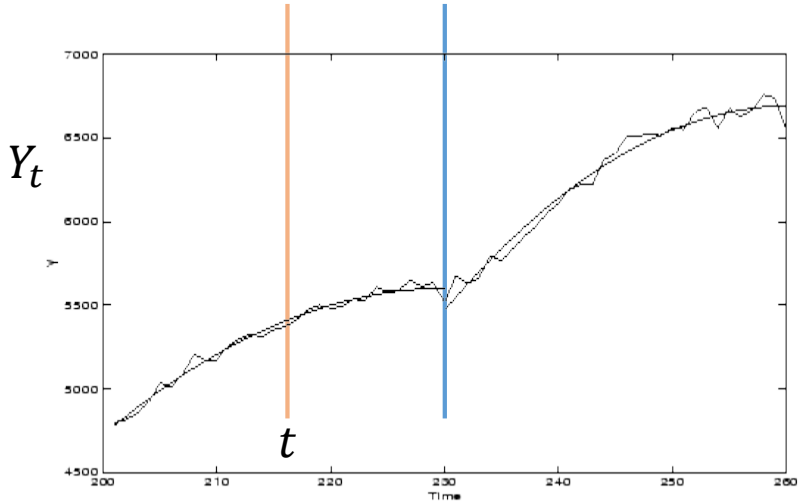
物質科学への応用

シリカ(SiO_2)の液相-ガラス相

SiO_2 を液体から急冷すると，ガラス状態になる.
目的: 液相からガラス相に転移する温度を特定したい.
データ: SiO_2分子動力学（MD）シミュレーション.
温度を変えて，80セットの３次元原子配置データを取得（ある時刻でのスナップショット）

- 原子の3次元配置データ（点集合）から，PD図を計算 (Nakamura et al 2015 Nanotechnology)
- Si と O 原子では，異なる半径の球を用いた.

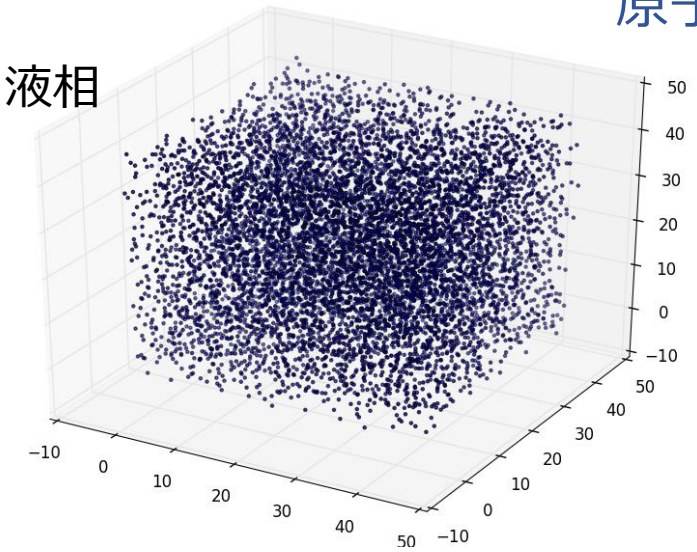
- 物理学的方法：エンタルピー曲線を描いて，微分の推定値の不連続点を推定. 正確な推定は難しい.
- 提案法：PD図のカーネル埋め込みに対する変化点検出問題として定式化. カーネル変化点検出法（Harchoui et al NIPS2008）を利用.



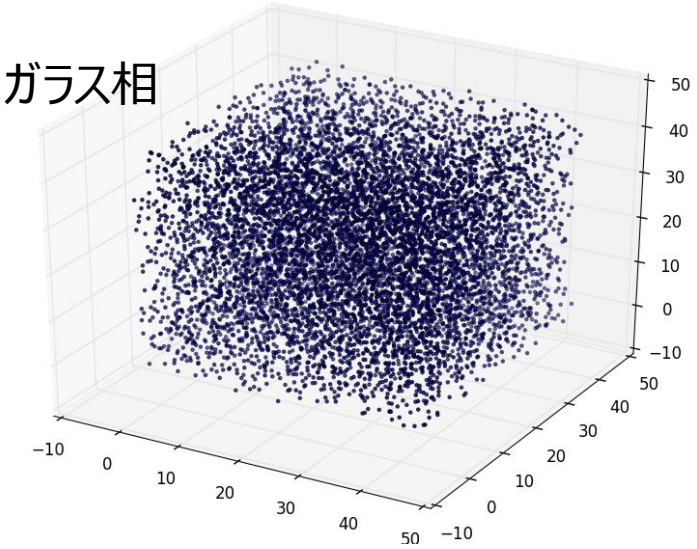
変化点

原子配置（ＭＤシミュレーション）

液相

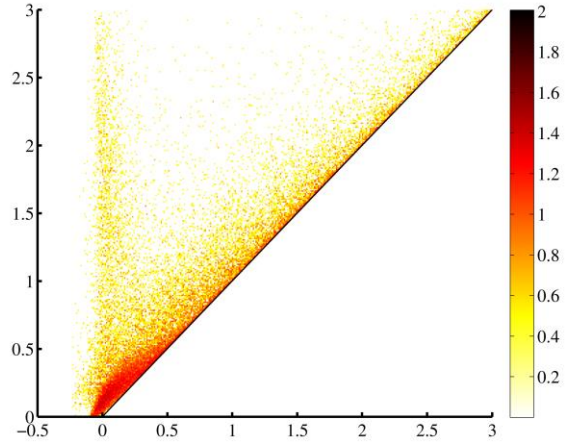


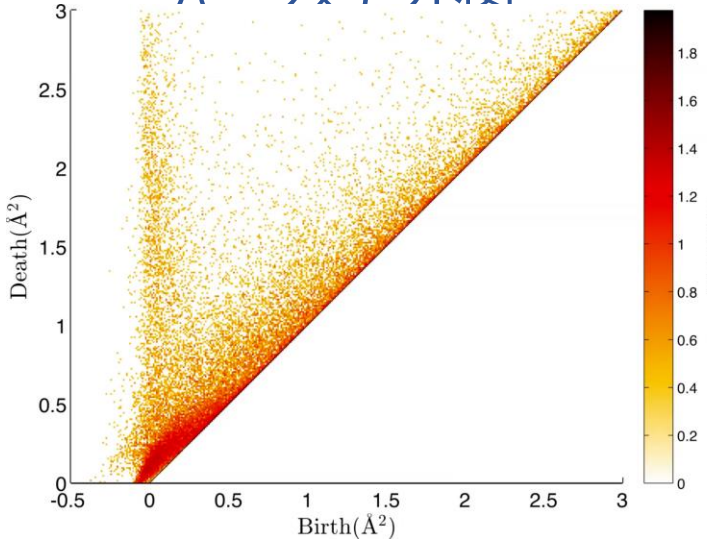
ガラス相



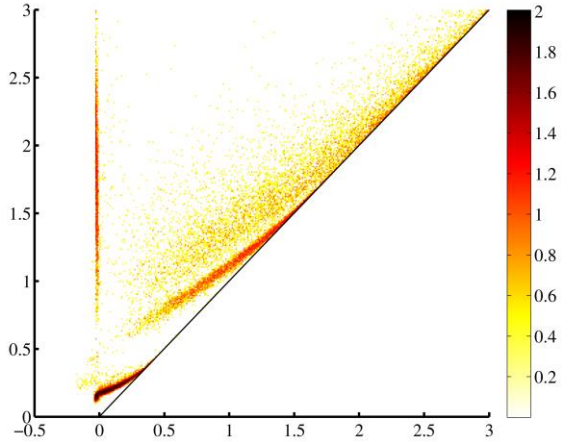
パーシステント図

液相

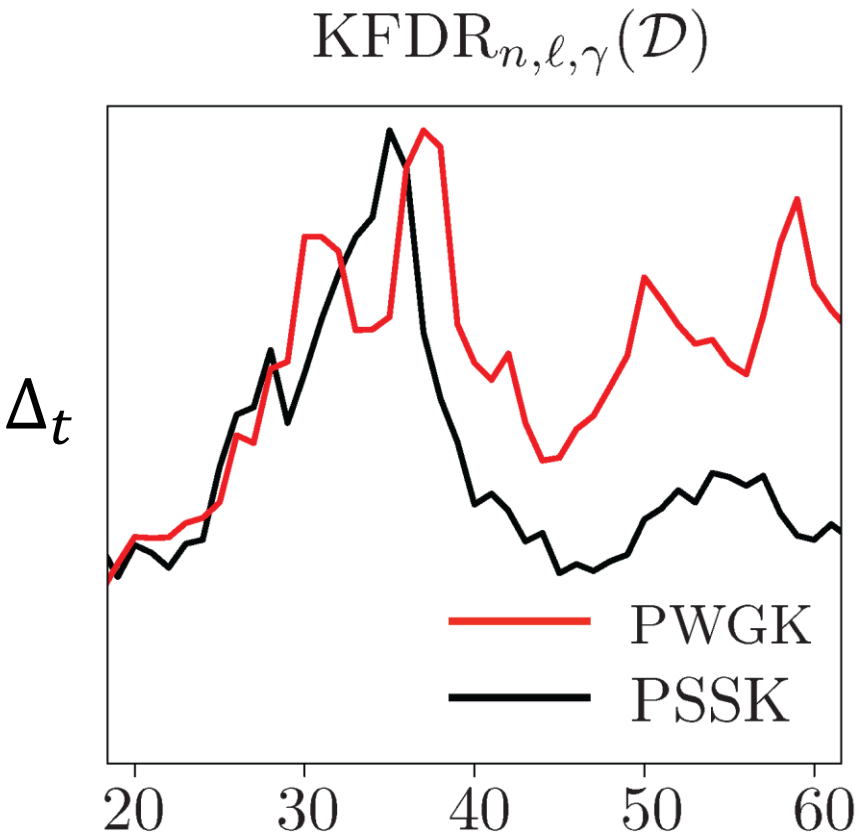




ガラス相



- SiO_2の相転移点検出



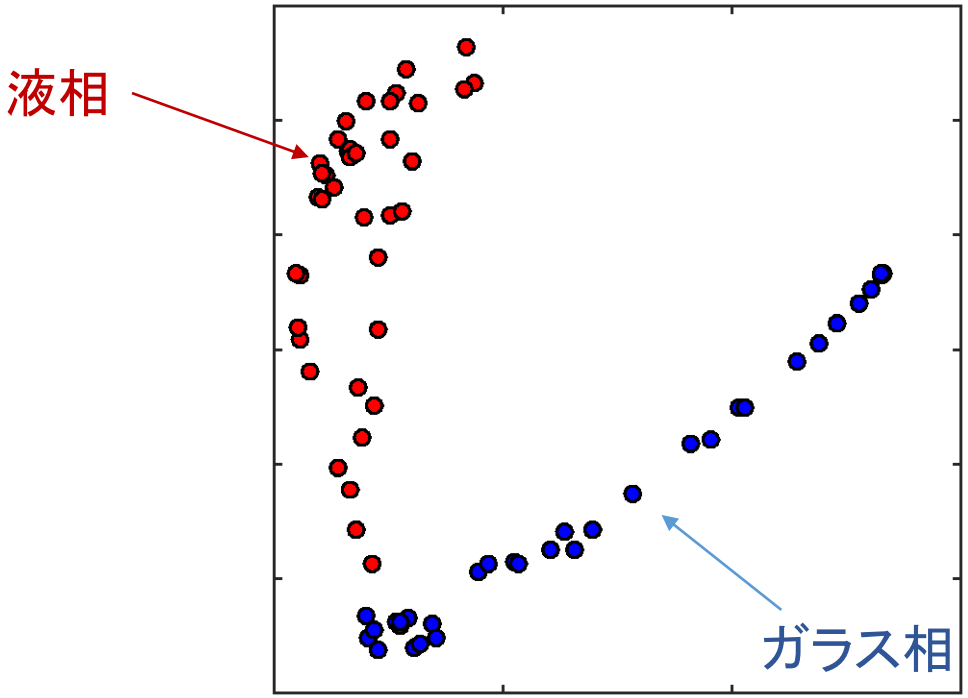
検出された変化点 = 3100K

Enthalpyによる方法: [2000K, 3500K]

PSSK: Reininghaus et al. (2015) の類似のカーネル
計算時間，識別能力でPWGKが有利.

- 低次元表現：　カーネル主成分分析

PDのカーネル埋め込み（ベクトル）に対して，さらにカーネル主成分分析を行い，２次元で表現.



液相-ガラス相は，変化点検出の結果に基づいて色付けした.

スナップショットからの液相-ガラス相判定が可能かは，物理学では未決着.

まとめ

- 位相的データ解析
  - キー技術：パーシステントホモロジー
  - 複雑な幾何的形状を持つデータに対し，幾何形状に基づく記述子を定める.
    - ε 球をすべての ε について考える
    - 位相以外の幾何情報も含む.

- 機械学習的アプローチ
  - 多数のパーシステント図に対する統計的データ解析
  - カーネル法により，パーシステント図を「ベクトル表現」
    - パーシステンス重み付きガウスカーネル：　ノイズの影響を考慮したカーネル
    - 多くの標準的データ解析手法が系統的に利用可能
      - 主成分分析，正準相関分析，判別分析，識別，クラスタリング，etc…
  - 解析の結果は比較的ブラックボックス

参考文献

Kusano, G., Fukumizu, K., Hiraoka, Y. (2016) Persistence weighted Gaussian kernel for topological data analysis. Proc. 33rd Intern. Conf. Machine Learning (ICML2016), pp. 2004–2013, 2016 http://jmlr.org/proceedings/papers/v48/kusano16.html

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002) Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

Carlsson, G. (2009) Topology and data. *Bull. Amer. Math. Soc.*, 46(2):255–308. http://dx.doi.org/10.1090/S0273-0979-09-01249-X .

平岡.　タンパク質構造とトポロジー：パーシステントホモロジー群入門.　共立出版,　2013.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., Matsue, K., and Nishiura, Y. (2015) Description of medium-range order in amorphous structures by persistent homology. *Proc. Natl. Acad. Sci. USA*. vol. 113 no. 26, 7035–7040. doi: 10.1073/pnas.1520877113

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015) Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26 (304001).

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015) A stable multi-scale kernel for topological machine learning. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* pp. 4741–4748.

Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. (2015) Statistical topological data analysis - a kernel perspective. *Advances in Neural Information Processing Systems 28*, pp. 3052–3060.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014) Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009) Kernel change-point analysis. *Advances in Neural Information Processing Systems 22*, pp. 609–616.