様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１６　日現在

機関番号：　６２６０３
研究種目：　挑戦的萌芽研究
研究期間：　2014～2016
課題番号：　２６５４００１６

研究課題名（和文）代数的位相幾何の方法による統計的データ解析の新たな展開

研究課題名（英文）New developments of statistical data analysis with algebraic topology

研究代表者
福水　健次（Fukumizu, Kenji）

統計数理研究所・数理・推論研究系・教授

研究者番号：６０３１１３６２

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：（１）パーシステント図をベクトル化するためのカーネル法を提案した．それをシリカの液相ガラス相転移温度の推定に応用し有効性が確認された．（２）多様体学習による低次元射影とクラスタリングを用いた，遺伝子クラスタリングに基づく進化系統樹解析の枠組みを提案した．（３）オイラー標数法を応用して，回帰問題の信頼区間，および行列の最大固有値の分布に関して新しい近似法を得た．

研究成果の概要（英文）：(1) Kernel methods for vectorizing persitence diagrams has been proposed, and they have been applied to detecting the phase transition temperature between glass and liquid. (2) A framework for phyogeny analysis has been proposed based on manifold learning and gene clustering. (3) The method of Euler number has been applied to derive reliability intervals of regression problems as well as the distribution of maximum eigenvalue of random matrices.

研究分野：　統計的機械学習

キーワード：　多変量解析　代数的位相幾何　多様体学習

１．研究開始当初の背景

高次元データの解析では、データの持つ低次元多様体構造の獲得が有効となることが多い。それにはデータが分布する多様体の連結性やオイラー数など位相幾何的情報の抽出が重要であるが、位相構造を議論する位相幾何学の理論は、ランダム性やノイズを持つ対象を想定しておらず、従来は統計的データに内在する位相幾何的特徴を安定的に捉えることが困難であった。

近年、パーシステントホモロジーという代数的位相幾何学の方法が発展し、幾何的対象の代数的表現（チェイン複体と呼ばれる）の１次元系列に対し、その系列内での位相構造の持続性が議論可能となった。これによりノイズに頑健な位相情報の抽出が期待され、国内でも分担者・平岡や連携者・松江がタンパク質構造や材料科学のデータ解析に応用している。一方、正しい統計的解析には、データのランダム性を反映したパーシステントホモロジーの統計的性質の解析が重要であるが、そのような研究は海外においても未開拓である。また、多次元ガウス変数の最大値を求める幾何的方法であるチューブ法に対し、その近似計算にパーシステントホモロジーが有効であることが、チューブ法の専門家である分担者・栗木らの予備的考察により予想されていた。

２．研究の目的

本研究は、近年発展している代数的位相幾何学の方法を統計的なデータ解析に応用することにより、多変量解析に新たな方法論を創成することを目的とする。より具体的には、位相的データ解析（文献[1]）の分野で発展しているパーシステントホモロジー（文献[2]）を統計的データに適用した際の、ホモロジー生成元の統計的性質を解明する。これによりパーシステントホモロジー自体をデータ表現とみなした解析が可能となる。さらに多変量解析への応用として、高次元データの低次元構造抽出法を高度化、多重比較などに有用なアブストラクトチューブ法に対する新しい近似計算手法の開発を行う。これらを通して、統計的データ解析に代数的位相幾何統計学というべき新しい方法論を確立する。

本研究では、特に以下の３つの点を明らかにしようとした。

【課題 1】統計的データから定まるパーシステントホモロジーの統計的性質の解明とその応用：

統計的データから定まるホモロジー群生成元の持続性を確率的に評価する。それを用いて、ホモロジー生成元の持続性自体をデータ表現とみなしたデータ解析法の構築を行う。

【課題 2】多様体学習への応用：

【課題 1】の結果に基づき、多様体学習において重要となる、データの近傍を定義する閾値や次元の決定方法を開発する。それを画像データ解析、系統樹データ解析に応用する。

【課題 3】チューブ法を用いた多重比較問題への応用：

パーシステントホモロジーの方法を、多次元ガウス分布の最大値を求めるためのアブストラクトチューブ法に応用し、多重比較問題で有用な数値計算法を開発する。

３．研究の方法

【課題 1】パーシステントホモロジーの統計的性質は、代表者の福水と、パーシステントホモロジーとその応用に関して実績の高い分担者・平岡が担当する。【課題 2】多様体学習への応用は、次元削減法などに実績のある代表者・福水と、代数統計およびその遺伝情報処理への応用に実績の高い Yoshida の協力のもとに研究を推進する。【課題 3】チューブ法による多重比較への応用は、チューブ法の第一人者である分担者・栗木および連携者・松江により研究を進める。

以上のように、「２．研究の目的」で述べたように統計科学と代数的位相幾何の両方に知識のある代表者・福水が主導し、幾何的なアプローチによる統計科学で高い実績を持つ参加者と、パーシステントホモロジー研究を主導する数学者との協働により研究を推進する体制を取る。

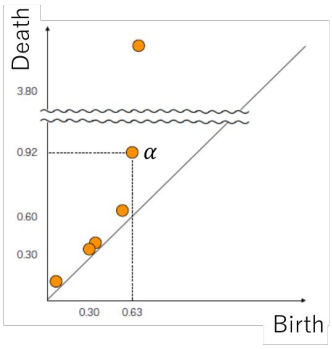

図１：パーシステント図の例

４．研究成果

（１）パーシステント図のベクトル化による統計的解析法の確立：

パーシステントホモロジーの計算結果は、ホモロジー生成元の生成・消滅時刻を表した、2 次元のパーシステント図（図１）で表現することが多く、実際のデータ解析においてもよく用いられている（文献[3,4]）。パーシステント図は、元の点集合の幾何的・位相的情報をコンパクトに表しているものの、それは 2 次元の点集合として表されており、通常のベクトルデータのように多くの統計的データ解析手法が適用可能な表現となっていない。

この問題に対して、カーネル法によるベクトル化による機械学習的な解決法を示し，特にパーシステント図に対して適した「パーシステント重み付きガウスカーネル」を提案した．このカーネルによってパーシステント図がベクトル化され，多くのデータ解析手法が包括的に適用可能となる．また，対角線付近のノイズ由来と思われる生成元の効果の割引率をコントロールすることが可能である．

基礎的検討として，人工データによる２値識別問題に対して，ベクトル化した後にサポートベクタマシンを用いることによって識別器を構成して識別性能を評価したところ，既存のベクトル化手法であるパーシステントランドスケープやパーシステントイメージなどに比べて，パーシステンス重み付ガウスカーネルによるベクトル化が優位性を持っていることが示された（表１）．

Kernel for embedding		2nd-level kernel	
Kernel	Weight	Linear	Gauss
	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
	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

表１：識別率の比較．赤字部分が提案手法

実問題への応用として，二酸化ケイ素（シリカ）の液相-ガラス相転移温度を，統計的な変化点検出の手法によって推定する問題に応用した．８０点の異なる温度設定によるMD シミュレーション（文献[3,4]）によって得られたシリカの原子配置データ（図２）から，球近傍によってフィルトレーションを構成し，そのパーシステント図（図３）からパーシステント重み付きカーネルを用いて，カーネル法による変化点検出法（文献[5]）によって，温度を変えたときに前後でパーシステント図が最も大きく変化する温度を求めた（図４）．その結果，物理学分野でエンタルピー曲線の変曲点推定によって求められている区間推定の範囲に入る転移温度が推定された．このことは，物理的知識を使わずに，原子配置に対するデータ解析だけから妥当な相転移温度が推定されたことを意味しており，興味深い結果である．さらにカーネル主成分分析を用いてパーシステント図のベクトル化データを３次元空間に射影してプロットしたところ，推定された転移温度の前後で急峻な変化が生じていることが確認された（図４）．物理的に結論がでていない時間を固定したスナップショットから転移温度が決定できるかどうかは，物理学でも未解決の問題であるそうだが，今回の結果は，限定的な数値実験であるが，肯定的な結論を示唆している．

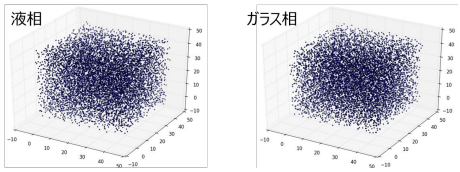

図２：シリカの原子配置（文献[3,4]）

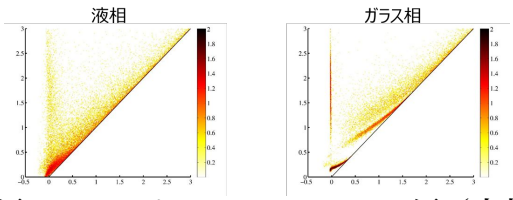

図３：シリカのパーシステント図（文献[3,4]）

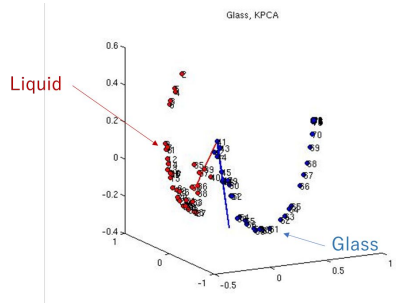

図４：カーネル PCA による３次元プロット．赤点が液相，青点がガラス層に対応．

以上の成果をまとめた論文（発表論文 1）が，機械学習分野のトップ国際会議である International Conference on Machine Learning に採択された．また，内容を拡充した論文（arXiv 版：文献[6]）を論文誌に投稿した．

（２）多様体学習による系統樹クラスタリング

遺伝子配列データから進化系統樹を推定する問題では，従来，多くの遺伝子を結合したひとつの系列データを作成し，生物種の間でその類似性をみることにより，ひとつの系統樹を推定する方法が標準的であった．しかしながら，遺伝子水平伝播などの現象が知られるようになり，単純な単一木構造では十分表現できない問題を知られるようになってきた．

本研究では，遺伝子ごとに異なる系統樹を構成し，それらを低次元表現する多様体学習を行ったあとにその低次元表現をクラスタリングし，遺伝子クラスタごとに系統樹を構成する方法を提案した．系統樹間の距離は，木の間の距離として標準的な Billera-Holmes-Vogtmann (BHV)距離を用い，

次元削減としては t-SNE，カーネル主成分分析が有効であった．また，クラスタリング法としては normalized cut が有効であることが，人工的なデータによる基礎実験によって確認された．この方法を，魚類と四足動物の橋渡しとなった種が何かという未解決問題に関連した解析として，シーラカンス，肺魚，四足動物を含む１０種の生物の遺伝子データにたいし，クラスタリング＋系統樹構成の方法を適用した．この結果，遺伝子に２つのクラスタがみられ，それぞれの遺伝子クラスタは多少異なる系統樹を示唆することがわかった．

以上の研究結果をまとめた論文が Annals of Operations Research から出版されている．

（３）確率場の最大値分布

確率場の最大値の分布をサブレベル集合（エクスカージョン集合）のベッチ数の交代和（オイラー標数）の期待値で近似する方法は，統計学ではオイラー標数法として知られている．本研究では，確率過程の最大値の分布に関して従来の結果を拡張し，オイラー標数法を用いて，回帰曲線の信頼性評価を行う方法を提案した．この結果は Journal of Multivariate Analysis によって出版された（発表論文 3）．またオイラー標数法を用いて，行列サイズが有限，無限の両方の場合について，実・複素ランダム行列（ガウス，ウィシャート）の最大固有値の分布の近似を導いた．また，その近似誤差が指数的微小量となることを確認した．

＜引用文献＞

[1] Carlsson, G. (2009) Topology and data. Bull. Amer. Math. Soc., 46(2):255-308.

[2] Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002) Topological persistence and simplification. Discrete and Computational Geometry, 28(4):511-533, 2002.

[3] 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).

[4] Hiraoka, Y., T. Nakamura, A. Hirata, E.G. Escolar, K. Matsue, and Y. Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. Proc. Nat. Acad. Sci. Vol. 113 no. 26, 7035-7040.

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

[6] Kusano, G., Hiraoka, Y. and Fukumizu, K. (2016) Persistence weighted Gaussian kernel for topological data analysis. arXiv:1601.01741 [math.AT]

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

1. Kusano, G., Hiraoka, Y. and Fukumizu, K. (2016) Persistence weighted Gaussian kernel for topological data analysis, Proceedings of The 33rd International Conference on Machine Learning (ICML2016), pp. 2004-2013.2016 査読あり
2. Yoshida, R., Fukumizu, K., Vogiatzis, C. Multilocus phylogenetic analysis with gene tree clustering. Ann Oper Res (March 2017). 査読あり
3. Lu, X. and Kuriki, S. (2017) Simultaneous confidence bands for contrasts between several nonlinear regression curves, Journal of Multivariate Analysis, 155, 83-104 査読あり

〔学会発表〕（計８件）

① Kusano, G., Hiraoka, Y. and Fukumizu, K. (2016) Persistence weighted Gaussian kernel for topological data analysis, The 33rd International Conference on Machine Learning (ICML2016)

② 福水健次. Persistence weighted Gaussian kernel for topological data analysis. ERATO 感謝祭（情報学研究所）2016/8/10

③ 福水健次. 位相的データ解析への機械学習的アプローチ．京都大学情報学研究科・統計数理研究所・公開シンポジウム「データサイエンス　－ 情報と統計が創造する未来 －」（京都大学）2017/3/17

④ 福水健次. Kernel Methods for Topdogical Data Anaiysis. 2016 International Workshop on Spatial and Temporal Modeling from Statistical, Machine Learning and Engineering perspectives （統計数理研究所）2016/7/22

⑤ 栗木哲.「チューブ法の理論・応用とその周辺」（統計学会賞受賞講演）2016 年度統計関連学会連合大会（金沢大）2016.9.6

⑥ 栗木哲.「期待オイラー標数法によるランダム行列の最大固有値分布の近似」行列解析の展開（名古屋大）2017.3.2

⑦ Kuriki, S. “ The volume-of-tube method

for simultaneous inference in nonlinear regression models”（招待講演）Mathematische Kolloquium（Ulm University）2016.5.2

⑧ Kuriki, S.“Some distributions associated with the cone of positive semidefinite matrices and their applications” IMS-APRM 2016 Hong Kong. 2016.6.28

〔図書〕（計　0件）

〔その他〕
ホームページ：
http://www.ism.ac.jp/~fukumzu/

６．研究組織
(1)研究代表者
福水健次（FUKUMIZU, Kenji）
統計数理研究所・数理・推論研究系・教授
研究者番号：60311362

(2)研究分担者
平岡裕章　（HIRAOKA, Yasuaki）
東北大学・材料科学高等研究所・教授
研究者番号： 10432709

(3)研究分担者
栗木哲　（KURIKI, Satoshi）
統計数理研究所・数理・推論研究系・教授
研究者番号： 90195545

(4)連携研究者
松江要（MATSUE, Kaname）
九州大学・マス・フォア・インダストリ研究所・助教
研究者番号：　70610046