2016年6月17日　統計数理研究所　オープンハウス

位相的データ解析とその物質科学への応用

福水　健次　　数理・推論研究系　教授　　（東北大・AIMR・平岡裕章先生，草野元紀氏との共同研究）

■ 位相的データ解析(TDA)

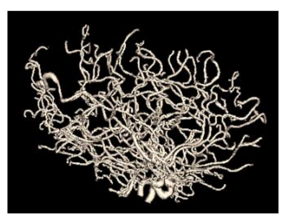
データの位相的・幾何的情報を抽出するための新しい方法論

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）

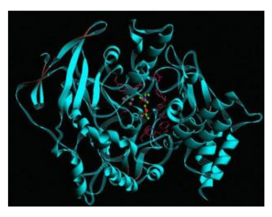
- すでに様々な応用がなされている

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 2014)

生化学

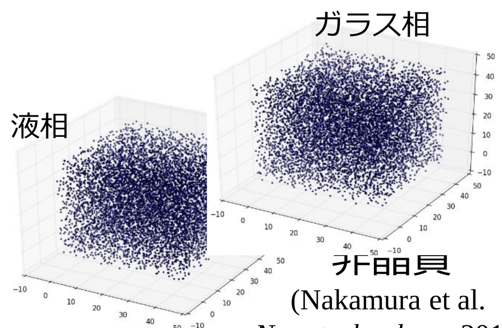


タンパク質の構造変化
構造による機能の違い
(Kovacev-Nikolic et al 2015)

コンピュータビジョン

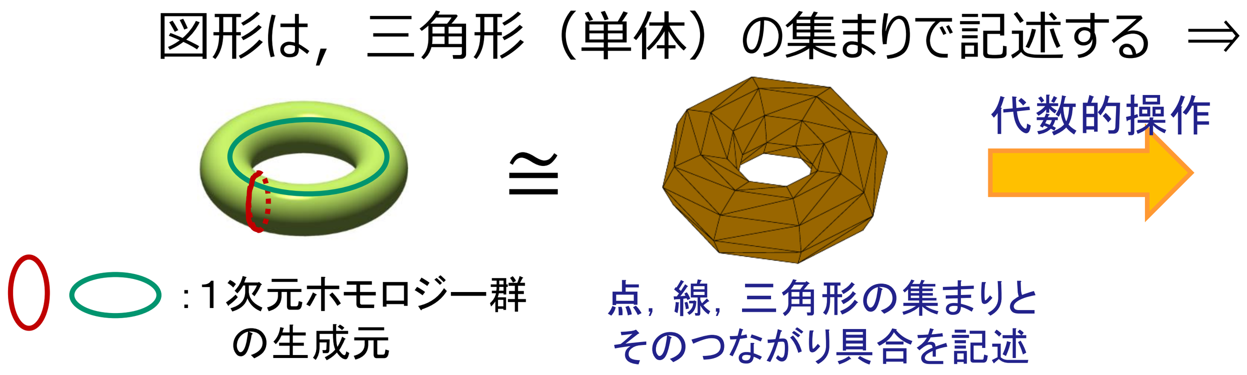
形状の記述
(Freedman & Chen 2009)

物質・材料科学



(Nakamura et al. *Nanotechnology*, 2015)

- ホモロジー群
  位相的不変量
  図形は，三角形（単体）の集まりで記述する　⇒ 代数的な扱い．



代数的操作　位相的不変量

：1次元ホモロジー群の生成元　　点，線，三角形の集まりとそのつながり具合を記述

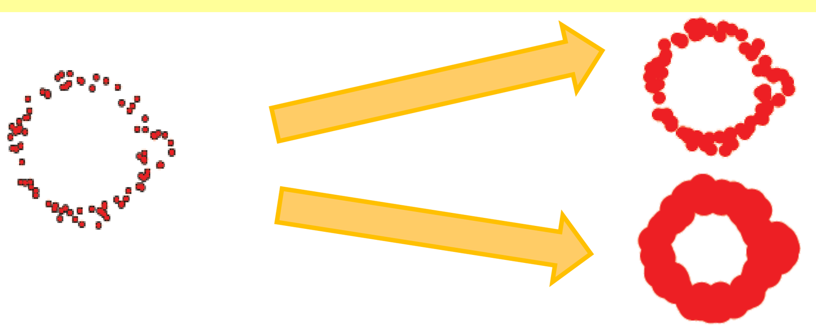
- ホモロジー群：位相的不変量として「穴」を群として表す．
  - 0次元＝連結成分　$H^0(X)$
  - 1次元＝リング　$H^1(X)$
  - 2次元＝空洞（cavity）$H^2(X)$　　・・・

ホモロジー群の生成元：連続に移り合えない「穴」の代表元

- 統計的推論における位相情報の利用
  有限データからの位相の特定は，それほど容易ではない．

半径 ε の球(ε - ball)により真の構造を捉えよう

ノイズのある有限個のサンプル

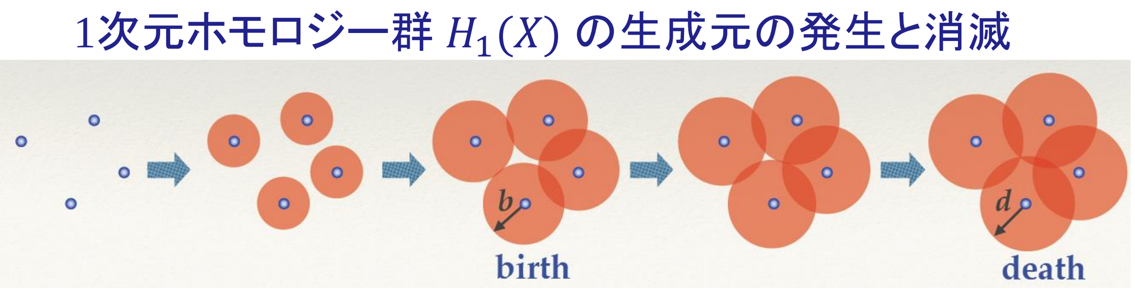


小さい ε
→ 不連続な集合

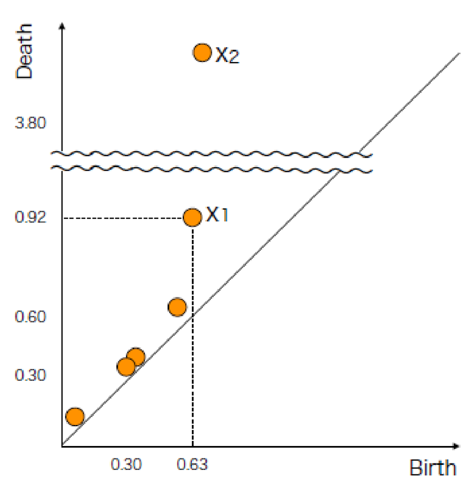
大きい ε
→ 小さいリングはつぶれてしまう

- パーシステントホモロジー　（Persistence Homology，PH）
  すべての ε を同時に考える．
  - 点集合　$X = \{x_i\}_{i=1}^m \subset \mathbf{R}^d$，　$X_\varepsilon := \cup_{i=1}^m B_\varepsilon(x_i)$
  - 位相空間の増大列
    $\mathfrak{X}: \ X_{\varepsilon_1} \subset X_{\varepsilon_1} \subset \cdots \subset X_{\varepsilon_L} \quad (\varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_L)$
  - 異なるパラメータ $\varepsilon_i < \varepsilon_j$ に対し，ホモロジー生成元の関係づけが可能（新たに発生，継続，消滅）．　（厳密な定義はCarlsson 2009; 平岡2013）
  - 各生成元の発生と消滅時刻が定まる．

1次元ホモロジー群 $H_1(X)$ の生成元の発生と消滅



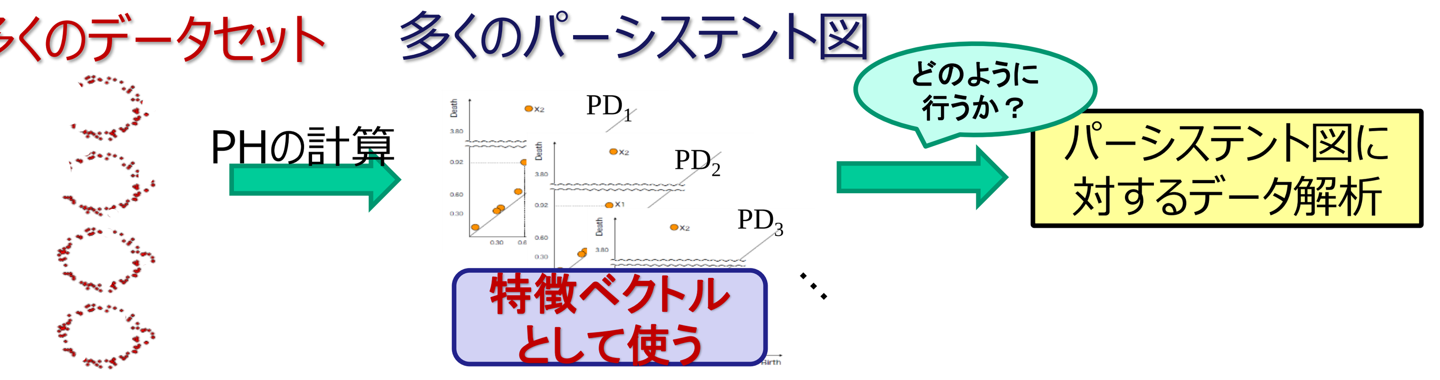
- パーシステント図（PD，生成，消滅の表現）
  各PH生成元の発生(b)，消滅(d)時刻を，2Dグラフ上の点 (b,d) で表したもの．



複雑な幾何的データの特徴ベクトル／記述子として使おう！

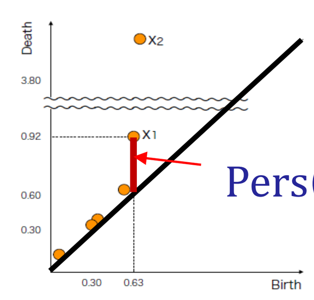
■カーネル法によるパーシステント図のベクトル化

- PDのデータ解析

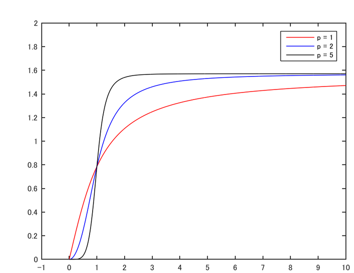


- カーネル法によるベクトル化
  PD ＝ 離散測度と思う　$\mu_D := \sum_{x_i \in D} \delta_{x_i}$　　$D = \{x_i\}$ 生成・消滅時刻
  PDのRKHSへの埋め込み
  $$\mathcal{E}_k: \quad \mu_D \mapsto \int k(\cdot, x) d\mu_D(x) = \sum_i k(\cdot, x_i) \quad \in H_k,$$
  $$\text{e.g. } \sum_i \delta_{x_i} \mapsto \sum_i \exp\left(-\frac{\|y-x_i\|^2}{2\sigma^2}\right)$$
  再生核ヒルベルト空間の元として表現　→ PD間の距離，内積(類似度)

- Persistence Weighted Gaussian Kernel（Kusano, F.ukumizu, Hiraoka ICML2016）
  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
  $$k_{PWG}(x, y) = w(x)w(y)\exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right)$$
  - 重み関数 $w(x) = w_{C,p}(x) := \arctan(C\,\mathrm{Pers}(x)^p)$　　$(C,p>0)$



$\mathrm{Pers}(x) := d - b$



- 定理（Stability）
  M：$\mathbf{R}^d$のコンパクト集合．　$S \subset M$，$T \subset \mathbf{R}^d$：有限集合．
  $p > d+1$であれば，PWG カーネル (p, C, σ) は以下を満たす．
  $$\left\| \mathcal{E}_k\left(\mu_{D_q(S)}\right) - \mathcal{E}_k\left(\mu_{D_q(T)}\right) \right\|_{H_k} \le A\, d_H(S, T).$$
  A：M, p, d, C, σのみに依存する定数，$D_q(S)$: S のq次PD，d_H：Haussdorff距離

点集合がHausdorff距離の意味で微小に変化したとき，そのベクトル表現もRKHSの距離で微小にしか動かない．

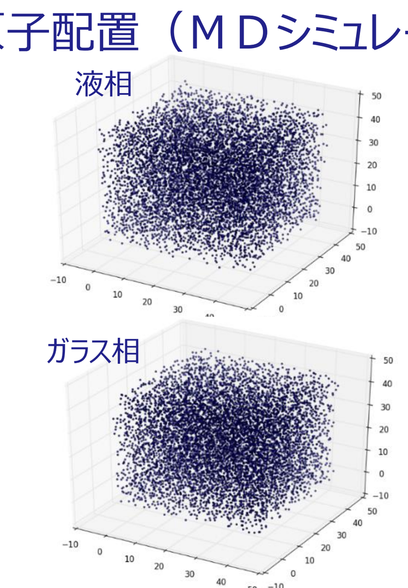
■物質科学への応用

シリカ(SiO2)の液相-ガラス相転移

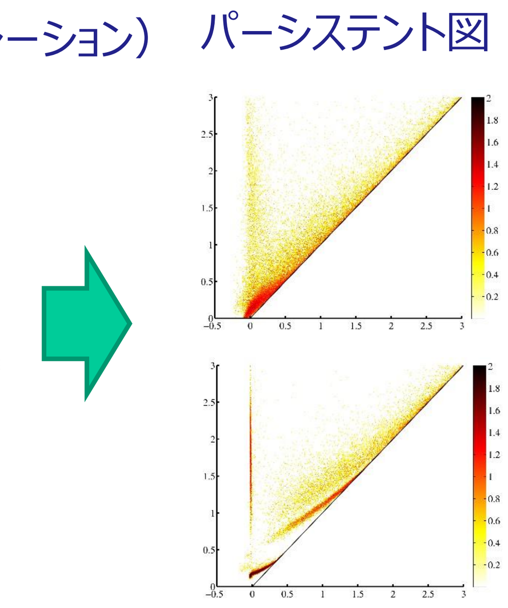
- 目的：液相からガラス相に転移する温度を特定したい．
- データ：SiO2分子動力学（MD）シミュレーション
  （Nakamura et al 2016 PNAS)

- 温度を変えて80セットの3次元原子配置データ（スナップショット）を取得
- 原子の3次元配置データから，PDを計算．

原子配置（ＭＤシミュレーション）

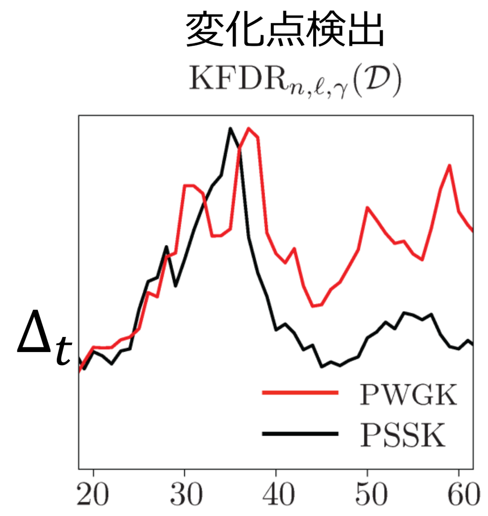


パーシステント図



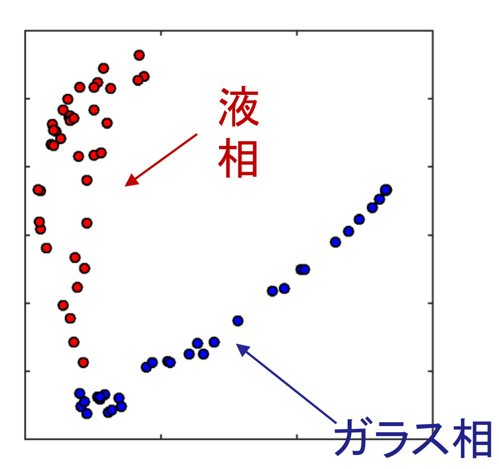
- 提案法：PDのベクトル化に対する変化点検出問題として転移点を推定

変化点検出



検出された変化点
＝3100K
物理的方法:
[2000K, 3500K]

主成分分析（カーネルPCA）



参考文献

Kusano, G., Fukumizu, K., Hiraoka, Y. (2016) Persistence weighted Gaussian kernel for topological data analysis. To appear in *Proc. ICML2016*