離散的モース理論に基づくパーシステントホモロジー群による mapper 解析の一般化

吉原凌[1]　佐藤好久[2]

概要：近年注目されている新しいデータ解析手法として、Topological Data Analysis（以下、TDA)があり、これは位相幾何学に基づく解析技術である。TDA はデータの持つ形に注目した解析手法で、データの分布モデルを仮定する必要がなく、ビッグデータなどの適切な分布モデルを仮定することができないデータに対して新しい知見を与えることが期待される。TDA の技術の中にはモース理論を基にした mapper 解析という技術がある。著者は本論文にて、離散的モース理論を基に mapper 解析を一般化した新たな TDA 解析を提案する。提案手法は mapper 解析の高次元パーシステントホモロジー群による拡張になっている。提案手法によって元データの特徴を捉えることができるか確認するために、人工的に作成した(位相的特徴の分かりやすい)データに対して本研究の提案手法を適用する。

キーワード：情報数学 – すべて、確率・統計、計算幾何学

Generalization of mapper by persistent homology based on discrete morse theory

Ryo Yoshihara[1]　Yoshihisa Sato[2]

Abstract: Topological Data Analysis (TDA) is a new data analysis method that has been attracting attention in recent years, and is an analysis technique based on topology. TDA is an analysis technique that focuses on the shape of data, eliminating the need to assume a distribution model for the data, and is expected to provide new insights into data for which it is impossible to assume an appropriate distribution model, such as big data. One of the TDA techniques is mapper analysis, which is based on Morse theory. In this paper, the author proposes a new TDA analysis that generalizes mapper analysis based on discrete Morse theory. The proposed method is an extension of mapper analysis with a macroscopic persistent homology group. In order to confirm whether the proposed method can capture the features of the original data, we apply the proposed method to artificially created data (with easily recognizable topological features).

Keywords: Informational Mathematics – All, Probability & Statistics, Computational Geometry

1　九州工業大学大学院情報工学府学術情報工学専攻
2　九州工業大学大学院情報工学研究院

1. はじめに

現在、世界各国で新型コロナウイルスの感染が広がっている。その感染力は凄まじく日本でも緊急事態宣言が発出されるほどである。その症状は年齢や健康状態など人によって様々であるが、最悪死に至るほどである。時間が経てばワクチンによる予防免疫の浸透でいずれは収束すると考えられるが、少しでも早く感染を抑え、経済活動の復興に役立つためのデータ解析を行うことが必要とされている。

近年注目されている新しいデータ解析手法として、Topological Data Analysis (以下、TDA)というものがあり[1]、これは位相幾何学に基づく解析技術である。有名なAYASDIの言葉として「Data has shape, Shape has meaning, Meaning drives value」という言葉があり、TDAの本質を表している。この言葉が意味するように、データは形を持っており、その形には意味があり、価値を生む。TDAはこのようなデータの持つ形に注目した解析手法で、データの分布モデルを仮定する必要がなく、また、ビッグデータなど視覚化できないような高次元のデータの解析も行うことができる。さらに、TDAの技術の中にはモース理論を基にしたmapper解析[2,3]と言われるものが存在する。

以上の背景から、本論文では、離散的モース理論に基づくmapper解析の考え方を用いて、データセットを作り、TDAを適用し、それを解析することにより新たなTDA解析を開発することを目的とする。提案手法はmapper解析の高次元パーシステントホモロジー群の拡張になっている。提案手法によって元データの特徴を捉えることができるか確認するために、人工的に作成した(位相的特徴の分かりやすい)データに対して提案手法を適用する。

2. パーシステントホモロジー群

パーシステントホモロジー群はTDAにおいて重要な概念である。パーシステントホモロジー群は図形の連結成分や輪っか、空洞といった構造に注目することで、データの形としての情報を抽出することができる。パーシステントホモロジー群の基本的な概念を説明する。

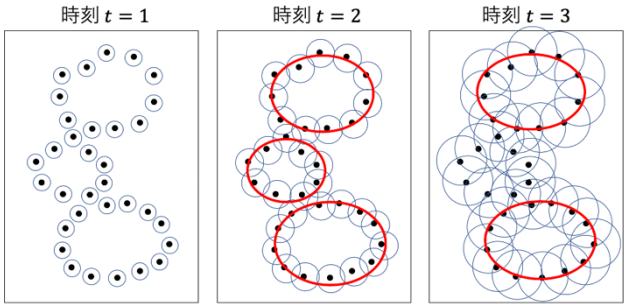

図 1：パーシステントホモロジー群の基本的な概念

図1左のように、点それぞれに半径を考える。時間経過で半径を大きくしていくと、時刻t=2の時点で3つの輪っかが生成される。時刻t=3では中央の輪っかが消滅している。このように図形の次元ごとに増大列(フィルトレーション)を考え、各次元の穴の発生時刻(birth)と消滅時刻(death)をペアリングして1つの図に表示したものをパーシステント図と呼ぶ。

3. モース理論とmapper解析

モース理論によれば、図形の位相的特徴を知るためにはその臨界点(特異点)を見ればよいことになる。mapper解析はそのモース理論を基にしたトポロジカルデータ解析の道具である。mapper解析の基本的な概念を、トーラス図形を例にして説明する。

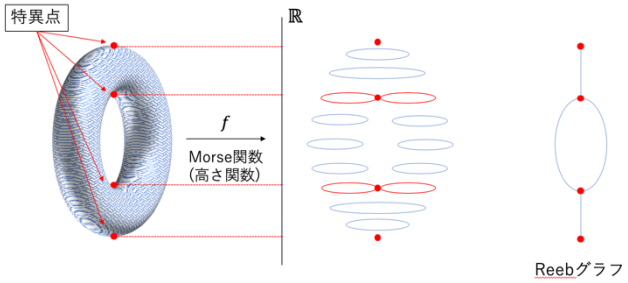

図 2：トーラス図形とReebグラフ

まず、図2のトーラス図形に対する特異点を考える。Morse関数(高さ)を考え、トーラス図形の入った水槽に水を張っていき、その切り口を見ていくイメージをすると、特異点の部分で切り口が変化していることがわかる。その切り口が図2の中央の図になる。Morse関数という1つの軸に沿ってクラスター分析をし、繋がりをエッジで結ぶことで図2の右の図ような Reeb グラフが生成される。これを離散データに対して適用して考えてみる。

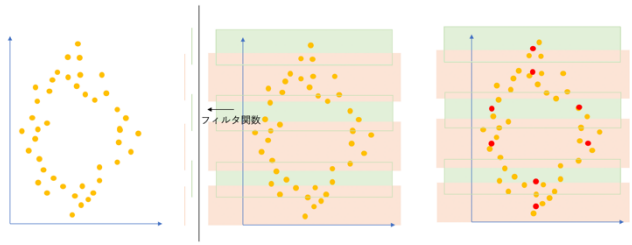

図 3：離散データに対するmapperグラフの生成(1)

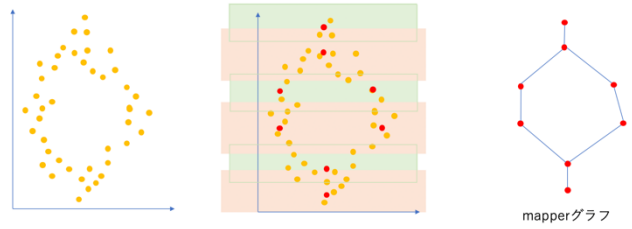

図 4：離散データに対するmapperグラフの生成(2)

まず、図3の左の図が示すような離散データに対して、フィルタ関数を考える。mapper解析において点群に対して与えられた関数をフィルタ関数という。図3の中央の図が示すように、離散データをいくつかに分割し、それぞれに被りが出るようにカバー範囲を考える。そうしてそれぞれの分割でクラスター分析することで、他の分割との繋がりも、繋がりが

あればエッジで結ぶことができる。図 4 の右の図がこの例の mapper グラフにあたる。mapper グラフでは、枝分かれのある頂点に対応する元データに特徴的なものがあると考える。

ここで、mapper 解析に用いられる関数の cover-level set を定義しておく。

実数値関数 $f:\mathbb{R}^n \to \mathbb{R}$ を考える。$r<s$ を満たす実数 $r,s \in \mathbb{R}$ に対して、

$$f^{-1}([r,s]) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n \mid r \le f(x) \le s\}$$

を関数 f の cover-level set という。また、$r<r'$, $s<s'$ に対して、

$$s-r=s'-r'$$
$$k=\frac{s-r'}{s-r} \quad (0<k<1)$$

を満たす。k を cover-level set $f^{-1}([r,s])$ と $f^{-1}([r',s'])$ のカバー比率とよぶ。

4. 提案手法への着想

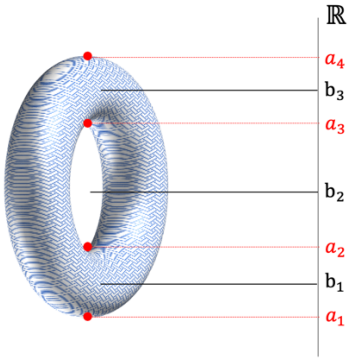

図 5：トーラス図形

図 5 のトーラス T^2 の高さ関数 $f:T^2 \to \mathbb{R}$ に対する cover-level set の 0 次元ホモロジー群 H_0 と 1 次元ホモロジー群 H_1 を考える。ホモロジー群の係数は環 $\mathbb{Z}_2$ とする。

$$M[a_1,b_1]=f^{-1}([a_1,b_1])=\{x \in T^2 \mid f(x) \in [a_1,b_1]\}$$
$$H_0(M[a_1,b_1])=\mathbb{Z}_2,\ H_1(M[a_1,b_1])=0$$
$$M[b_1,b_2]=f^{-1}([b_1,b_2])=\{x \in T^2 \mid f(x) \in [b_1,b_2]\}$$
$$H_0(M[b_1,b_2])=\mathbb{Z}_2,\ H_1(M[b_1,b_2])=\mathbb{Z}_2 \oplus \mathbb{Z}_2$$
$$M[b_1,b_3]=f^{-1}([b_1,b_3])=\{x \in T^2 \mid f(x) \in [b_1,b_3]\}$$
$$H_0(M[b_1,b_3])=\mathbb{Z}_2,\ H_1(M[b_1,b_3])=\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$$

包含写像 $\iota:M[b_1,b_2] \to M[b_1,b_3]$ を通じて 2 つの 1 次元ホモロジー群の関係を見ると、$H_1(M[b_1,b_2])$ の 2 つの生成元は $H_1(M[b_1,b_3])$ の生成元の一部として引き継がれ、新たにもう 1 つの $H_1(M[b_1,b_3])$ の生成元が生まれている。cover-level set のホモロジー群を比較することにより、cover-level set 間の位相的特徴を抽出することができる。この考えを与えられた点群に対して用いて、各 cover-level set のパーシステントホモロジー群の生成元の繋がり具合等を位相的特徴として捉える。

5. 提案手法

本研究では、mapper 解析のクラスター分析が 0 次元パーシステントホモロジー群に相等する。0 次元パーシステントホモロジー群を 1 次元パーシステントホモロジー群に置き換えたりして実験を行う。そのための提案手法の手順を示す。また、提案手法の簡単な流れを図 6 に示しておく。

(手順 0) 分割する区間とカバー比率を決定する

(手順 1) 離散データをフィルタ関数に基づいて有限個の cover-level set に分割する

(手順 2) 分割された各 cover-level set に対して k 次パーシステントホモロジー群を計算する

(手順 3) 手順 2 で得られた k 次パーシステント図をベクトル化する

(手順 4) 手順 3 のベクトルに対して異常検知技術を用いて、データの特異的なベクトル値を検出する

(手順 5) 特異的なベクトル値に対応する cover-level set を分析する

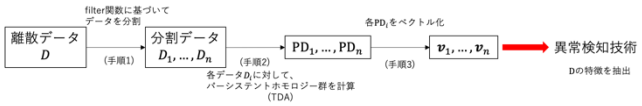

図 6：提案手法の流れ

手順 1 のデータの分割方法と手順 3 のベクトル化について詳しく説明する。

手順 1：離散データをフィルタ関数に基づいて分割し、その分割されたデータを時系列データとして考える。分割方法については cover-level set を用いる。

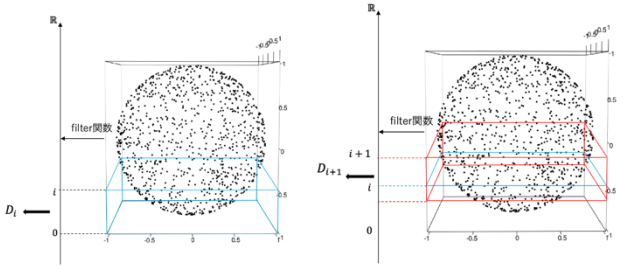

図 7：一般的な cover-level set の作成方法

図 3、4 で説明したように、フィルタ関数で考える 1 つの軸を有限個の区間に分割していく。この区間に対応する元データを集めてきたものをデータ D_i (図 7 左の図の青枠)、右の図の赤枠で切り取られたデータをデータ D_{i+1} とする。その際、離散モース理論に基づいてデータ D_i とデータ D_{i+1} に重なりをもたせる。こうして得られた分割データを時系列データとして考える。

手順 3：手順 2 で得られた k 次パーシステント図のベクトル化を行う。パーシステントホモロジー群を計算して得られるパースステント図は多重集合であるため、そのままの解析は

難しい。そこでパーシステント図をベクトル化することで解析を可能にする。

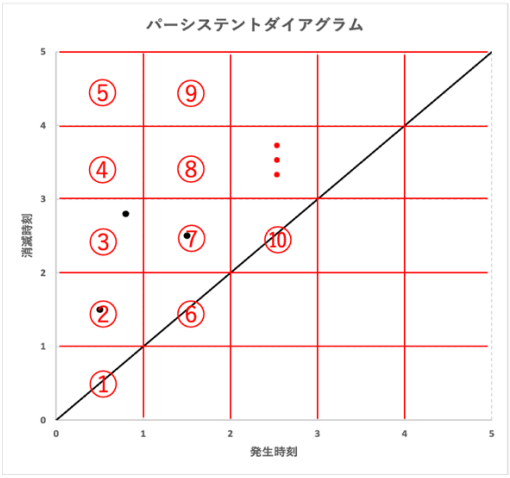

図 8：パーシステント図のベクトル化

図 8 のように、birth － death 領域をメッシュ化し各メッシュに番号をつける。パーシステントホモロジー群の次元ごとに、その各メッシュに含まれる点を数えることでベクトルを生成する。今回は 0 次元パーシステントホモロジー群と 1 次元パーシステントホモロジー群で、それぞれベクトル化する。

6. 人工的に作成したデータに対する検証実験

検証に使用する元データについて考える。元データとして位相的特徴がわかりやすい(モース理論による結果が予想しやすい)人工的に作成したデータを用いる。具体的には図 9 にように球面、トーラス、ダブルトーラス(種数 2 閉曲面)上に点を均等に分布したものを用いて検証を行う。トーラスについては、分布のさせ方が meridian 方向と longitude 方向で 2 通りあるため、両方用意している。また、ダブルトーラスは上記のトーラスを 2 つ連結したものとなっている。表 10 にデータの個数をまとめておく。

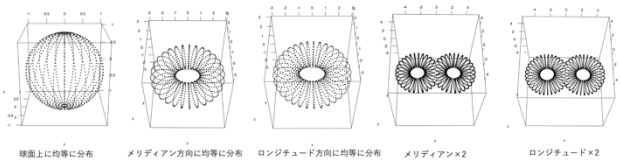

図 9：人工的に作成したデータ

表 10：人工的に作成したデータの種類と点の個数

	球面	meridian トーラス	longitude トーラス	meridian ダブルトーラス	longitude ダブルトーラス
点の数	1000	1000	1000	1926	1934

図 6 に示した提案手法の流れように、人工的に作成したデータを cover-level set で分割する(手順 1)。分割された各データに対してパーシステントホモロジー群を計算し(手順 2)、ベクトル化まで行う(手順 3)。最後は主成分分析を利用した異常検知技術を用いて解析し、従来の mapper 解析と比較しながら元データの特徴を抽出できたか検証する。主成分分析の結果から寄与率の高い主成分を取り出す。取り出した主成分が 1 つの場合は正規分布に基づく異常検知を行い、取り出した主成分が複数の場合はクラスター分析を用いる。正規分布に基づく異常検知では、異常度の定義と閾値の設定が必要であるため、簡単に述べる。

異常度として正規分布の負の対数尤度を採用する。確率変数を x としたとき、平均 μ、分散 σ^2 をもつ正規分布 $\mathcal{N}(x \mid \mu, \sigma^2)$ は、

$$\mathcal{N}(x \mid \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

で定義される。元データから最尤推定を用いて推定値 $\hat{\mu}, \hat{\sigma}^2$ を求めれば、確率モデル $\mathcal{N}(x \mid \hat{\mu}, \hat{\sigma}^2)$ が得られる。この式の対数尤度を計算すると、ある観測値 x' の異常度 $a(x')$ は、

$$a(x') = \left(\frac{(x'-\hat{\mu})}{\hat{\sigma}}\right)^2$$

で与えられる[6, p. 19]。

ホテリング理論から、この異常度 $a(x')$ は自由度 1、スケール因子 1 のカイ二乗分布に従う。このことから閾値は、カイ二乗分布から設定する。本研究では、閾値を 10%とする。

7. 実験結果と考察

データ 1：球面（filter 関数＝z 座標値）

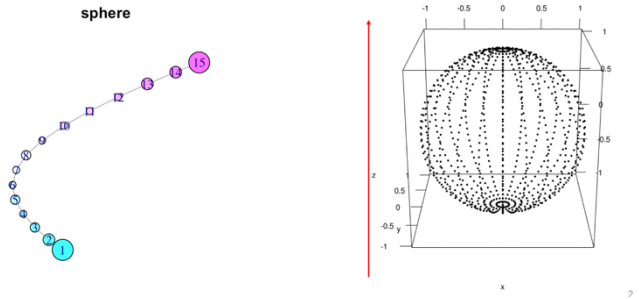

図 11：mapper 解析(球面 filter 関数＝z 座標値)

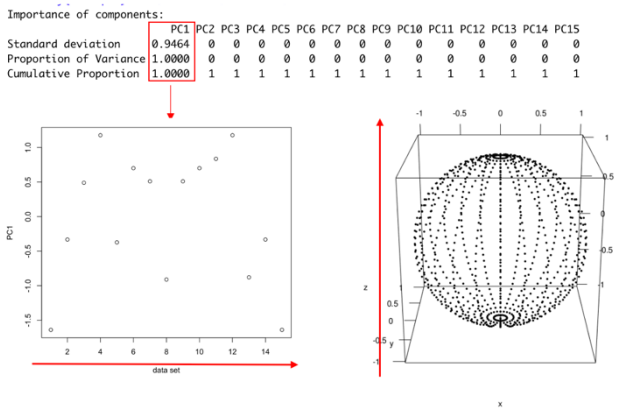

図 12：提案手法(球面 filter 関数＝z 座標値、$\mathbf{PH_0}$)

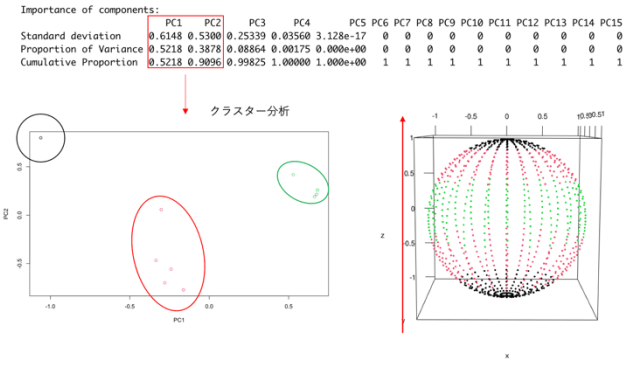

図 13：提案手法(球面 filter 関数＝z 座標値、$\mathbf{PH_1}$)

図 11~13 は球面(filter 関数＝z 座標値)に対する mapper 解析の結果と 0 次元パーシステントホモロジー群(PH_0)、1 次元パーシステントホモロジー群(PH_1)を用いた提案手法の結果となっている。

図 11 は従来の手法である mapper 解析の結果で、左の図は mapper グラフを、右の図は mapper グラフの分岐点に対応する元データを示している。今回は分岐点が見られないため、対応する元データに色分けをしていない。赤矢印はフィルタ関数の軸を示している。

図 12 はPH_0を用いた提案手法の結果を示している。提案手法の流れに沿って、球面上の点を cover-level set で分割し、分割したデータに対してパーシステントホモロジー群を計算する。計算結果からPH_0を用いてベクトル化し、異常検知技術として主成分分析を用いて特徴抽出を行なっている。図 12 の上部はその主成分分析の結果となっている。主成分分析の結果から寄与率の高い第一主成分を取り出し(図 12 左)、正規分布に基づく異常検知から異常度が閾値を超えたデータに対応する元データを色分けしている(図 12 右)。図 12 では閾値を超えた元データは存在していない。

図 13 はPH_1を用いた提案手法の結果を示している。提案手法のベクトル化の際にPH_1を用いてベクトル化し、そのベクトルに対して主成分分析を行なっている。図 13 上部の主成分分析の結果から、寄与率が高い第一主成分と第二主成分を取り出し(図 13 左)、クラスター分析を用いた分類を行なっている。図 13 右は、各クラスターに対応する元データを色分けしたものとなっている。1 次元 PH を用いているので、赤と緑のクラスターで、輪の大きさの違いが出てきたのではないかと考えられる。

データ 2：meridian トーラス（filter 関数＝z 座標値）

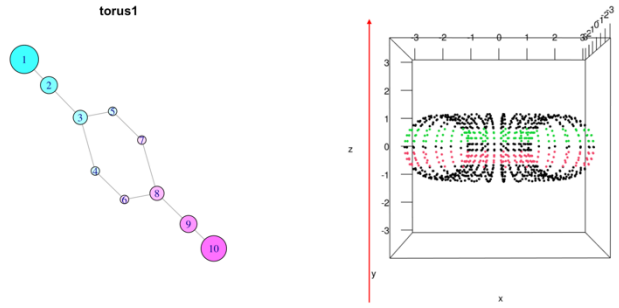

図 14：mapper 解析(meridian トーラス filter 関数＝z 座標値)

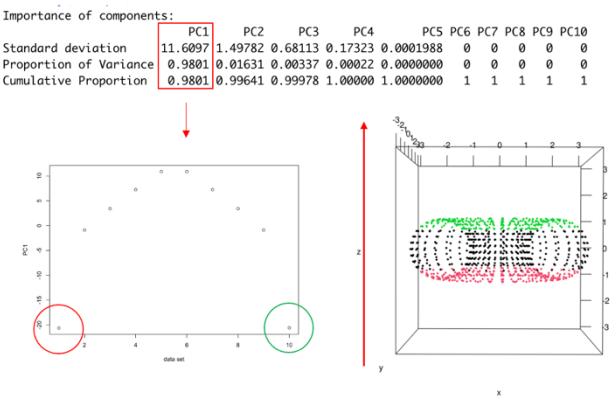

図 15：提案手法(meridian トーラス filter 関数＝z 座標値、$\mathbf{PH_0}$)

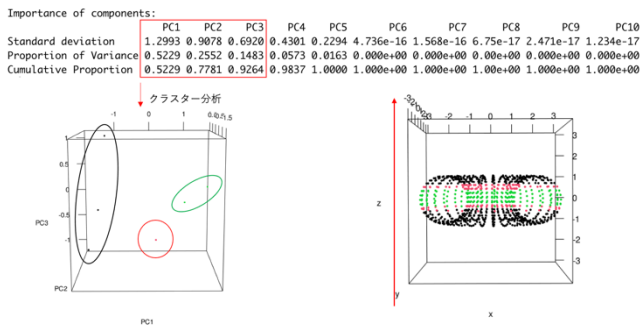

図 16：提案手法(meridian トーラス filter 関数＝z 座標値、$\mathbf{PH_1}$)

図 14~16 は meridian トーラス(filter 関数＝z 座標値)に対する mapper 解析の結果と提案手法(PH_0、PH_1)の結果となっている。mapper 解析では分岐点が 2 つ確認でき(図 14 左)、提案手法(PH_0)では異常値として 2 つ検出された(図 15 左)。mapper グラフの分岐点に対応する元データは何か、mapper グラフの生成手順を遡って解析することを逆解析という。それぞれ逆解析した結果を比較すると(図 14 右と図 15 右)、少し違った結果となっていることがわかる。0 次元パーシステントホモロジー群として考えると、z=0 付近のデータセットは内側と外側で 2 つの連結成分が、それ以外のデータセットは 1 つの連結成分が生成されることが予想できるため、提案手法の結果は不自然ではないと考えられる。提案手法(PH_1)の逆解析の結果(図 16 右)は、1 次元パーシステントホモロジー群として考えると、z=0 付近のデータセットでは 2 つの輪

が、それ以外のデータセットでは 1 つの輪が生成されることが予想できる。また、赤と緑のデータセットで輪の大きさ(発生時刻と消滅時刻)の違いが表されたと考えられる。

データ 3：longitude トーラス（filter 関数＝x 座標値）

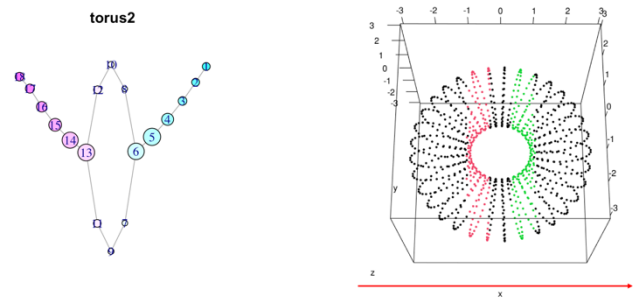

図 17：mapper 解析(longitude トーラス
filter 関数＝x 座標値)

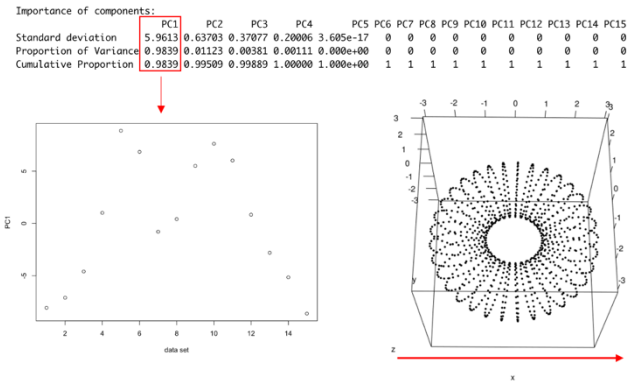

図 18：提案手法(longitude トーラス
filter 関数＝x 座標値、$\mathbf{PH_0}$)

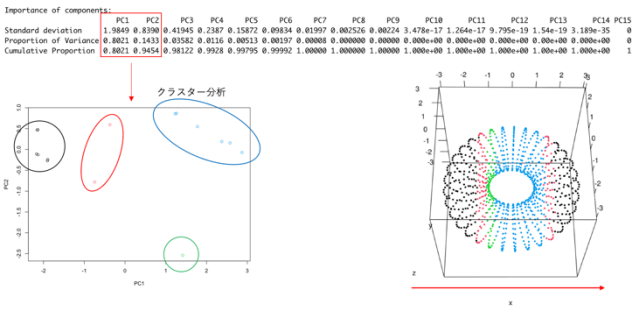

図 19：提案手法(longitude トーラス
filter 関数＝x 座標値、$\mathbf{PH_1}$)

図 17~19 は longitude トーラス(filter 関数＝x 座標値)に対する mapper 解析の結果と提案手法(PH_0、PH_1)の結果となっている。mapper 解析と提案手法(PH_0)の逆解析の結果を比較すると(図 17 右と図 18 右)、提案手法(PH_0)では位相的特徴が失われている。提案手法(PH_1)の逆解析の結果(図 19 右)は、1 次元パーシステントホモロジー群として考えると、filter 関数に対して生成元として特徴的に見られる輪の個数が 0→1→2→1→0 と変化することがわかるため、大まかに特徴を捉えているといえる。

データ 4：meridian ダブルトーラス （filter 関数＝x 座標値）

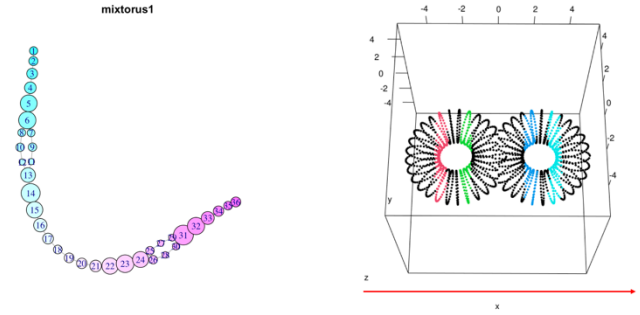

図 20：mapper 解析(meridian ダブルトーラス
filter 関数＝x 座標値)

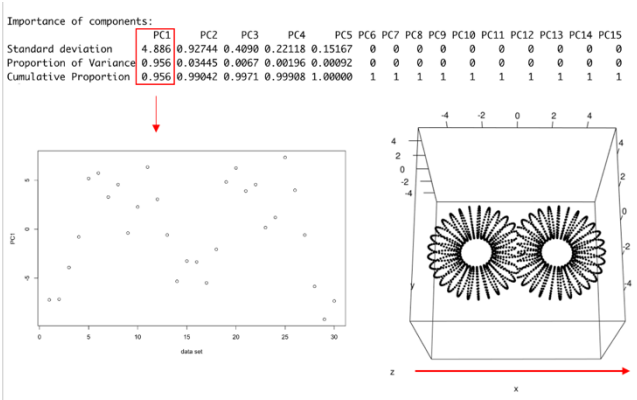

図 21：提案手法(meridian ダブルトーラス
filter 関数＝x 座標値、$\mathbf{PH_0}$)

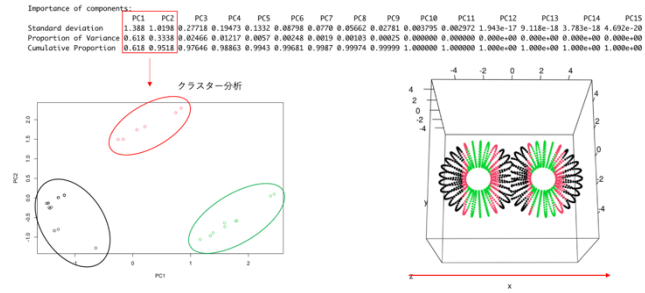

図 22：提案手法(meridian ダブルトーラス
filter 関数＝x 座標値、$\mathbf{PH_1}$)

図 20~22 は meridian ダブルトーラス(filter 関数＝x 座標値)に対する mapper 解析の結果と提案手法(PH_0、PH_1)の結果となっている。mapper 解析と提案手法(PH_0)の逆解析の結果を比較すると(図 20 右と図 21 右)、提案手法(PH_0)では位相的特徴が失われている。提案手法(PH_1)の逆解析の結果(図 22 右)は赤と緑のクラスターで図形の特徴を捉えていることがわかる。

データ 5：longitude ダブルトーラス （filter 関数＝y 座標値）

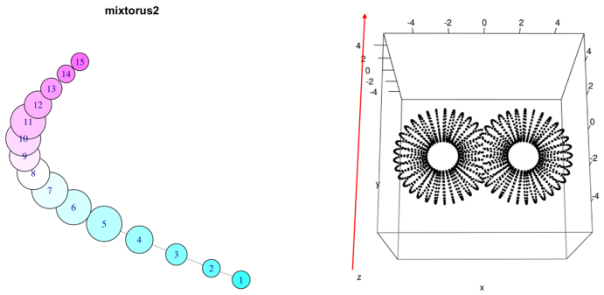

図 23：mapper 解析(longitude ダブルトーラス filter 関数＝y 座標値)

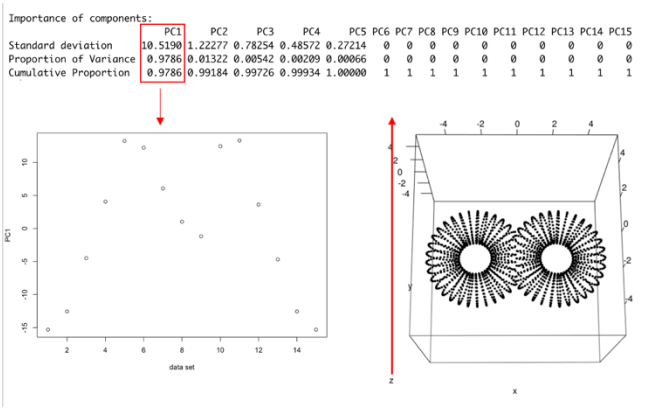

図 24：提案手法(longitude ダブルトーラス filter 関数＝y 座標値、$\mathbf{PH_0}$)

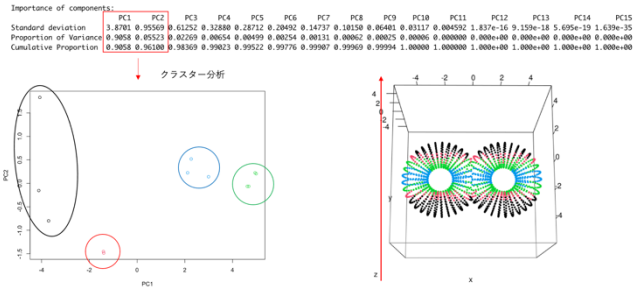

図 25：提案手法(longitude ダブルトーラス filter 関数＝y 座標値、$\mathbf{PH_1}$)

図 23~25 は longitude ダブルトーラス(filter 関数＝y 座標値)に対する mapper 解析の結果と提案手法($\mathbf{PH_0}$、$\mathbf{PH_1}$)の結果となっている。mapper 解析と提案手法($\mathbf{PH_0}$)の逆解析の結果を比較すると(図 23 右と図 24 右)、同じような結果となっているが、どちらも位相的特徴が失われている。0 次元パーシステントホモロジー群として考えると、filter 関数に対して連結成分の個数が 2→1→3→1→2 と変化することがわかるため、位相的特徴が失われているのは明らかである。提案手法($\mathbf{PH_1}$)の逆解析の結果(図 25 右)は図形の特徴を捉えている。

また、実データとしてバレーボール V リーグ選手データに対しても提案手法を適用し、Sabermetrics 的な選手の能力評価を行なってみたが、満足のいくような結果は得られなかった。

8. 結論

離散的モース理論を基にした提案手法($\mathbf{PH_1}$)は、人工的データに対してクラスターごとに位相的特徴を捉えることができた。この手法に mapper 解析では得られない位相的特徴を抽出することが期待される。提案手法($\mathbf{PH_0}$)は、mapper 解析と同じような結果が得られたり、位相的特徴が失われたり様々であった。原因として、仮定した分布モデルが適切でなかったことが挙げられる。正規分布ではなく混合正規分布をモデルとして用いたり、他の異常検知技術を用いれば、今回とは違った結果が得られる可能性がある。また、提案手法では cover-level set の作成、ベクトル化、異常検知手法にハイパーパラメータが存在するため、その調整も課題となる。

9. 参考文献

[1] VALUENEX, 数学の時代到来の予感、パーシステント・ホモロジー -VALUENEX 技術トレンドレポート-, 2021

[2] Gurjeet Singh, Facundo Memoli and Gunnar Carlsson, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Eurographics Symposium on Point-Based Graphics, 2007

[3] Emerson G. Escolar, Yasuaki Hiraoka, Mitsuru Igami, Yasin Ozcan, Mapping Firms' Locations in Technological Space: A Topological Analysis of Patent Statistics, 2021

[4] AYASDI, Machine Learning with Ayasdi, 2014

https://www.slideshare.net/Ayasdi/machine-learning-with-ayasdi

[5] 平岡裕章, 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

[6] 井出剛, 入門 機械学習による異常検知 - R による実践ガイド-, コロナ社, 2015