

パーシステント図に対する統計的機械学習

福水 健次

統計数理研究所

ENCOUNTER with MATHEMATICS

2017年12月22-23日 中央大学

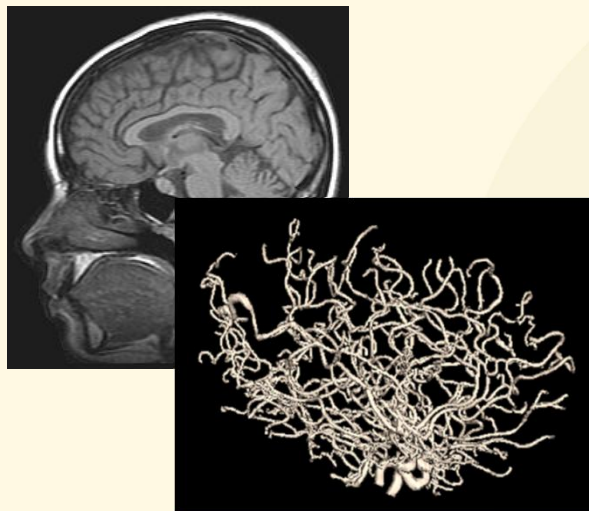
講演の概要

1. イントロ
2. データ解析：カーネル法を中心として
 - データ解析とは何か？
 - カーネル法の基礎
3. パーシステント図のデータ解析
4. おわりに

TDA: 拡大する実応用

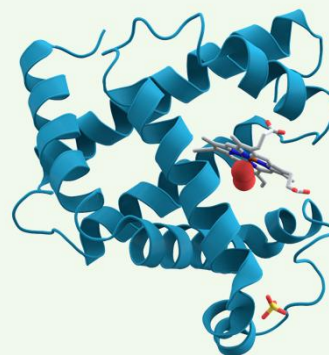
Big Dataの時代： 複雑な幾何構造を持つデータ → 特徴量/記述子の導入が困難

脳科学



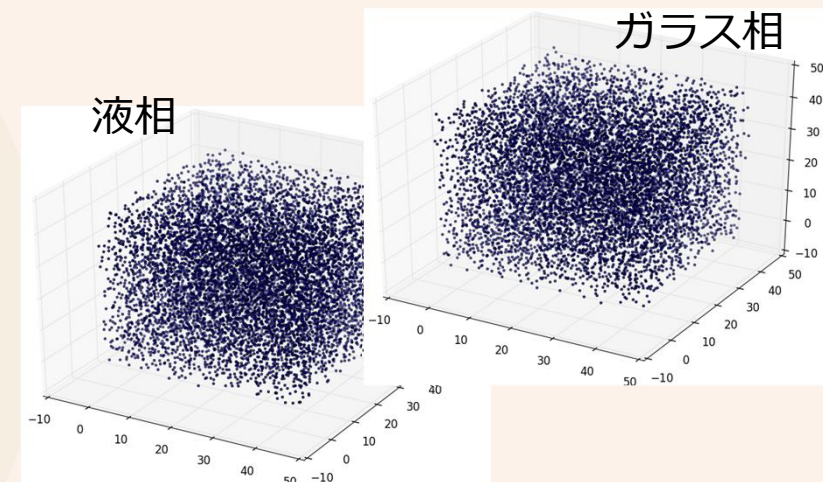
脳動脈の木構造
(Bendich et al 2014)

生化学



タンパク質の構造変化
(Kovacev-Nikolic et al 2015)

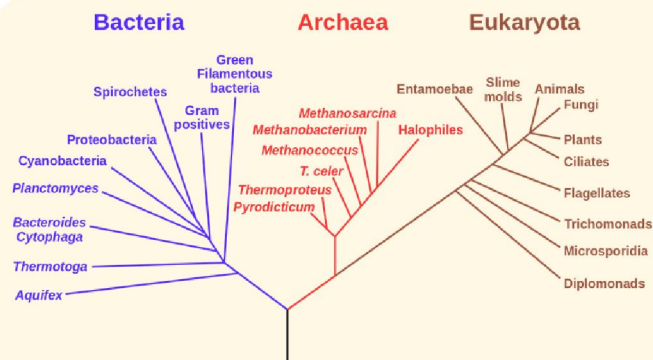
物質・材料科学



非晶質
(Nakamura et al. *Nanotechnology*, 2015;
Hiraoka et al. *PNAS* 2016)

生物学

Phylogenetic Tree of Life



遺伝子組み換え
(Cámara et al. *PLOS Comp. Bio.* 2016)

Figure omitted

葉の形のデータ
Li et al. *Nature* 2017

グラフィクス・画像

Figure omitted

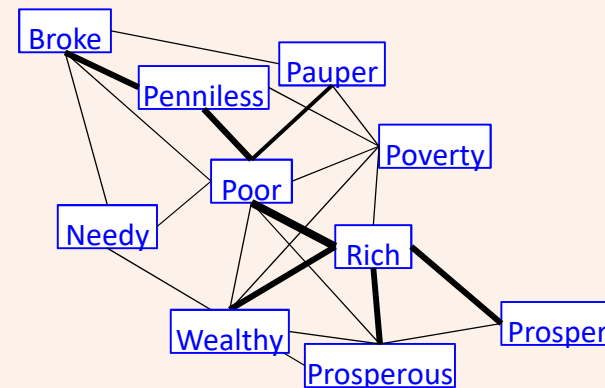
形状の記述
(Freedman & Chen 2009)

音楽

Figure omitted

Tonnetzによる音楽表現
(Bergomi et al. *CTIC2016*)

自然言語処理



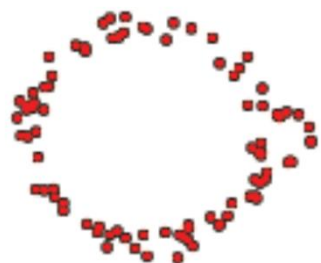
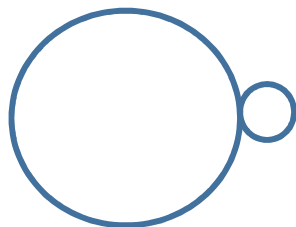
単語連想ネットワーク
(Pietrosanu et al. arXiv 2016)

統計的データへのトポロジーの応用？

8

8

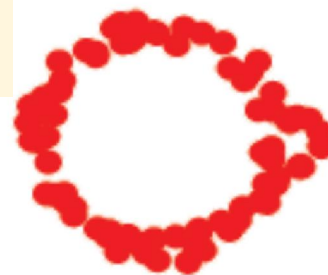
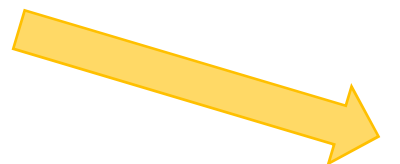
真の幾何的構造



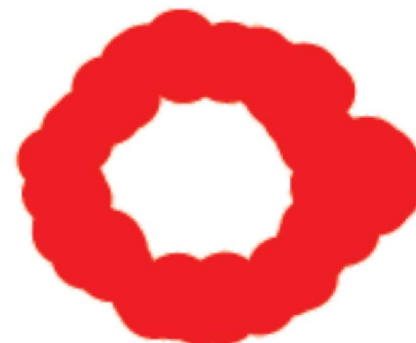
ノイズのある
有限個のサンプル

e.g. 多様体学習における近傍グラフ構成

半径 ε の球 (ε - ball)
を考えることにより
真の構造を捉えよう



小さい $\varepsilon \rightarrow$ 不連続な集合



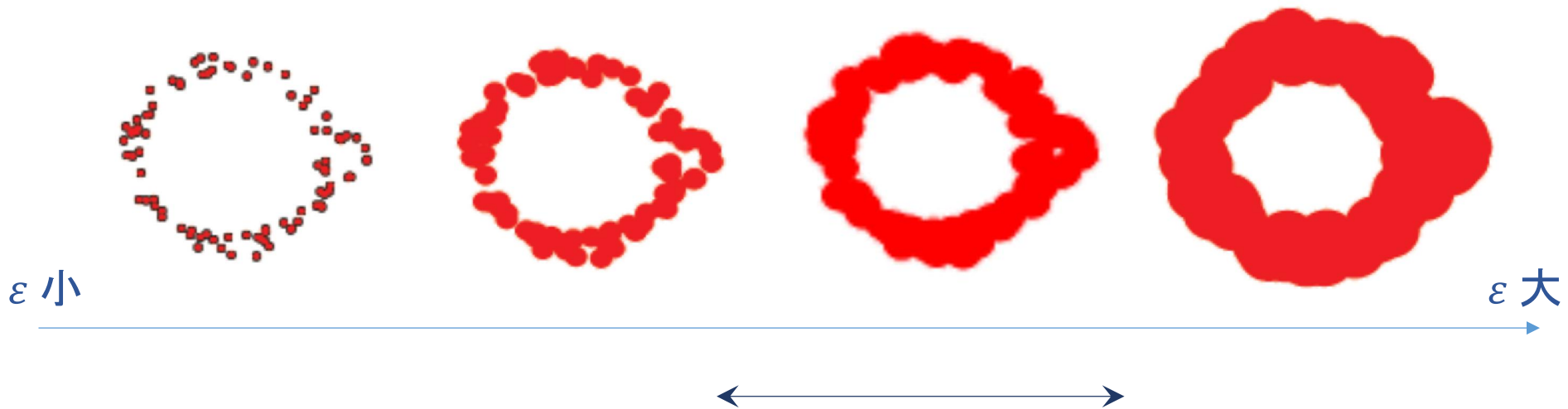
適切な ε (スケール) の設定は
難しい！

大きい $\varepsilon \rightarrow$ 小さいリングは潰れる

パーシステントホモロジー

- すべての ε (スケール) を同時に考えられる.

$$X_\varepsilon := \bigcup_{i=1}^m B_\varepsilon(x_i)$$

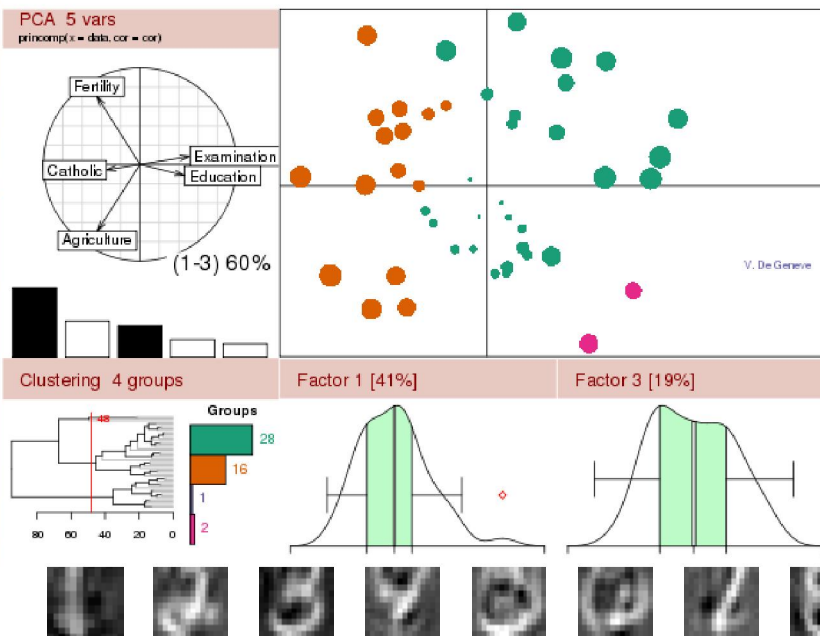
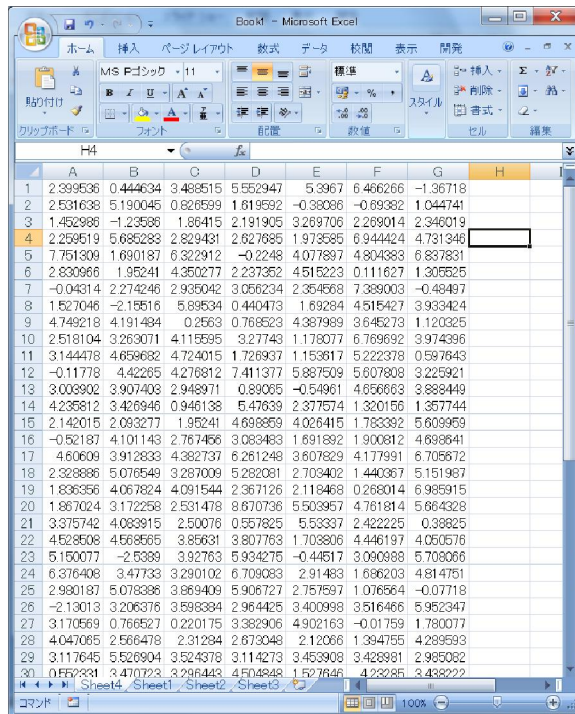


真の構造由来の2個のリング（1次元ホモロジー群の生成元）は
長いインターバルで存在するはず

データ解析とは何か？

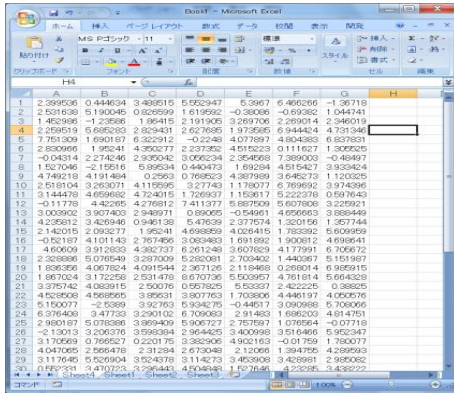
データ解析

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. -- *Wikipedia*



教科書的な（線形な）データ解析

- 数値のテーブル



	A	B	C	D	E	F	G
1	2.389536	0.444634	0.403615	5.552347	5.3967	6.495266	-1.05719
2	2.531638	5.190045	0.826599	1.619592	-0.38006	-0.69382	1.044741
3	1.452366	-1.23568	1.96415	2.191905	3.269706	2.259014	2.346019
4	2.259619	5.682352	2.35431	5.627655	1.972555	5.94444	4.751346
5	7.761509	1.690187	6.322912	-0.2248	4.077897	4.804383	6.837631
6	2.630965	1.95244	4.393277	2.237352	4.515223	0.111627	1.302025
7	-0.04314	2.274246	2.935042	3.095234	2.354568	7.389503	-0.48497
8	1.527046	-2.19516	5.99534	4.440473	1.69204	4.915427	3.932454
9	4.748218	4.191484	0.2553	0.769523	4.387989	3.648273	1.120325
10	2.551014	3.283011	4.115595	3.27740	1.710077	6.769692	3.974396
11	3.144438	4.696652	4.754015	7.259577	1.153817	5.222778	0.597643
12	-0.11778	4.42285	4.276812	7.411377	5.687509	5.607808	3.225921
13	3.003932	3.907403	2.949571	0.95905	-0.54961	4.655653	3.888449
14	4.225612	3.426946	0.946138	5.47639	2.377574	1.320156	1.357744
15	2.142015	2.898277	1.95241	4.698889	4.034415	1.783982	5.009959
16	-0.52187	4.101143	2.767456	3.093483	1.691892	1.900812	4.695641
17	4.60609	3.912833	4.362737	6.261248	3.607829	4.177991	6.706672
18	2.353886	5.076549	3.257039	5.252091	2.703402	4.442867	5.151867
19	1.836356	4.067824	4.091544	2.367126	2.116488	0.266014	6.985915
20	1.987024	3.172259	2.531476	6.670736	5.603557	4.761814	5.664328
21	3.375742	4.065915	2.50076	0.597625	5.53357	2.422225	0.38825
22	4.82959	4.565955	3.99531	3.807763	7.03806	4.461167	4.092576
23	5.150077	-2.5389	3.92763	5.934275	-0.44517	3.090988	5.700056
24	6.376408	3.47733	3.230102	6.709093	2.91483	1.686203	4.614761
25	2.860187	5.076386	3.894909	5.606727	2.767697	1.07654	-0.07716
26	-2.13013	3.206376	3.593894	2.964425	3.400988	3.516486	5.952347
27	3.170569	0.766527	2.223175	5.362306	4.902163	-0.01759	1.700077
28	4.047095	2.566478	2.31284	2.673048	2.12066	1.394795	4.289593
29	3.117645	5.526904	3.524378	5.114273	3.453938	3.428851	2.365082
30	0.187231	4.470723	3.594443	4.548486	1.577636	4.23252	3.438222

行列による表示

$$X = \begin{pmatrix} X_1^{(1)} & \dots & X_m^{(1)} \\ X_1^{(2)} & \dots & X_m^{(2)} \\ \vdots & \ddots & \vdots \\ X_1^{(N)} & \dots & X_m^{(N)} \end{pmatrix}$$

m : 次元 (列)
 N : データ数 (行)

- 線形代数によるデータ解析

- 相関
- 線形回帰
- 主成分分析 etc.

有益な情報

- 例1：主成分分析 (Principal component analysis, PCA)

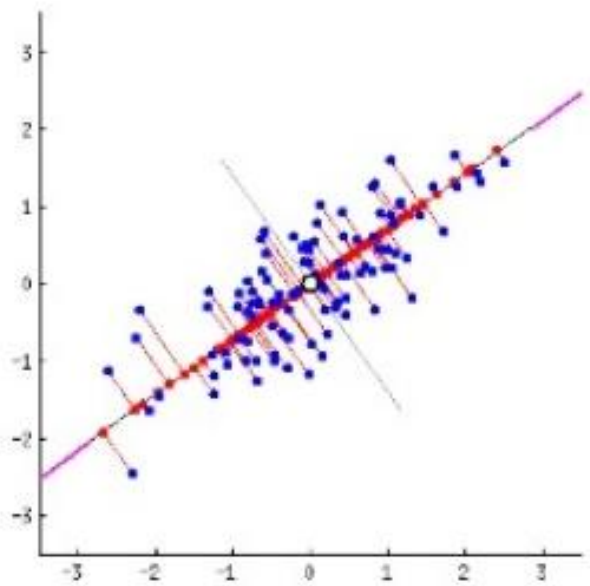
PCA: データを低次元表現する方法
分散を最大にする部分空間に正射影する.

1st direction = $\operatorname{argmax}_{\|a\|=1} \operatorname{Var}[a^T X]$

$$\operatorname{Var}[a^T X] = \frac{1}{N} \sum_{i=1}^N \left\{ a^T \left(X^{(i)} - \frac{1}{N} \sum_{j=1}^N X^{(j)} \right) \right\}^2 = a^T V_{XX} a.$$

$$V_{XX} = \frac{1}{N} \sum_{i=1}^N \left(X^{(i)} - \frac{1}{N} \sum_{j=1}^N X^{(j)} \right) \left(X^{(i)} - \frac{1}{N} \sum_{j=1}^N X^{(j)} \right)^T$$

X の分散共分散行列 ($m \times m$ matrix)



- 第1主成分方向 $= \arg \max_{\|a\|=1} a^T V_{XX} a$
 $= \operatorname{argmax}_{\|a\|=1} a^T V_{XX} a$
 $= u_1$ (最大固有値に対する固有ベクトル)
- 第 p 主成分方向 = 第 p 固有値に対する固有ベクトル

PCA $\Rightarrow$ 分散共分散行列 V_{XX} の固有値問題

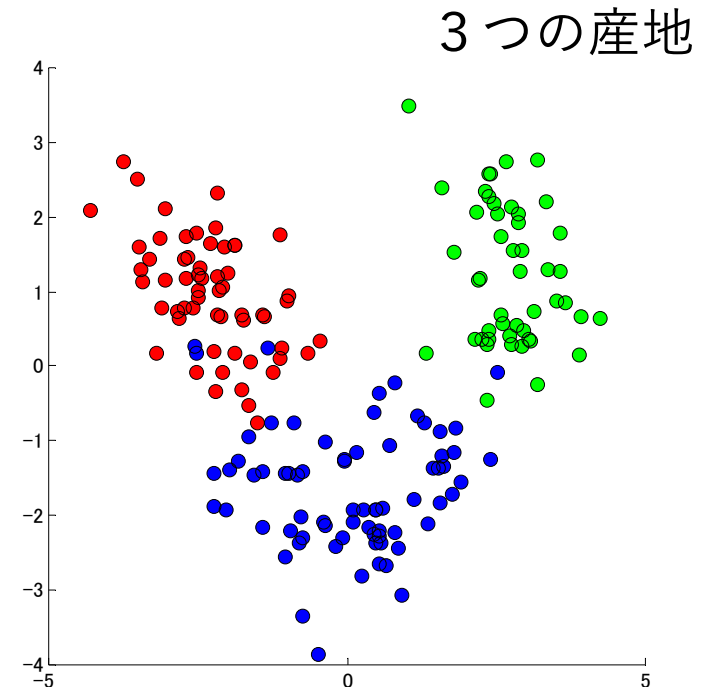
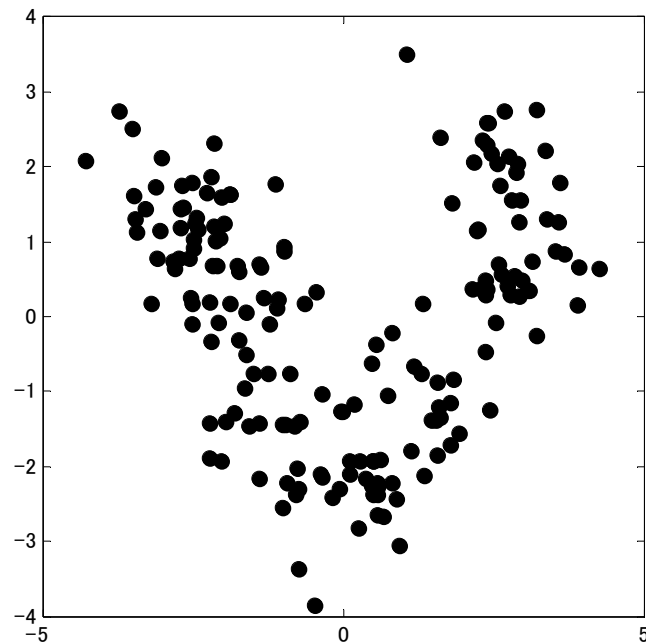
- 具体例：Wine データ (taken from UCI Machine Learning Repository)

178種のイタリアワイン

13個の化学成分を計測

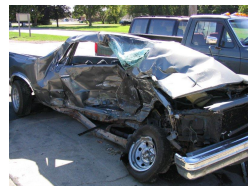
$N = 178, m = 13.$

→ PCA: 2次元の射影による可視化



- 例2：線形識別

- 2クラス識別問題
画像の識別：



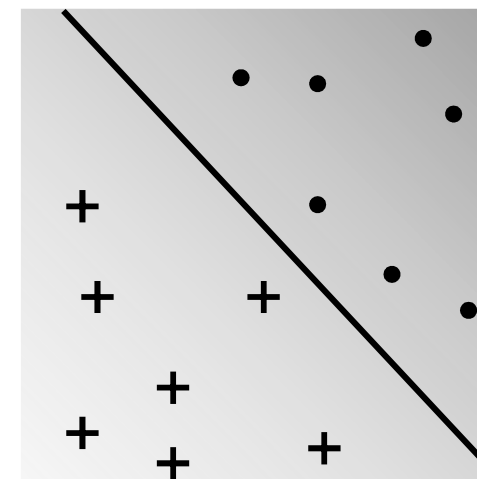
Car? → Yes/No

- 訓練データ

Input data

Class label

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \dots & X_m^{(1)} \\ X_1^{(2)} & \dots & X_m^{(2)} \\ \vdots & \ddots & \vdots \\ X_1^{(N)} & \dots & X_m^{(N)} \end{pmatrix} \quad Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(N)} \end{pmatrix} \in \{\pm 1\}^N$$



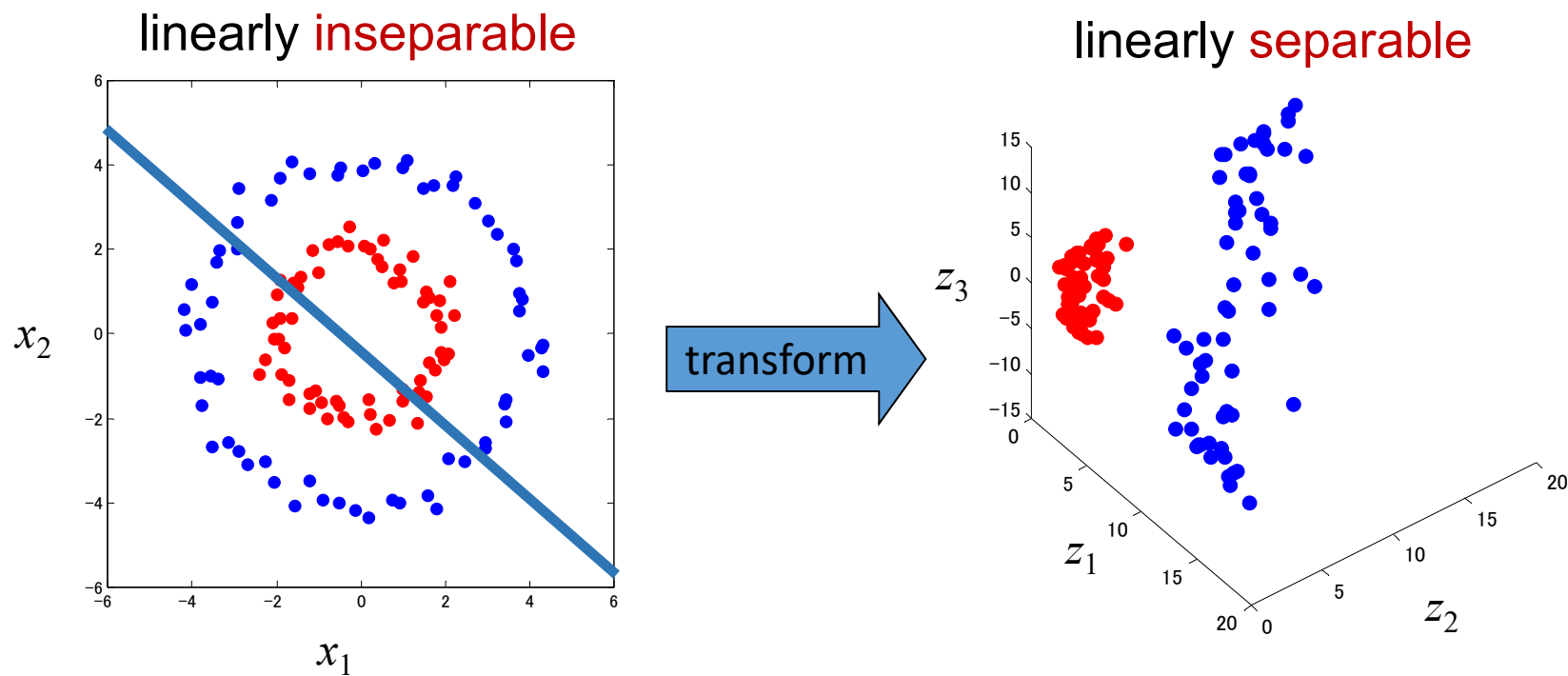
- 学習：線形識別関数を求める

$$h(x) = \text{sgn}(a^T x + b)$$

such that $h(X^{(i)}) = Y^{(i)}$ for all (or most) i .

例) Fisher判別関数, Support Vector Machine (SVM), ロジスティック回帰

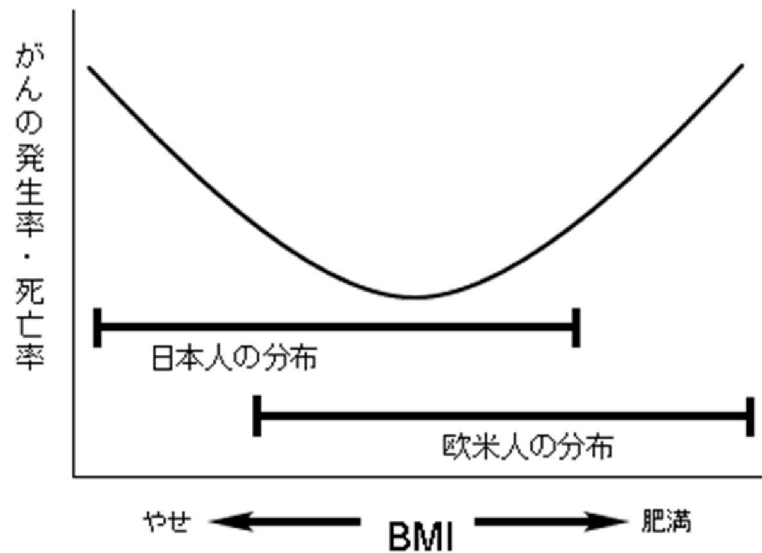
線形で十分か？



$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

Body-Mass-Index vs Cancer Risk (death rate)

肥満とがんとの関係



喫煙状況別にみた肥満度BMIとがん全体の死亡との関連(男)



Japan Public Health Center-based Prospective Study,
National Cancer Center

非線形変換は有用

カーネル法の基礎 ー 非線形データ解析 ー

データの非線形変換

- 高次情報の抽出

$$(X, Y, Z) \rightarrow (X, Y, Z, X^2, Y^2, Z^2, XY, YZ, ZX, \dots)$$

- 元の空間の次元が高いと 計算は実現できない!

e.g. 10000 次元のデータ,
2 次までの特徴

$$10000C_1 + 10000C_2 = 50,005,000$$

- 計算量爆発.
より効率的な方法が必要 → カーネル法

非ベクトルのデータ

- データ形式の多様化

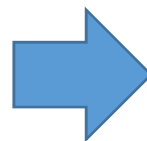
- 文字列 (String)

She keeps her room clean.

Alice gave a present to Bob.

- ヒストグラム

- 画像のRGBヒストグラム表現



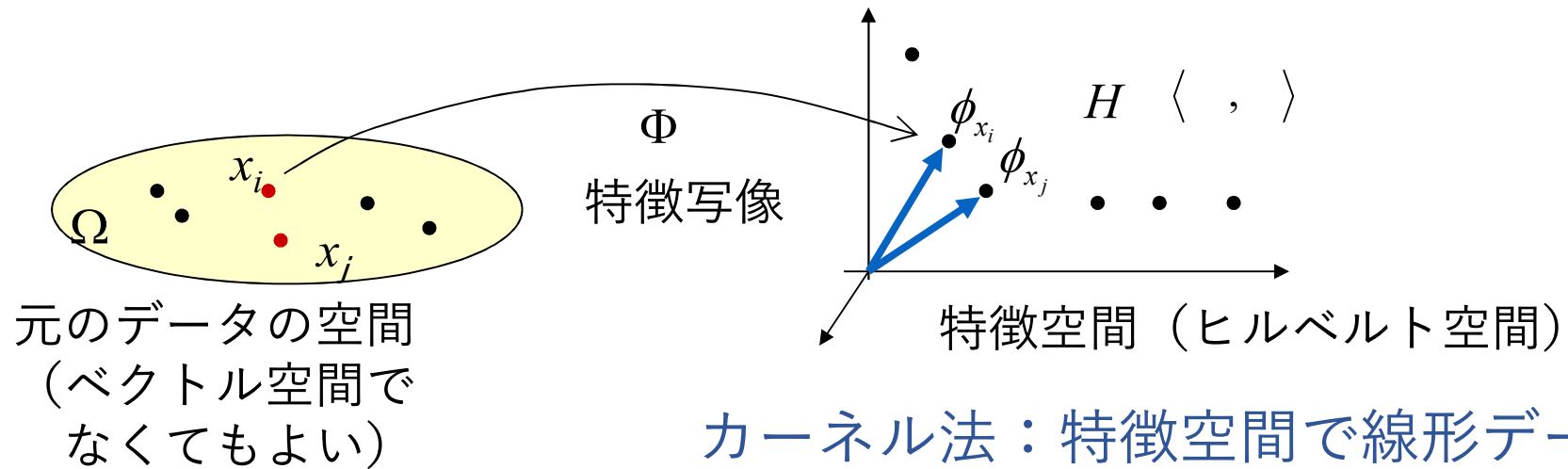
標準的なデータ解析手法はベクトルデータを想定.

他の形式にはそのままでは適用できない



カーネル法

カーネル法の概念図



カーネル法：特徴空間で線形データ解析を適用

$$\Phi: \Omega \rightarrow H$$

- 内積計算が陽に可能.

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

kernel trick

- 多くの線形データ解析の計算は内積に依拠している.

正定値カーネル

Def.

Ω : set. $k: \Omega \times \Omega \rightarrow \mathbf{R}$ is a **positive definite kernel** if it satisfies

1) (symmetry) $k(x, y) = k(y, x)$

2) (positivity) for arbitrary points $x_1, \dots, x_n$ in Ω , the

Gram matrix $\begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$ is positive semidefinite,

$$\text{i.e., } \sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0 \quad \text{for any } c_i \in \mathbf{R}.$$

- $\mathbf{R}^m$ 上の正定値カーネルの例

- Euclid内積

$$k(x, y) = x^T y$$

- 多項式カーネル

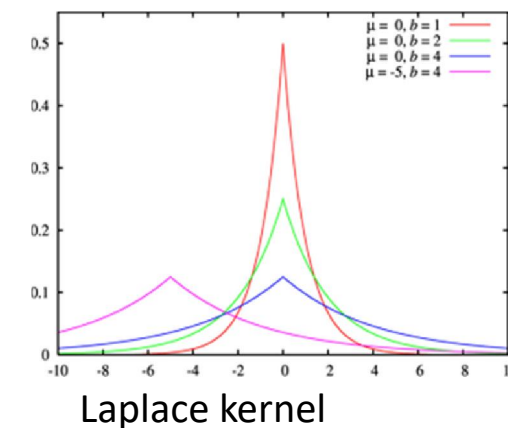
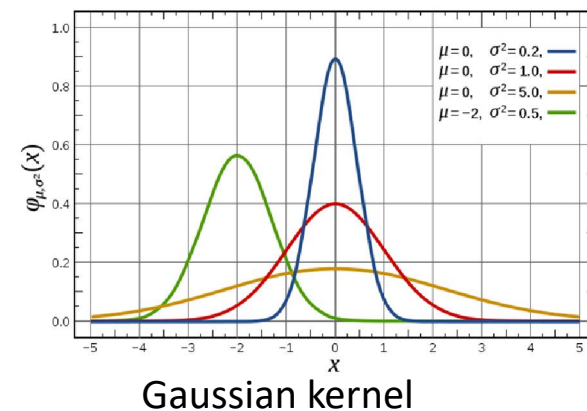
$$k_P(x, y) = (c + x^T y)^d \quad (c > 0, d \in \mathbf{N})$$

- Gaussian カーネル

$$k_G(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right) \quad (\sigma > 0)$$

- Laplace カーネル

$$k_L(x, y) = \exp\left(-\alpha \sum_{i=1}^m |x_i - y_i|\right) \quad (\alpha > 0)$$



命題

H を内積 $\langle \cdot, \cdot \rangle$ を持つベクトル空間とし, $\Phi: \Omega \rightarrow H$ を写像(特徴写像)とする. $k: \Omega \times \Omega \rightarrow \mathbf{R}$ を

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

により定義すると, $k(x, y)$ は正定値である.

- カーネルトリック (内積の陽な計算) を成り立たせる関数は, 正定値カーネルである.

定理 (Moore-Aronszajn)

Ω 上の正定値カーネル k に対し, Ω 上の関数からなる Hilbert空間 H_k が一意に存在して, 次が成り立つ.

- 1) $k(\cdot, x) \in H_k$ ($\forall x \in \Omega$).
- 2) $\text{span} \{ k(\cdot, x) \mid x \in \Omega \}$ は H_k で稠密
- 3) (再生性)

$$\langle f, k(\cdot, x) \rangle = f(x) \quad \text{for any } f \in H_k, x \in \Omega.$$

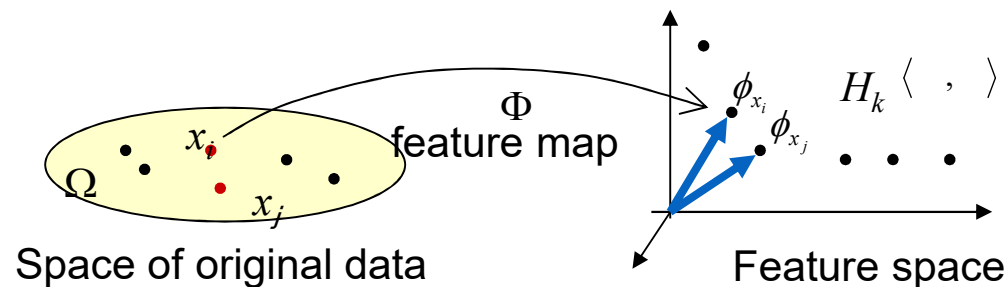
- 上の H_k を k が定める再生核ヒルベルト空間 (reproducing kernel Hilbert space, RKHS) という. *c.f.* L^q 空間
- 再生性より特に, $\Phi(x) = k(\cdot, x)$ とおくと,

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y).$$

正定値カーネルによるデータ解析

- 正定値カーネル k を用意
- 特徴写像: $\Phi: \Omega \rightarrow H_k, x \mapsto k(\cdot, x)$

$$X_1, \dots, X_n \mapsto k(\cdot, X_1), \dots, k(\cdot, X_n)$$



- カーネルトリック（再生性）：

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j)$$

非ベクトルデータの
ベクトル化が可能

- 正定値カーネルを与えれば十分.
 - 多くのデータ解析手法は、内積計算ができれば適用可能.
 - 特徴写像, 特徴ベクトルを陽に知る必要はない.
 - カーネル法の計算は、グラム行列 $\left(k(X_i, X_j) \right)_{ij}$ による計算となる.

カーネルPCA

- PCA: 線形な次元削減 → カーネルPCA: 非線形な次元削減 (Schölkopf et al. 1998).
- RKHSでPCAを行う

$$\max_{a \in \mathbf{R}^m, \|a\|=1} \text{Var}[a^T X] = \frac{1}{N} \sum_{i=1}^N \left\{ a^T \left(X^{(i)} - \frac{1}{N} \sum_{j=1}^N X^{(j)} \right) \right\}^2$$

普通のPCA

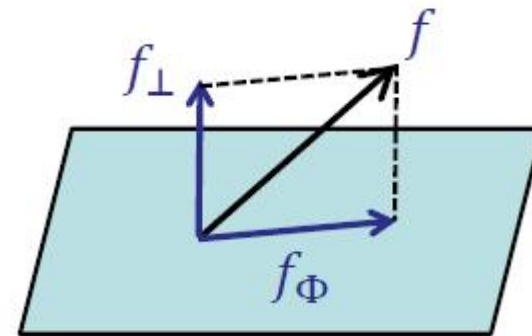


$$\max_{f \in H_k, \|f\|_{H_k}=1} \text{Var}[\langle f, \Phi(X) \rangle] = \frac{1}{N} \sum_{i=1}^N \left\{ \left\langle f, \Phi(X^{(i)}) - \frac{1}{N} \sum_{j=1}^N \Phi(X^{(j)}) \right\rangle \right\}^2$$

カーネルPCA

次の形の f を考えれば十分

$$f = \sum_{i=1}^N c_i \left(\Phi(X^{(i)}) - \frac{1}{N} \sum_{j=1}^N \Phi(X^{(j)}) \right)$$



(直交する方向は分散に効いてこない!)

→ (再生性を使うと)

$$\max \quad \text{Var}[\langle f, \Phi(X) \rangle] = \frac{1}{N} c^T \tilde{K}_X^2 c$$

$$\text{subject to} \quad \|f\|=1 \Leftrightarrow c^T \tilde{K}_X c = 1$$

$$\begin{aligned} (\tilde{K}_X)_{ij} &= k(X^{(i)}, X^{(j)}) - \frac{1}{N} \sum_{b=1}^N k(X^{(i)}, X^{(b)}) \\ &\quad - \frac{1}{N} \sum_{a=1}^N k(X^{(a)}, X^{(j)}) + \frac{1}{N^2} \sum_{a,b=1}^N k(X^{(a)}, X^{(b)}) \end{aligned}$$

(中心化Gram行列)

- カーネルPCAのアルゴリズム:

- 中心化Gram行列 $\tilde{K}_X$ の計算
- $\tilde{K}_X$ の固有分解 $\tilde{K}_X = \sum_{i=1}^N \lambda_i u_i u_i^T$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N \geq 0 \quad \text{eigenvalues}$$

$$u_1, u_2, \dots, u_N \quad \text{unit eigenvectors}$$

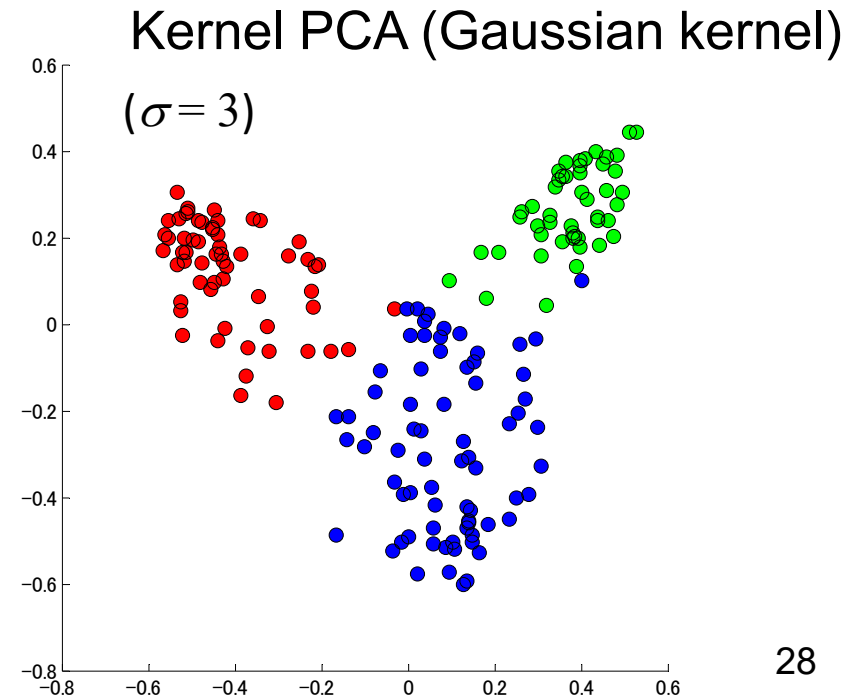
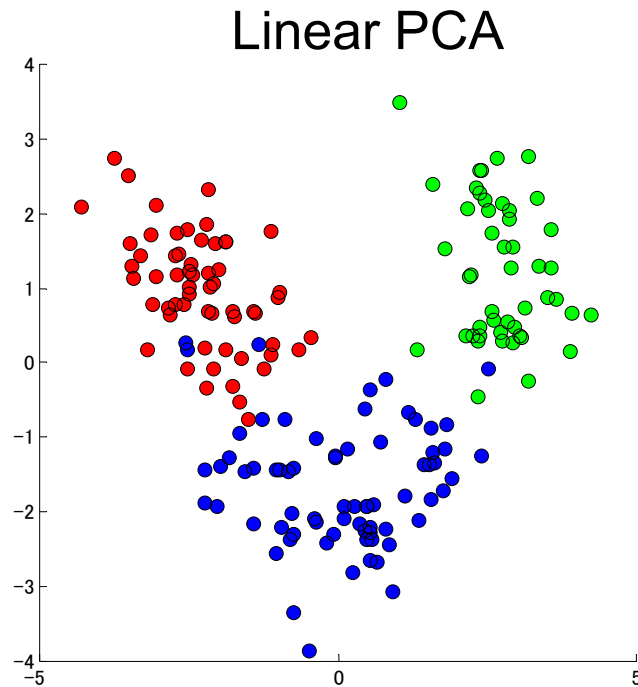
- 第 p 主成分方向 $f_p = \sum_j \frac{1}{\sqrt{\lambda_p}} u_{pj} \tilde{\Phi}(X^{(j)}),$

$$\tilde{\Phi}(X^{(j)}) = \Phi(X^{(j)}) - \frac{1}{N} \sum_{b=1}^N \Phi(X^{(b)}): \text{中心化特徴ベクトル}$$

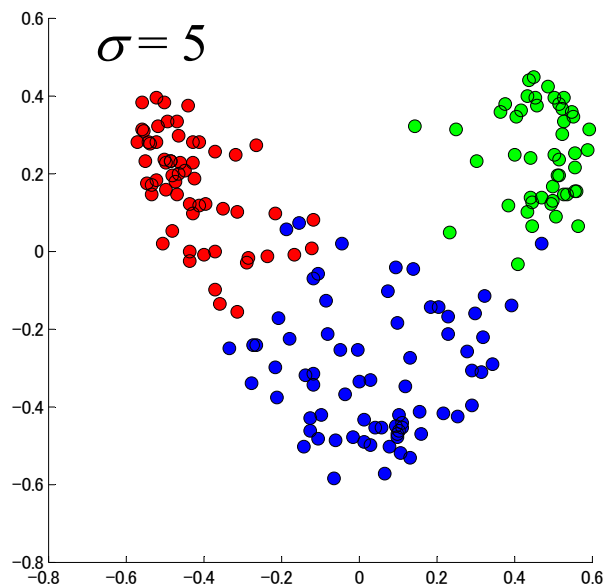
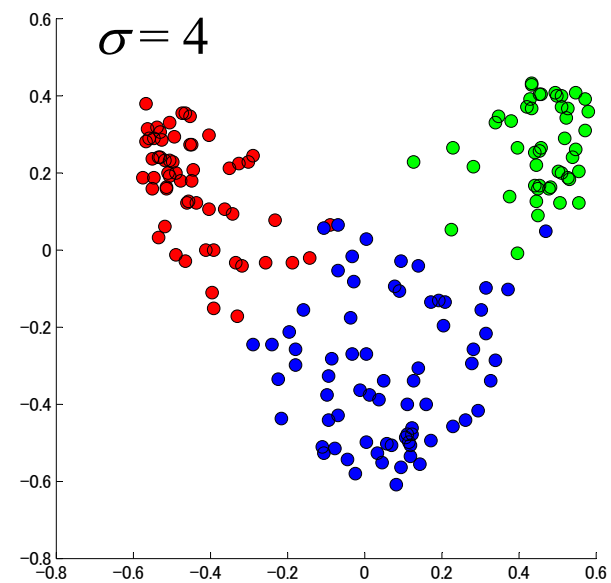
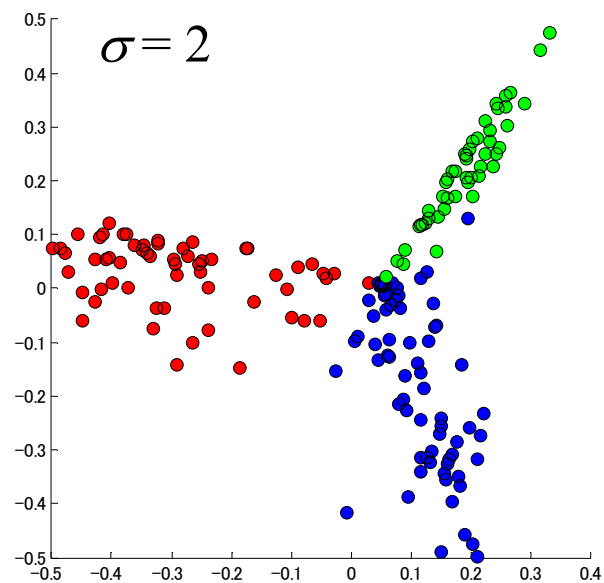
- $X^{(i)}$ の第 p 主成分 $= \langle f_p, \tilde{\Phi}(X^{(i)}) \rangle = \sum_j \frac{1}{\sqrt{\lambda_p}} u_{pj} \tilde{K}_{ji} = \sqrt{\lambda_p} u_{pi}$

カーネルPCAの例

- Wine データ (UCI repository)
クラスの情報はカーネルPCAには用いていない



Kernel PCA (Gaussian)



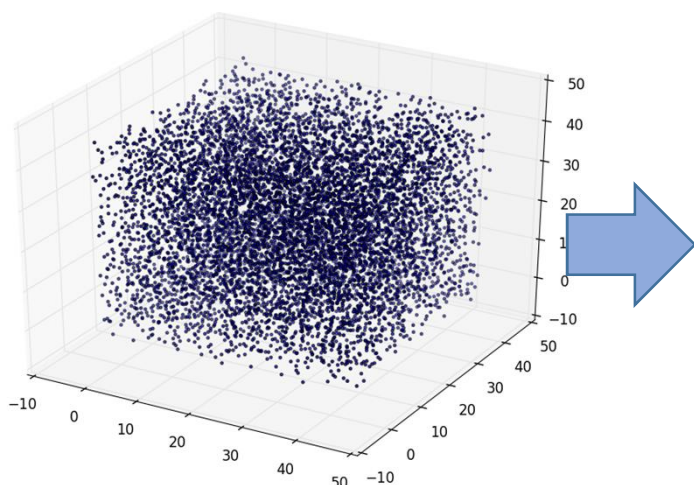
$$k_G(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$$

パーシステント図のデータ解析

PHによる統計的データ解析

- 単純な位相的データ解析

データ



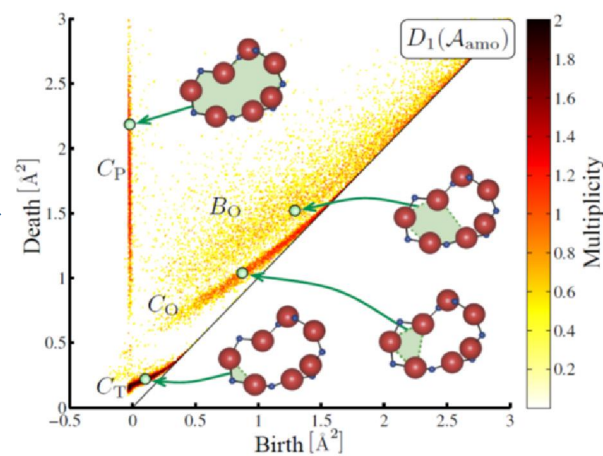
e.g. MDシミュレーション

PHの計算



ソフトウェア

視覚化 (PD)



専門家による分析



• 統計的な位相的データ解析

(Kusano, Fukumizu, Hiraoka ICML2016; Reininghaus et al CVPR 2015; Kwitt et al NIPS2015; Fasy et al 2014)

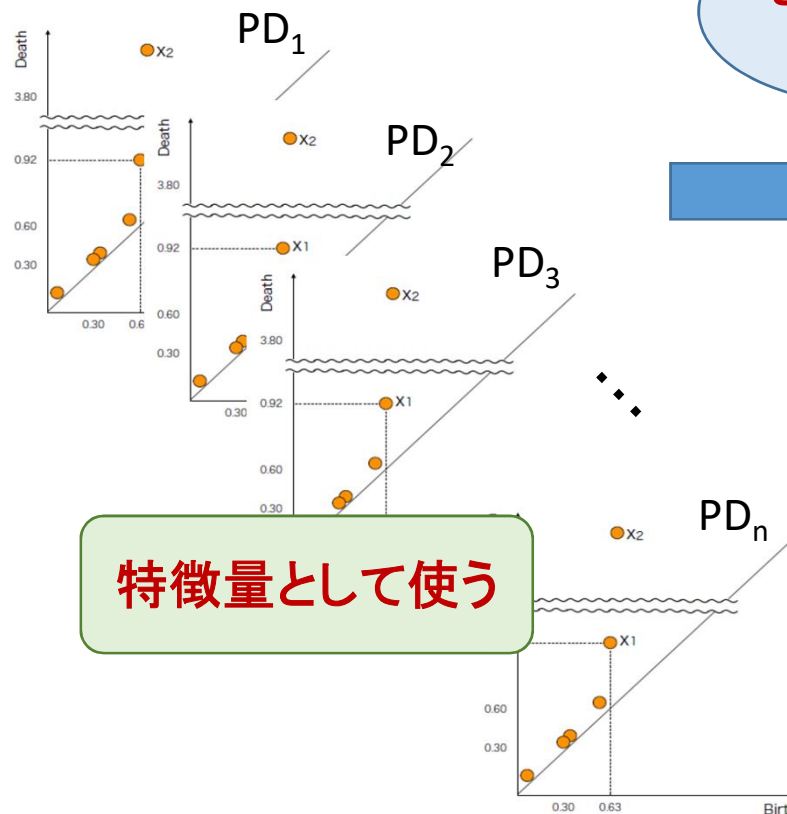
多くの Point Cloud



PHの計算



多くのパーシステント図



どのように
行うか？

カーネル法

パーシステント図に
対するデータ解析

パーシステント図のベクトル化

$$\Pi := \{(b, d) \in \bar{\mathbf{R}}^2 \mid d > b\}, \quad \Delta = \partial\Pi.$$

パーシステント図 $D = D_o \cup \Delta$

$D_o = \{x_i\}$: multiset on Π . (重複度込み)

Γ : $|D_o| < \infty$ なるパーシステント図全体

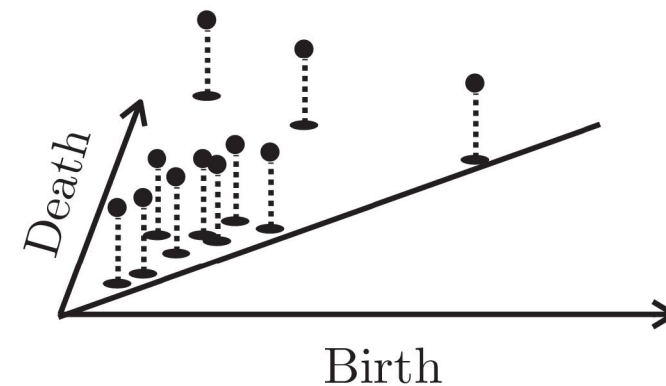
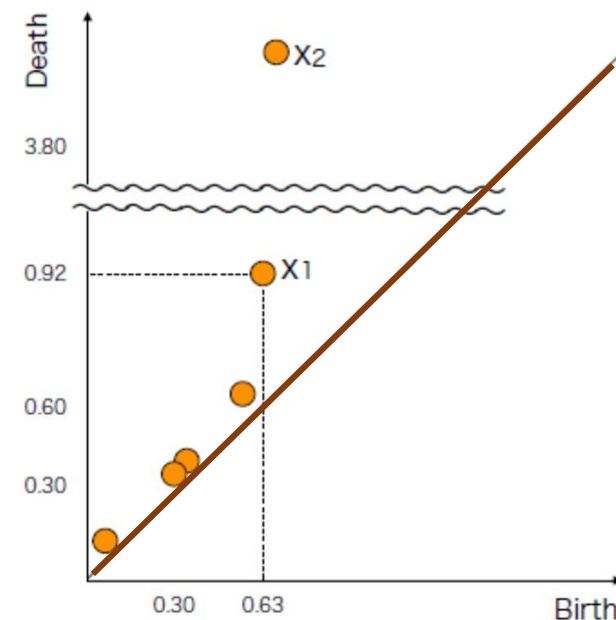
- PD = 離散測度と同一視できる

$$D \leftrightarrow \sum_{x_i \in D_o} \delta_{x_i} \quad \delta: \text{Diracのデルタ関数}$$

- カーネル埋め込み

k : Π 上の正定値カーネル

$$\begin{aligned} \varepsilon_k: \Gamma &\rightarrow H_k, \\ \sum_i \delta_{x_i} &\mapsto \sum_i k(\cdot, x_i) \in H_k \end{aligned}$$



カーネル埋め込み (Muandet, Fukumizu, Sriperumbudur, Schölkopf 2017)

(Ω, B) : 可測空間, k : Ω 上の可測で有界な正定値カーネル.

$\mathcal{M}$: (Ω, B) 上の有限測度の族. 以下の写像を **カーネル埋め込み** と呼ぶ.

$$\varepsilon_k : \mathcal{M} \rightarrow H_k, \quad \mu \mapsto \int k(\cdot, x) d\mu(x)$$

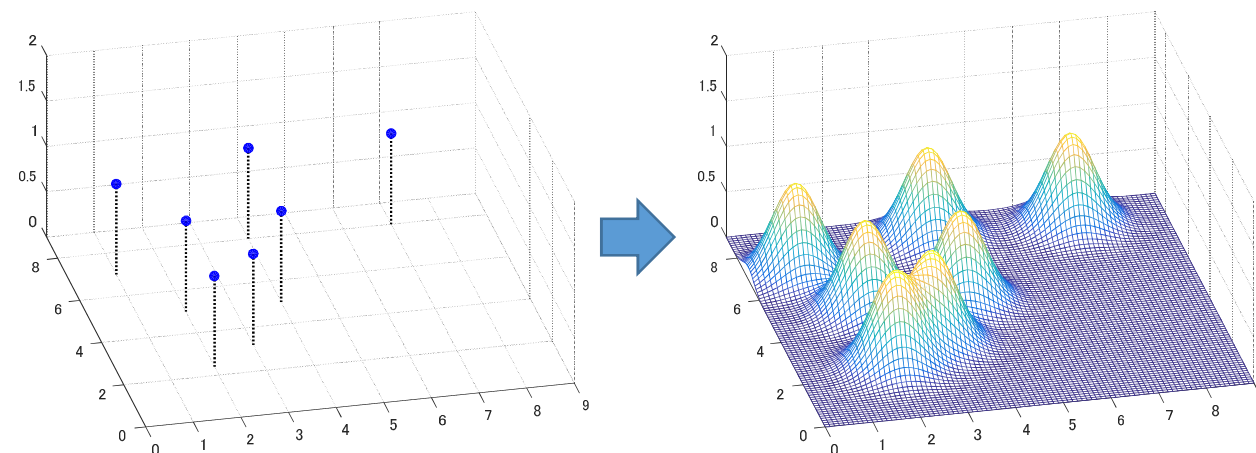
Def. 可測で有界な正定値カーネル k が **$\mathcal{M}$ -特性的 (Characteristic)** であるとは, ε_k が単射であることをいう.

- Ω が $\mathbf{R}^m$ の部分集合で $\mathcal{M}$ がBorel測度全体の場合は, 単に「特性的」ということにする.

例) ガウスカーネルの場合

$$w(x)dx \mapsto \int \exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right) w(x)dx \quad (\text{積分変換})$$

$$\sum_{i=1}^{\ell} \delta_{x_i} \mapsto \sum_{i=1}^{\ell} \exp\left(-\frac{\|y-x_i\|^2}{2\sigma^2}\right)$$



- ベクトル化の実用的メリット

- ガウスカーネルなどを用いると，単射なので測度の識別性は保たれる．
- ベクトルデータに対する多くのデータ解析手法が適用できる
- カーネルトリックにより，計算はグラム行列計算に還元される．

特性的なカーネルの特徴付け

定理 (Bochner)

ψ を $\mathbf{R}^m$ 上の連続関数とする時, ψ (**C**-値)正定値関数であることと, $\mathbf{R}^m$ 上の有限非負Borel測度 Λ が存在して

$$\psi(z) = \int \exp(\sqrt{-1}\omega^T z) d\Lambda(\omega)$$

と書けることは同値である. このとき Λ は一意に決まる.

定理 (Sriperumbudur, Gretton, Fukumizu, Scholkopf, Lanckriet 2011)

$\mathbf{R}^m$ 上の連続で平行移動不変な(**C**-値)正定値カーネル $k(x, y) = \psi(x - y)$ (ψ は正定値関数で上の積分表現を持つ) に対し, k が特性的であることと $\text{Supp}(\Lambda) = \mathbf{R}^m$ とは同値である.

Proof idea: $\psi * \mu = 0 \iff \hat{\psi}\hat{\mu} = 0$. Example: ガウスカーネル

Persistence Weighted Gaussian Kernel : PD用のカーネル (Kusano, Fukumizu, 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
→ 重みを小さくする

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

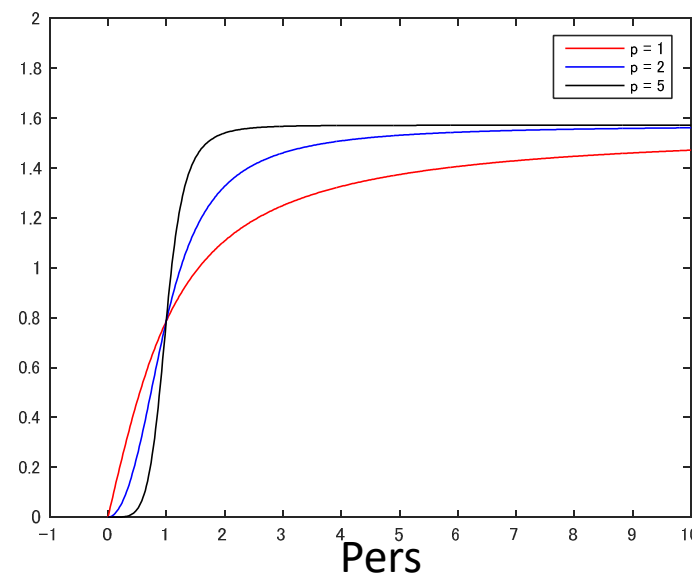
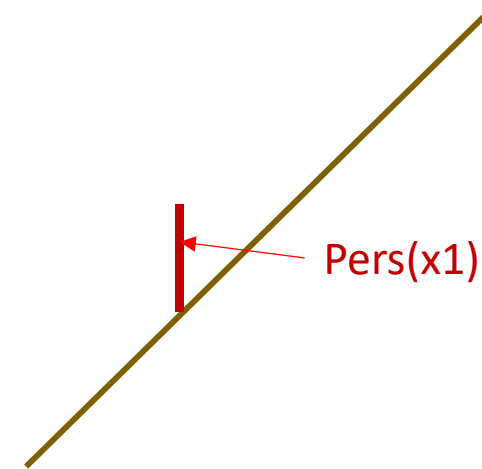
$$x = (x_1, x_2), y = (y_1, y_2) \in \mathbf{R}^2$$

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

$(C, p > 0)$

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

PWKGによる埋め込みも特性的

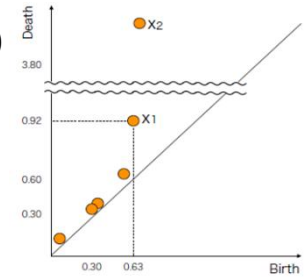


距離関数の安定性

Point Cloud



PD



点集合の微小な変化はPDの距離に微小な変化し与えない

- 安定性(stability)

d : PD上の距離 e.g. bottleneck distance

点集合上の距離として Hausdorff距離 D_H を考える
定義

d が安定とは, ある定数 L が存在し, 任意の点集合 $X = \{x_i\}, Y = \{y_j\}$ に対し

$$d(PD(X), PD(Y)) \leq L \cdot D_H(X, Y)$$

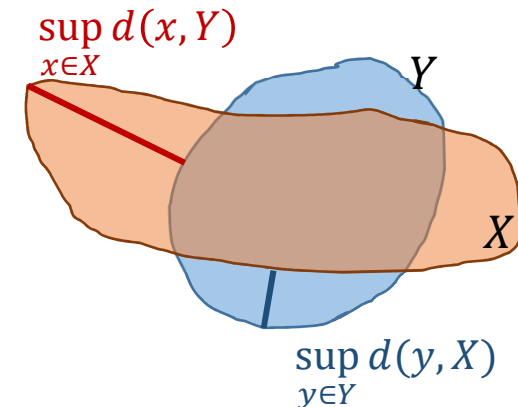
が成り立つことをいう (Lipschitz連続性).

- Hausdorff距離

X, Y : 距離空間 (Ω, d) の集合,

D_H : Hausdorff距離

$$D_H(X, Y) = \max\left\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(y, X)\right\}$$



- 定理 (Bottleneck距離の安定性 Cohen-Steiner et al 2007; Chazal et al 2014)
 X, Y : 有限集合, $D_q(X), D_q(Y)$: 対応する q 次パーシステント図 (q 任意)

$$d_B(D_q(X), D_q(Y)) \leq D_H(X, Y).$$

- PWGK の決める距離 $d_k(D_1, D_2) := \|\varepsilon_k(D_1) - \varepsilon_k(D_2)\|_{H_k}$

定理 (PWGK距離の安定性. Kusano, Hiraoka, Fukumizu ICML2016)

$M: \mathbf{R}^d$ のコンパクト集合. $X \subset M, Y \subset \mathbf{R}^d$: 有限集合.

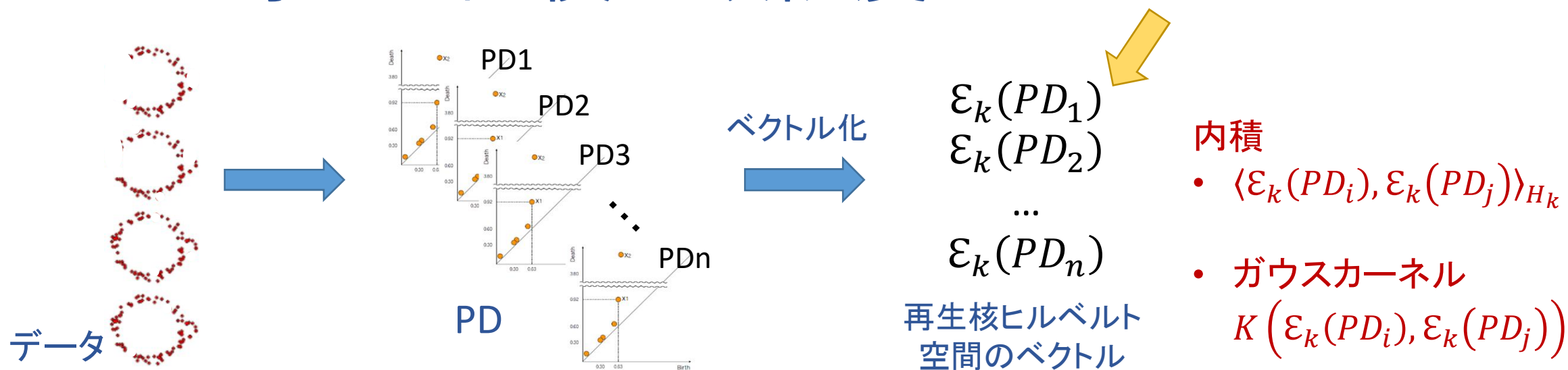
$p > d + 1$ ならば, PWG カーネル $k_{p,C,\sigma}$ に対し

$$d_k(D_q(X), D_q(Y)) \leq A D_H(X, Y).$$

ここで A は M, p, d, C, σ のみに依存する定数 (X, Y には依存しない)

- ガウスカーネルでは安定性は知られていない.

PDに対する内積・類似度



- 再生核ヒルベルト空間上のガウスカーネル $K(\varphi_1, \varphi_2) = \exp\left(-\frac{\|\varphi_1 - \varphi_2\|_{H_k}^2}{2\tau^2}\right)$

$$K(PD_i, PD_j) = \exp\left(-\frac{\|\epsilon_k(PD_i) - \epsilon_k(PD_j)\|_{H_k}^2}{2\tau^2}\right)$$

PD_i, PD_j : Persistence diagrams

PWKGの効率的近似計算

- 計算量の問題

PDの点 (PHの生成元) の数が多いこともある ($\geq 10^3, 10^4$)

m 個のPDの解析 $PD_i = \{x_a^{(i)}\}_{a=1}^{N_i}$

$K(PD_i, PD_j) = \exp\left(-\frac{\|\varepsilon_k(PD_i) - \varepsilon_k(PD_j)\|_{H_k}^2}{2\tau^2}\right)$ の計算には

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

$\exp\left(-\frac{\|x_a - x_b\|^2}{2\sigma^2}\right)$ の計算回数 = $O(m^2 N^2)$ $\rightarrow N \approx 10^4$ では計算量莫大

$$N = \max\{N_i | i = 1, \dots, m\}$$

- Random Fourier feature による効率的近似 (Rahimi & Recht 2008)

Bochnerの定理

ガウス分布 $=: Q_\sigma$

$$\exp\left(-\frac{\|x_a - x_b\|^2}{2\sigma^2}\right) = C \int \underbrace{e^{\sqrt{-1}\omega^T(x_a - x_b)}}_{\substack{\left(\frac{\sigma^2}{2\pi}\right) e^{-\frac{\sigma^2 \|\omega\|^2}{2}}}} d\omega$$

積分（平均）をMonte Carlo サンプリングで近似: $\omega_1, \dots, \omega_L: i.i.d. \sim Q_\sigma$

$$\exp\left(-\frac{\|x_a - x_b\|^2}{2\sigma^2}\right) \approx C \frac{1}{L} \sum_{\ell=1}^L e^{\sqrt{-1}\omega_\ell^T x_a} \overline{e^{\sqrt{-1}\omega_\ell^T x_b}}$$

$$\sum_{a=1}^{N_i} \sum_{b=1}^{N_j} k(x_a^{(i)}, x_b^{(j)}) \approx \frac{C}{L} \sum_{\ell=1}^L \underbrace{\sum_{a=1}^{N_i} w(x_a^{(i)}) e^{\sqrt{-1}\omega_\ell^T x_a^{(i)}}}_{L \text{ dim.}} \overline{\sum_{b=1}^{N_j} w(x_b^{(j)}) e^{\sqrt{-1}\omega_\ell^T x_b^{(j)}}}$$

$L \text{ dim.}$

計算量 $O(LN)$ → 2層目のGram 行列計算 $O(mLN + m^2L)$. c.f. $O(m^2N^2)$
 $L, m \ll N$ なら大きなゲイン

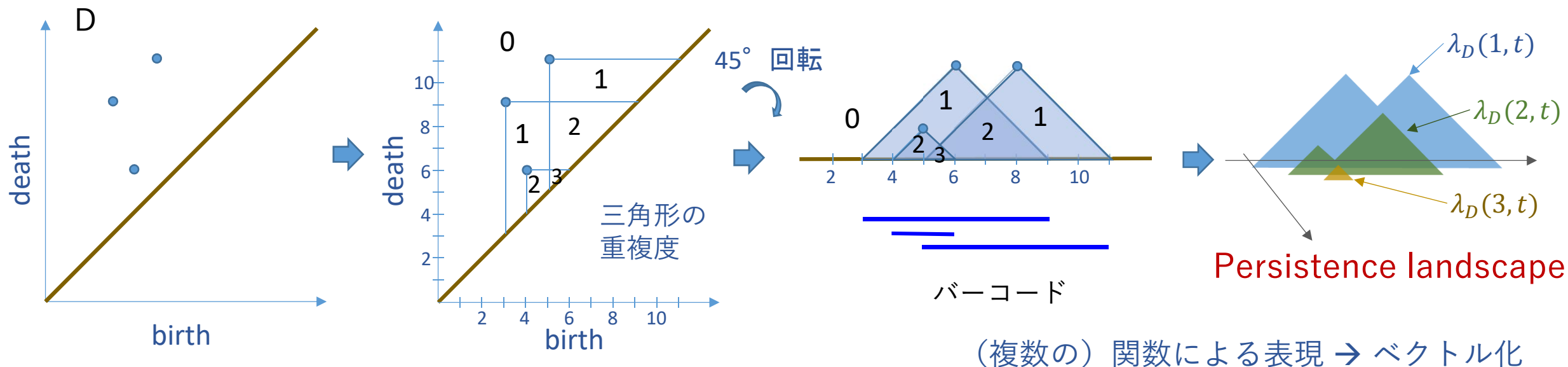
他のベクトル化：Persistent Landscape

- Persistent Landscape (Bubenik, JMLR 2015)

$$\lambda_D(k, t) = \text{kmax}_{(b,d) \in D} \Lambda_{(b,d)}(t)$$

kmax: k 番目に大きい値

$$\Lambda_{(b,d)}(t) := \begin{cases} t - b, & t \in [b, (b+d)/2] \\ d - t, & t \in [(b+d)/2, d] \\ 0, & \text{otherwise} \end{cases}$$



- Persistence landscape に基づく内積・距離
 $D_1, D_2 : (q\text{次元})$ パーシステント図

$$(D_1, D_2)_{PL} := \sum_{k=1}^{\infty} \int \lambda_{D_1}(k, t) \lambda_{D_2}(k, t) dt \quad (\text{L2内積})$$

$$d_{PL,p}(D_1, D_2) := \left(\sum_{k=1}^{\infty} \left\| \lambda_{D_1}(k, t) - \lambda_{D_2}(k, t) \right\|_p^p \right)^{1/p} \quad (\text{Lp距離})$$

- Persistence landscape 一つの計算量 $= O(N^2)$ N : 生成元の数
 (Bubenik & Dłotko, 2017)

Persistence Scale-Space Kernel

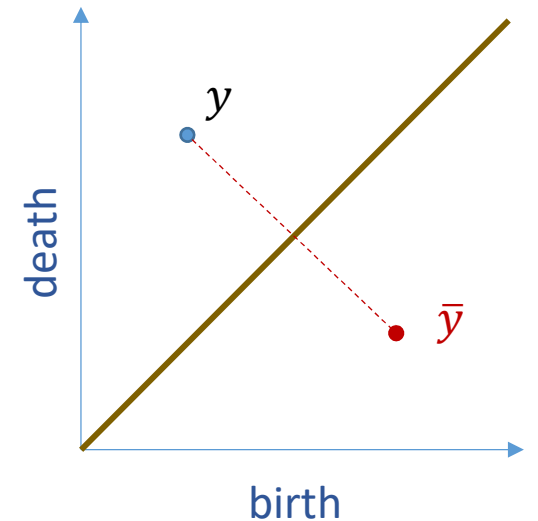
Reininghaus, S. Huber, U. Bauer, and R. Kwitt, CVPR2015

- PSSカーネル：熱方程式から着想

$$k_{PSS}(D_1, D_2) = \frac{1}{8\pi t} \sum_{x_i \in D_1} \sum_{y_j \in D_2} \exp\left(-\frac{\|x_i - y_j\|^2}{8t}\right) - \exp\left(-\frac{\|x_i - \bar{y}_j\|^2}{8t}\right)$$

$y = (b, d)$ に対して $\bar{y} = (d, b)$

- x_i または y_j が対角線に近づくと和の成分は0に近づく.
PWGKと同様の効果
- 実は, PWGKでweightを特別に取ったものに一致.
- ガウスのバンド幅と寿命による重みが1つの
パラメータ t で制御されている → 自由度低い (59p参照)



さまざまなPDの距離

- Bottleneck distance d_B (計算量 $O(N^{2.5})$ (d'Amico 2006))
- カーネル埋め込みによる距離 d_k (計算量 $O(N^2)$, 近似により $O(N)$)
- Persistence Landscapeによる距離 d_{PL} (計算量 $O(N^2)$) N : 生成元の数

- Wasserstein distance of degree p (一般の非負測度に対する距離)

$$W_p(D_1, D_2) := \inf_{\gamma: D_1 \rightarrow D_2} \left(\sum_{x \in D_1} \|x - \gamma(x)\|_\infty^p \right)^{\frac{1}{p}}$$

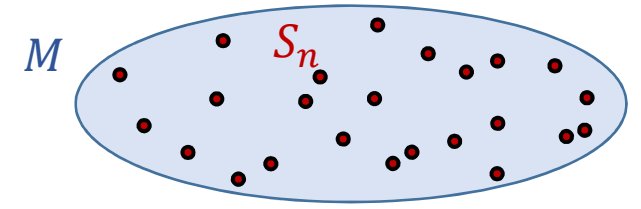
γ : 対角線および重複度を込めた全単射

Fact: $W_\infty = d_B$

計算量: $O(N^3)$ (N : 生成元の数, Kuhn-Munkres アルゴリズム,

- 安定性: W_p, d_{PL} に対しても安定性は知られているが, 弱い形 (厳密な Lipschitz 連続ではない. 詳細は Cohen-Steiner et al 2010; Bubenik 2015) .

信頼集合



- リサンプリングによる信頼集合の構成

P : 確率測度, 連続な密度関数 $p(x)$. $M = \text{Supp}(P)$. $S_n: X_1, \dots, X_n \sim P, \text{i.i.d.}$

Goal: $PD(S_n)$ は確率変数 $\rightarrow d_k(PD(S_n), PD(M))$ を評価せよ.

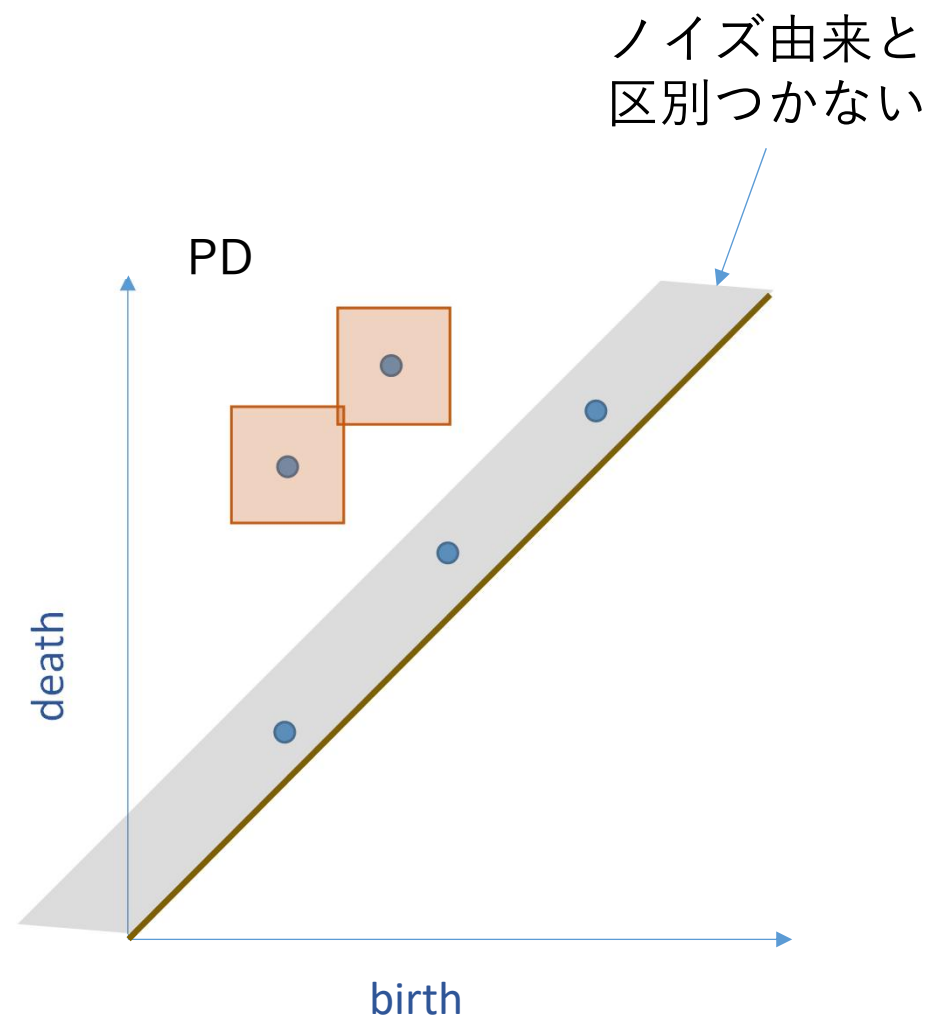
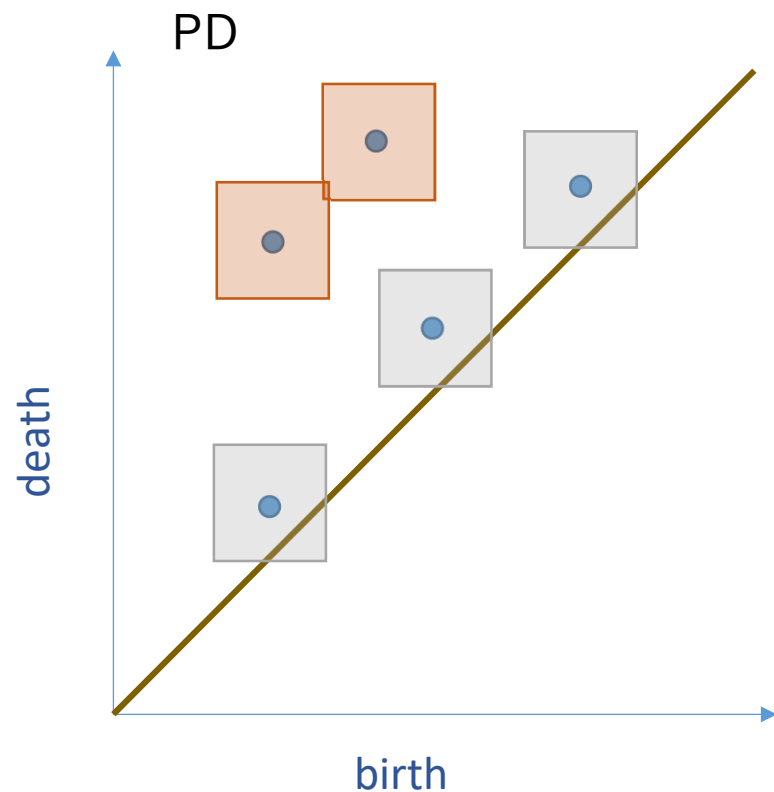
Idea: 安定性 $d_k(PD(S_n), PD(M)) \leq A d_H(S_n, M)$ を用いる.

$S_{b,n}^{(\ell)} := \{X_{i_1}, \dots, X_{i_b}\}$: S_n からのサブサンプル. $\ell = 1, \dots, \binom{n}{b} =: N$ (サイズ b サブサンプルの総数)

n が大きいとき, 高い確率で $d_H(S_n, M) \leq d_H(S_{b,n}^{(\ell)}, S_n)$.

$\tilde{L}_b(t) = \frac{1}{N} \sum_{\ell=1}^N I(d_H(S_{b,n}^{(\ell)}, S_n) > t)$ リサンプリングによる d_H の累積分布関数は推定可能.

確率 $\geq 1 - \alpha + o(1)$ で,
各生成元 $z \in PD(S_n)$ に対し $x \in PD(M)$ が
存在する.



PWKGでも同様の議論が可能

カーネル法によるPD解析の応用例

人工データでの基礎実験

- 2値識別問題

- 点集合：大きな円 $S1$ 上の点
小円 $S0$ 上の点は、ある場合とない場合がある.

- 正解 $Y = \text{XOR}(Z_1, Z_2)$

- Z_1 : $S0$ が存在するか Yes/No

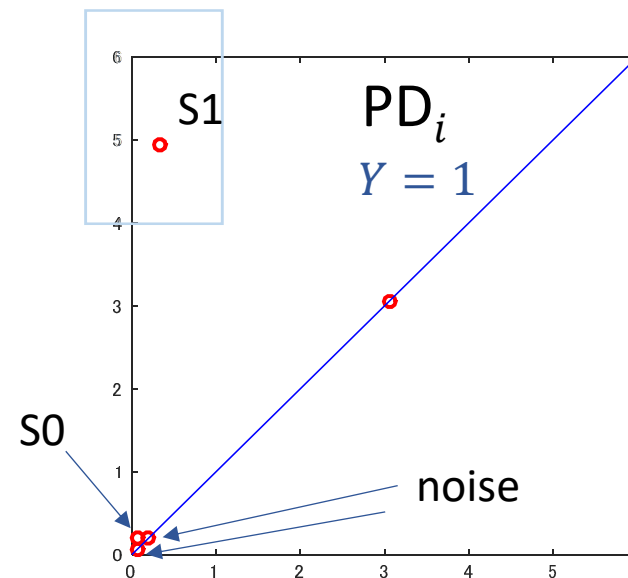
- Z_2 : $S1$ に対応する生成元が

$((b(S1) < 1 \ \&\& \ d(S1) > 4)? \text{ Yes/No}$ 大きなスケール

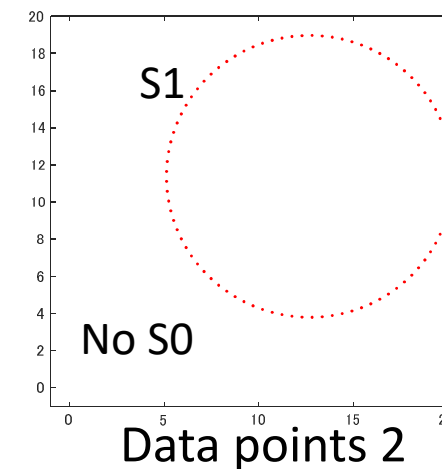
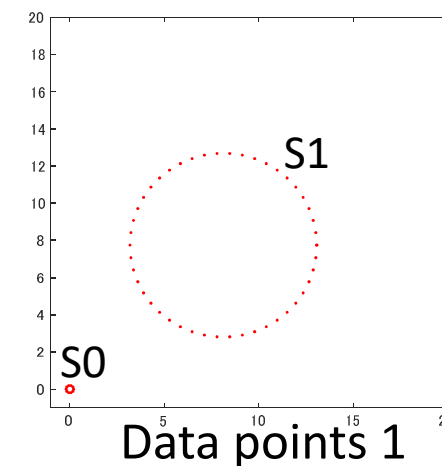
- ノイズを付加.

- 訓練データ100セット,
テストデータ100セット

- SVM**で識別機構成



小さなスケール



- テストデータに対する正解率

埋込み用のカーネル		2層目のカーネル	
Kernel	Weight	Linear	Gauss
$k_{PWG}(x, y) = w(x)w(y)\exp\left(-\frac{\ y-x\ ^2}{2\sigma^2}\right)$	PWGK: $p = 1$	51.5	83.8
	PWGK: $p = 5$	50.5	84.8
	Weighted Gauss PWGK: $p = 10$	49.4	84.8
	w_{pers}	56.5	57.5
	$w = 1$ (Gauss)	51.2	51.5
$k_{Lin}(x, y) = w(x)w(y)x^T y$	atan $p = 1$	49.4	52.5
	atan $p = 5$	50.5	50.5
	Weighted Linear atan $p = 10$	50.5	51.5
	w_{pers}	49.4	51.5
	$w = 1$ (Linear)	48.4	57.5
PSSK (K_{PSS})		50.5	58.5
Persistence Landscape (K_{PL})		52.5	54.5

$w_{pers} := Pers(x)$

応用例 1：タンパク質の識別

- タンパク質の幾何的形状 → 機能に大きな役割
- 幾何的形状をパーシステンスホモロジー
／パーシステント図（PD）で表現
- PDをデータとして識別器を構成.
サポートベクターマシン（SVM）
識別問題の defacto standard.

- データ A : Protein-drug 結合

Figure omitted

- A 型インフルエンザウイルスのM2 channel
抗インフルエンザ薬のターゲット

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

- Rimantadine が結合しているか否かを,
M2 channel のNMR(核磁気共鳴)による立体構造データから識別.
 - 結合／非結合それぞれ15データ.
 - ランダムに選んだ20データ (各10) でS V Mを訓練. 残りでテスト.
1 0 0 ランダムセット行う.

- データ B : ヘモグロビンの 2 状態の識別

- X線回折による立体構造データを用いて, 2 状態 Relaxed (R) / Taut (T) を識別.

- R : 9データ, T : 10データ.

- 各クラス 1 データずつをテストに残し,
残りで訓練, 2 データでテスト.
全ての組み合わせに対して行う.

Figure omitted

Relaxed (R)

Taut (T)

Cang, Mu, Wu, Opron, Xia, Wei, *Molecular Based Mathematical Biology* (2015) Fig. 4

- 比較：

Cang et al (2015)では, PHに対して, Molecular Topological Fingerprint (MTF) と呼ぶ経験的な13個の記述子を定義.

#	Dim	Description
1	0	2番目に長い生成元の寿命
2	0	3番目に長い生成元の寿命
3	0	寿命の総和
4	0	寿命の平均
5	1	最長生成元の生成点
6	1	最長生成元の寿命
7	1	長さ1.5Å以上の中の最短生成元の生成点
8	1	長さ1.5Å以上の生成元の間点の平均
9	1	[4.5, 5.5]Åに存在する生成元の数／総原子数
10	1	[3.5, 4.5)Å と (5.5, 6.5]Åに存在する生成元の数／総原子数
11	1	寿命の総和
12	1	寿命の平均
13	2	最初の生成元の生成点

- PWGK + ガウスカーネルSVM

PWGK : $p = 5$, C とカーネルバンド幅はクロスバリデーションで選択

- 識別結果

クロスバリデーションによる平均正答率

	A. Protein-Drug	B. Hemoglobin
PWGK	100	88.90
MTF*	(nbd) 93.91 / (bd) 98.31	84.50

- MTF の結果は, Cang et al. *Molecular Based Mathematical Biology* (2015) より抜粋.
- PWGKは, 1次元PHのみ使用.

応用例 2 : 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。

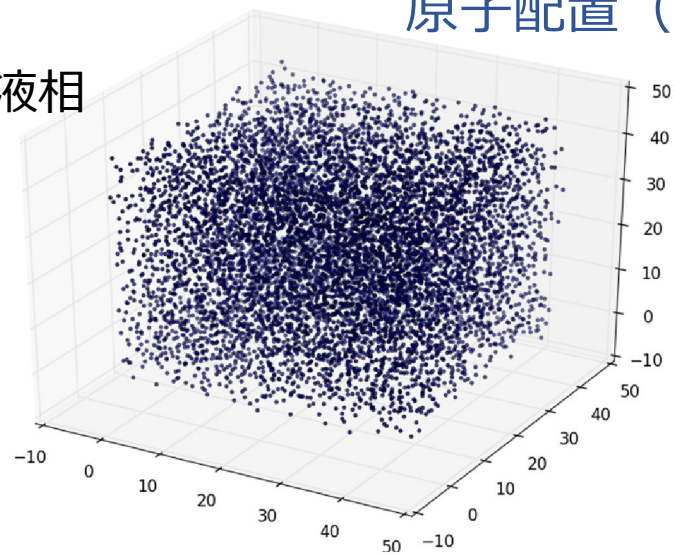
アモルファスの記述子：
標準的なものがない

データ: SiO_2 分子動力学 (MD) シミュレーション
(Nakamura et al 2015 Nanotechnology; Hiraoka et al 2016 PNAS)

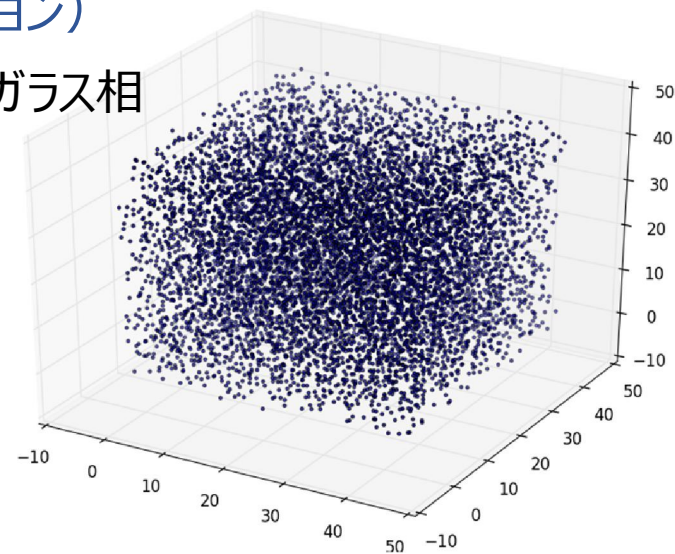
- 80個の温度で、3次元原子配置データを取得 (スナップショット)
- 原子の3次元配置データから、PDを計算。
- Si と O 原子では、異なる半径の球を用いている。
- **物理学的方法**: エンタルピー曲線を描いて、微分の推定を行い、その不連続点を推定する。正確な推定は難しい。
- **提案法**: PD図のカーネル埋め込みに対する、変化点検出問題として転移点を推定する。 (Kusano et al ICML2016)

原子配置 (MDシミュレーション)

液相

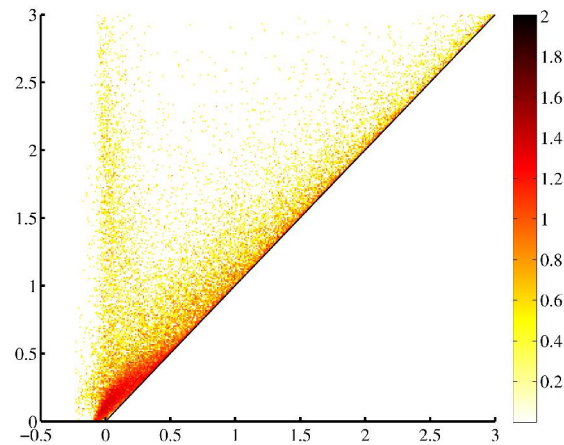


ガラス相

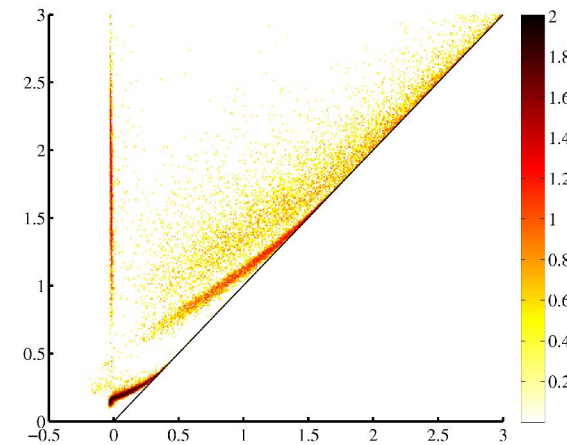


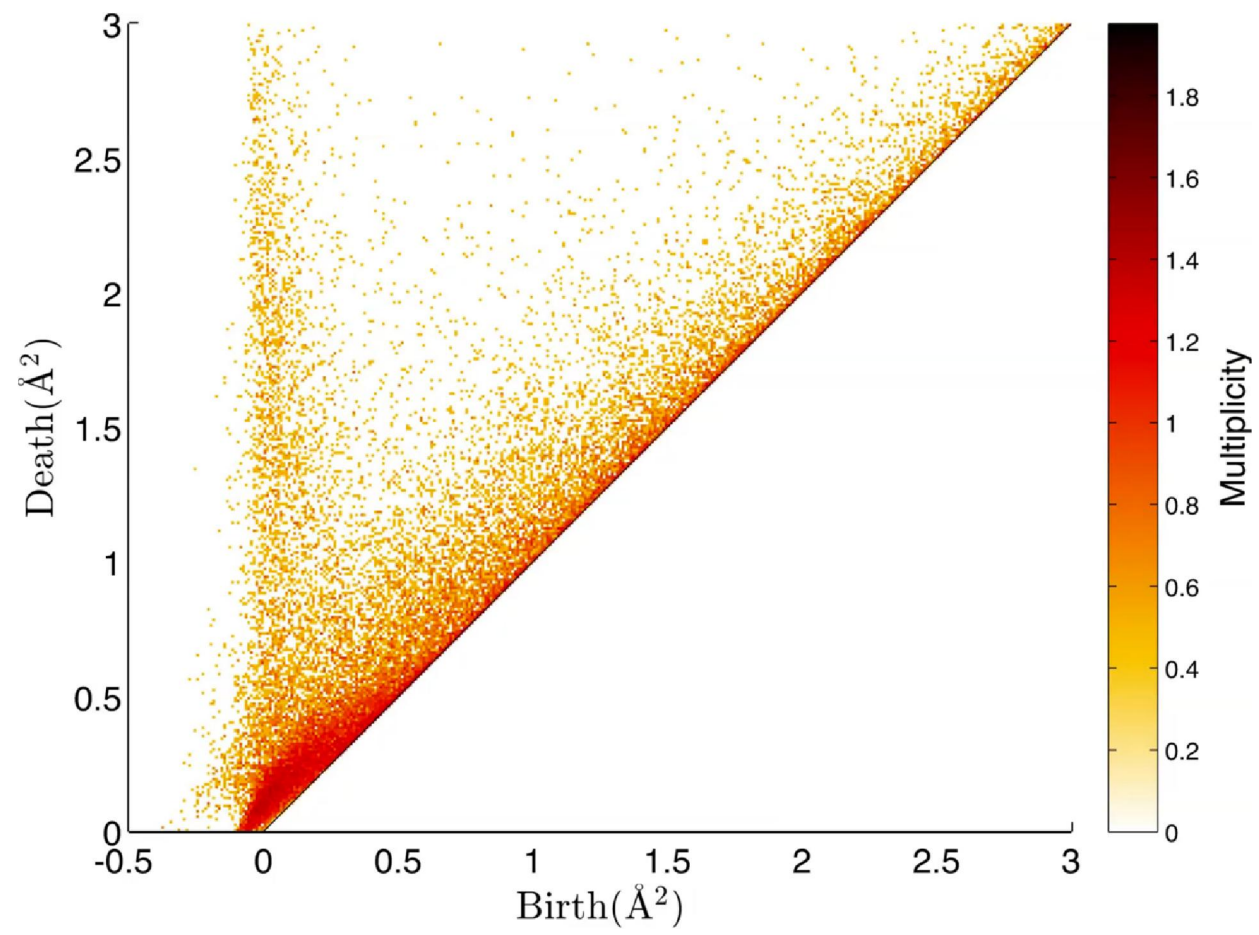
パーシステント図

液相



ガラス相





• 変化点検出（一般的な話）

- パラメータ t に沿ったデータが変化する t_* を見つける.

$$Y_t, t = 1, \dots, T.$$

- カーネル変化点検出法 (Harchoui et al 2009)

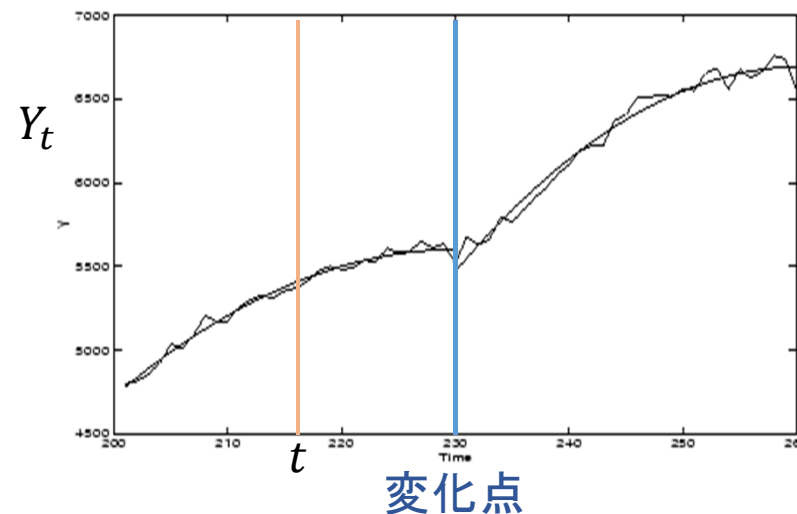
Fisher判別スコアを用いる

- 各 t に対し前後で平均を計算 : $\hat{m}_{1:t} = \frac{1}{t} \sum_{i=1}^t \Phi(X_i)$, $\hat{m}_{t+1:T} = \frac{1}{T-t} \sum_{i=t+1}^T \Phi(X_i)$.

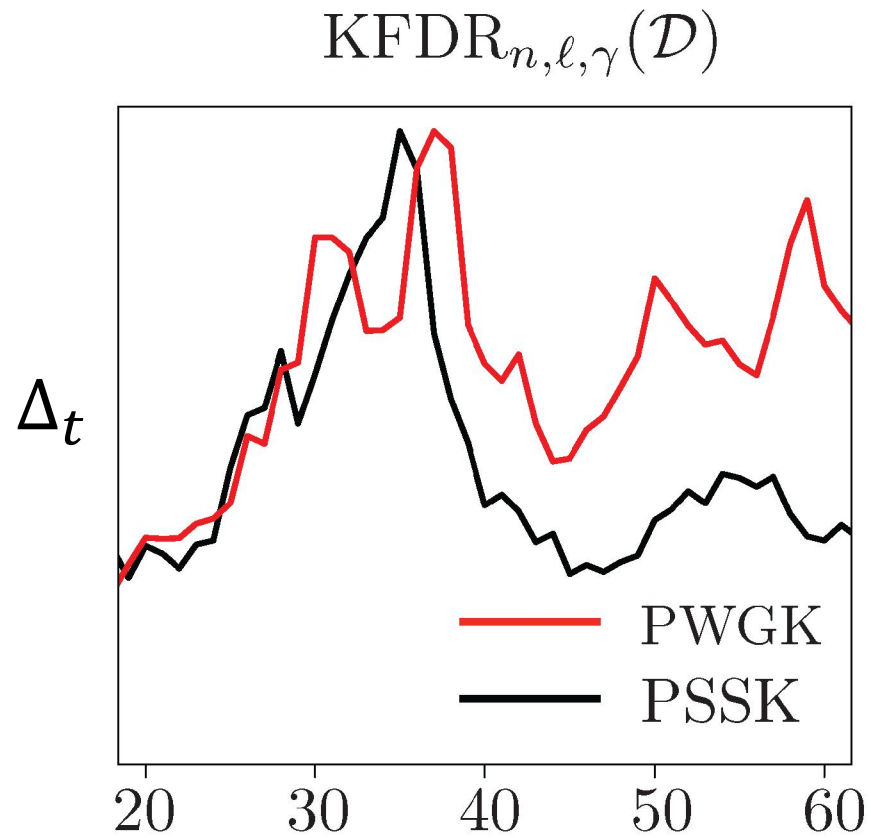
- $\Delta_t := \left\| (V_{1:t} + V_{t+1:T} + \gamma I)^{-\frac{1}{2}} (\hat{m}_{1:t} - \hat{m}_{t+1:T}) \right\|_{H_k}^2$. ($V_{1:t}, V_{t+1:T}$: 前後の分散)

- Find $\max_t \Delta_t$.

- ガラス-液相転移点問題の場合は, $Y_t = \varepsilon_k(D_{\varepsilon_t})$ ($t = 1, \dots, 80$). 無限次元



- SiO_2 の転移点検出

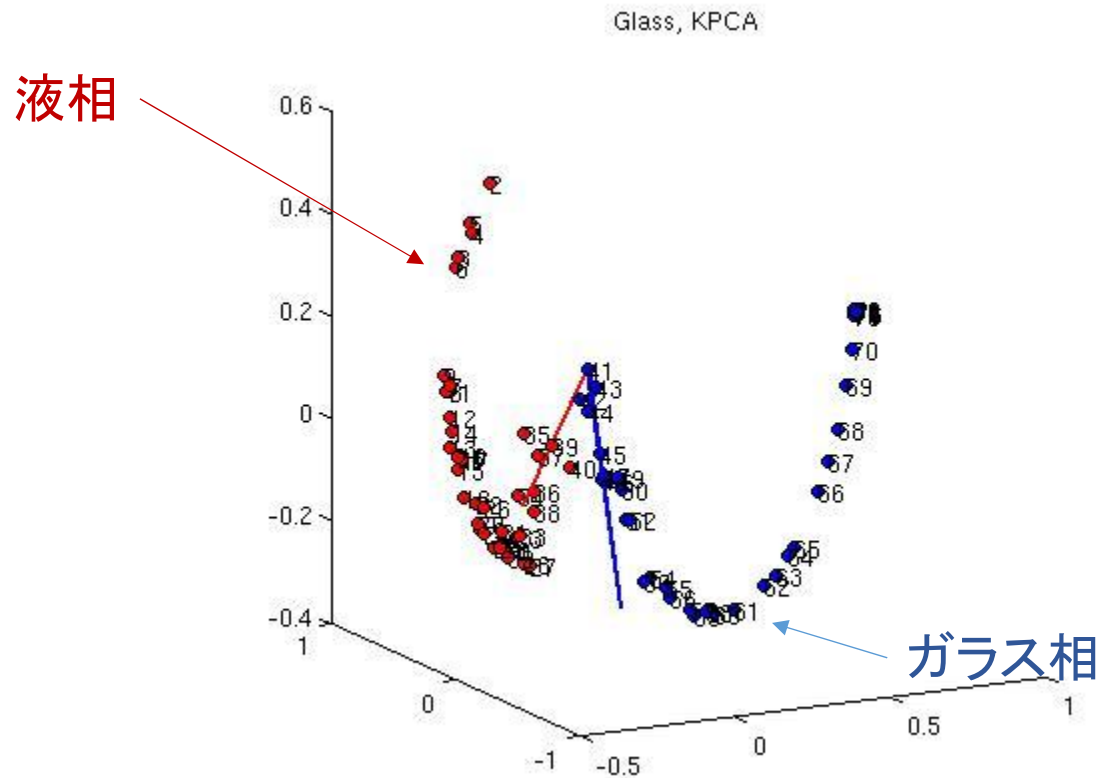


検出された変化点 = 3100K

Enthalpyによる方法: [2000K, 3500K]

- 低次元表現： カーネル主成分分析

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， 3次元で表示.



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.

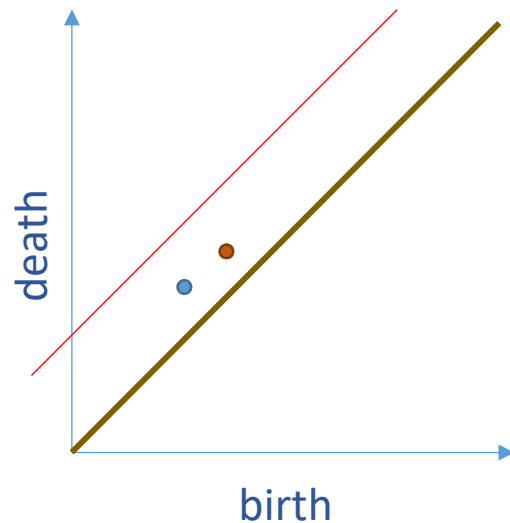
まとめと展望

まとめ

- カーネル法によるPDのベクトル化 → さまざまなデータ解析
 - 多数のパーシステント図に対する統計的データ解析
 - PWGK：生成元のlifetimeを考慮した柔軟なカーネル
 - さまざまな標準的データ解析手法が利用可能
 - 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 …

展望

- PDの信頼度
 - 現状の信頼集合の議論はPDの距離に基づく
 - 短い寿命だが統計的に安定な生成元の信頼度は計れない.



各生成元の信頼度をどうはかるか？

- 安定性の不等式を超えた解析（ランダムPDの分布）が可能か？
Signal/noiseの分離

- 時系列

- 代数的： 

- 統計的： $PD_t, PD_{t+1}, PD_{t+2}, \dots$  統計的な時系列解析

Reference

- Persistence weighted Gaussian kernel for topological data analysis.
Genki Kusano, Yasuaki Hiraoka, Kenji Fukumizu.
[Conference paper] Proceedings of The 33rd International Conference
on Machine Learning, PMLR 48:2004-2013, 2016.
<http://proceedings.mlr.press/v48/kusano16.html>
[Full journal version] [arXiv:1601.01741](https://arxiv.org/abs/1601.01741) [math.AT]