研　究　報　告　書

「データ空間の幾何学的特徴を活用する解析手法と統計理論」

研究タイプ：通常型
研究期間：　平成26年10月～平成30年3月
研　究　者：　小林　景

1．　研究のねらい

統計学や機械学習などにおいて解析対象となるデータは，数ベクトルの形をしていることがほとんどである．よって自然にユークリッド空間上の点集合と同一視できる．しかし，実際にはデータはユークリッド空間に局在する部分空間（データ空間）にのみ存在する場合が多く，そのことを仮定した上で，データ空間の情報を用いた統計解析を行うことにより，解析精度を大きく向上させることができる．特に，高次元データの解析では，データ空間自体は低次元な部分空間や部分多様体となる場合も多く，計算量と汎化誤差のいずれの観点からも，データ空間の情報は重要な役割を果たす．また，ネットワーク構造，グラフ構造などをもつデータの解析では，自然に導出されるデータ空間（ネットワーク，グラフなどの集合）が，良い性質をもつ多様体や多面体的複体になることも多く，その特徴を利用した統計的解析が可能である．

本研究では，このようなデータ空間の幾何学的特徴づけ，及びそれを用いた統計的解析の新手法の開発，そしてその理論的解析を目的とする．さらには，その過程において数学的にも普遍的に意味があるような結果を残すことも目標とする．より具体的には，以下の様な目的がある．

（A）データ空間が非正の曲率を持つとき，標本および母集団分布のいずれの Fréchet 平均も，一意に存在することが知られている．このように，データ空間の曲率とその上の統計的解析については深い関係がある．本研究では，それをさらに進め，高次元空間に埋め込まれたより複雑なデータ空間の曲率と，それを用いたデータ解析手法の評価，提案を行う．

（B）実際にデータ解析を行う際には，データ空間は未知である場合が通常であり，何らかの方法で近似する必要がある．その中で一般的な方法の一つが，経験グラフを用いてデータ空間を近似する方法である．本研究では，確率的な手法などでこの理論評価を行うことにより，経験グラフの構成法に関する指針を見つける．

（C）遺伝系統樹分析などに応用される tree space は，CAT(0)という数学的性質を持つことが証明され，既存の統計手法の応用や，アルゴリズムの開発などのブレークスルーとなった．本研究では，より一般的な幾何学構造をもつデータに関しても，同様にデータ空間を構成できるのかを探る．

（D）データ空間に幾何学的制約がある場合の情報幾何学を構成し，推定，予測理論などに応用する．

2. 研究成果

（1）概要

本研究ではデータがその周辺に分布するような「データ空間」に着目する．まず，測地距離を連続的に変換することにより CAT（k）性とよばれる曲率を積極的にコントロールし，判別分析やクラスタリングなどの統計解析の精度を向上する方法を提案した．また，新しく提案した距離を用いて定義される一般化分散を用いて，各年のイギリス降雨量データの分散の時系列の変化を計算することにより，通常のユークリッド距離を用いた分散ではみられない近年の分散の増大を確認できた．さらに，気候データのもつ年周期性から生じる幾何学的構造をまず視覚化し，その時間的，空間的な変化をとらえるための R 言語のソフトウェアを開発した．本ソフトウェアをはじめ，開発されたプログラムおよびアメダス全国データ等を用いた解析結果はホームページにまとめられ，一般に公開された．

上記の新手法をはじめ，多様体学習等の手法においては，データ空間の距離近似として経験グラフを用いてデータ空間を近似する．その際にデータ空間上での最短距離を経験グラフによる最短経路長で近似した場合に，どの程度の近似精度がえられるかは重要な課題であった．本研究では，高次元性によりこの近似精度が落ちるひとつの原因である「ハブ現象」について，局所的に中心化された距離を用いるとハブが削減されることを実験的に示し，またその理由を解析した．一方，測地距離の経験グラフ最短経路近似を用いた場合の近似評価については，2 次元ドロネーグラフを用いた場合の Fréchet 平均の一致性を証明し，より一般的な場合の証明の足掛かりができた．

一方，木グラフのデータ空間は CAT(0)という幾何学的特徴をもつため，その上での最短経路長計算や，平均，分散の定義ができるという既存の事実を拡張して，新しい統計解析手法を提案することを目指し，デンドログラムという限られた幾何学的構造をもつデータの場合には部分的に成功した．さらに，その得られた結果を用いて，外国語学習者の心内辞書（mental lexicon）の相違を明らかにするという実験データ解析のための，並べ替え検定を用いた新しい検定統計量を提案した．また，検定統計量の漸近的な一致性や有効性を証明し，特に群平均法とよばれるデンドログラム構成法が，射影性や局所線形性など好ましい性質を持つことを示した．さらに，折田充（熊本大）らの研究グループと共同で，単語のオンライン学習プログラムを開発した．

（2）詳細

研究テーマ（A）「データ空間の曲率を利用した統計的解析」

本さきがけ研究までのデータ空間の曲率解析は，単純な図形や多様体を仮定したものであった．また，曲率は固定されているため，精度保証という受動的な形で曲率の情報は利用されてきた．本研究ではデータ空間の測地距離を連続的に変換することにより CAT（k）性とよばれる曲率を積極的にコントロールし，判別分析やクラスタリングなどの統計解析の精度を向上する方法を提案した．具体的には，α 距離変換と呼ばれる測地距離への変換ののち，β 距離変換とよばれる距離の変換を行い，必ずしも測地距離ではない一般の距離へと変換する．特に β 距離変換は，計量錐とよばれる測地距離空間に埋め込んだうえで，その計量錐の曲率を変換していることに対応することを証明した（論文１及び arXiv1401.3020v5）．また，新しく

提案した距離を用いて定義される一般化分散を用いて，各年のイギリス降雨量データの分散の時系列の変化を計算することにより，通常のユークリッド距離を用いた分散ではみられない近年の分散の増大を確認できた．本成果については Royal Statistical Society Annual Conference 2016 で発表された．

さらに，上記の研究の過程で，気候データのもつ年周期性が幾何学的構造をもち，それが幾何学的な特徴量を構成する動機となることに気付いたため，その幾何学的構造をまず視覚化し，その時間的，空間的な変化をとらえるためのソフトウェアを開発した．また，開発した R 言語のソフトウェアおよび，実際にアメダス全国データを用いた解析結果の GUI を備えたホームページを作成し，一般に公開した．

研究テーマ(B)「経験グラフの構成手法とその理論評価」

本研究テーマ(A)による α 距離を計算する際の最小経路の計算においては，もとのデータ空間が未知であり，何らかの方法で近似する必要がある．その中で一般的な方法の一つが，経験グラフを用いてデータ空間を近似する方法であるが，その際にデータ空間上での最短距離を経験グラフによる最短経路長で近似した場合に，どの程度の近似精度がえられるかは重要な課題であった．本研究では，高次元性によるハブ現象解析と，ドロネー三角化による経験グラフの最短経路長解析という二つのアプローチからその課題に挑戦し，部分的な解決を得た．

まず，データ空間のグラフ近似精度評価で重要なハブ現象の解析に関しては，局所的に中心化された距離を用いるとハブが削減されることを国際学会にて共著発表し（Similarity Search and Applications 2015），また，この方法によりなぜハブが削減されるのかという理由を，実験結果とシンプルな理論で説明し，38th Annual ACM SIGIR Conference および AAAI Conference on Artificial Intelligence 2016 にて共著発表し，その査読付き Proceeding が出版された（論文2，4，5）．

一方，測地距離の経験グラフ最短経路近似を用いた場合の近似評価については，二次元ドロネーグラフを用いた場合の内測平均（Fréchet mean）の一致性について，ポアソン過程についてのドロネーグラフ上の最短経路に関する確率理論および二次元ドロネーグラフが平面スパナーであるという事実を用いて証明し，研究集会「大規模統計モデリングと計算統計 III」において結果を紹介した．α 距離による経験グラフは α の値を小さくするとドロネーグラフの部分グラフとなることが経験的にわかっているので，二次元ポアソン過程という限られた場合であっても，ドロネーグラフの最短経路長について確率的な保証が得られたことは，今後のより一般的な理論評価への足掛かりとなることが期待される．

研究テーマ(C)「データの幾何学的構造に則したデータ空間の構成と，その上での統計解析」

本研究では，木グラフのデータ空間(tree space)が CAT(0)という幾何学的特徴をもつため，その上での最短経路長計算や平均，分散の定義ができるという既存の事実を拡張して，新しい統計解析手法を提案することを目指し，デンドログラムという限られた幾何学的構造をもつデータの場合には部分的に成功した．さらに，その得られた結果を用いて，外国語学習者の心内辞書（mental lexicon）の相違を明らかにするという実験データ解析を行った．これは，各

被験者の単語分類結果を一つのデンドログラムとみなし，各群内においての各被験者のデンドログラムの分布の違いについて，並べ替え検定を用いた新しい検定統計量を提案するというものである．また，検定統計量の漸近的な一致性や有効性を証明し，特に群平均法とよばれるデンドログラム構成法が，射影性や局所線形性など好ましい性質を持つことを示した．本研究については，国際学会 ICAPM 2016 で成果を報告し，論文としても投稿中である．

さらに，折田充（熊本大）らの研究グループと共同で，単語のオンライン学習プログラムを開発し，実際に学習効果による被験者のデンドログラムの変化を，並べ替え検定を応用することにより確認した．本成果については，論文3をはじめ複数の論文と学会発表によって公表された．

研究テーマ（D）「データ空間上の情報幾何学の理論と応用」

本研究の目的はパラメータ空間や確率分布空間における情報幾何学ではなく，データ空間上に情報幾何学の理論を構成し，それを応用して新しい統計学，データ解析手法を提案することであった．結論から言うと，さきがけ研究期間中に一般的なデータ空間については情報幾何学を構成することはできなかった．一方，指数型分布族，特に相関行列のもつ双対性に着目し，相関行列の新しい変換手法についての研究を進めた．その際に，空間データの地点間の相関関係を，地図を変形することで視覚化する問題を思いつき，実際に地点データ，各地点での観測ベクトル値および地図データが与えられれば，地図を変形した「相関地図」を作成できるようなソフトウェアを開発した．このソフトウェアおよび手法の説明や，実際にアメダスデータをもとに日本地図を変形した結果は，ホームページ上で公開された．

3．　今後の展開

今回提案された手法は，主に気象データ解析に用いることが多かったが，それ以外のデータにも応用することができる．例えば一般化分散の変化検出は，一般的な時系列データにあてはめることができ，例えば地震の発生データなどがその例である．また，相関による地図の変形ソフトウェアは，交通機関での都市間の移動時間をもとにする地図の変形に応用することが可能である．また，本研究においては気候データのも年周期構造は視覚化やデータの幾何学的構造の確認に用いられていたが，より細かい幾何学的な特徴量を構成できれば，異常気象の検知や予測などにつながる可能性がある．経験グラフによるデータ空間の距離近似精度については，今回非常に単純な場合のみ理論的に評価できたが，近年注目されている研究テーマであり，新しく発表されている成果を参照しながら，より一般的な場合の理論の構築をめざす．さらに，木構造やデンドログラム以外のデータ空間上の統計学や，一般のデータ空間上の情報幾何学は，今回真剣に取り組んでみて，非常に難しいと同時に数学的に奥深いテーマであると気付いたので，群が作用する幾何学的構造などについての知識を深めることにより挑戦を続けていきたい．

4．　評価

（1）自己評価

（研究者）

研究目的については，当初予定していたもののいくつかは達成することができた．特に，研究テーマの（A）「データ空間の曲率を利用した統計的解析」や（B）「経験グラフの構成手法とその理論評価」についての新しい成果は，複数の国際学会において発表，査読付きプロシー

ディングとして発表した．また，その解析プログラムや手法の解析を公開するためのホームページとして作成し，データ空間の統計学という新種の理論や手法の紹介を行うことができた．本領域の特徴として，研究のアウトリーチ活動にも重点を置いており，このホームページがその一端を担えたと考える．一方，（C）「データの幾何学的構造に則したデータ空間の構成と，その上での統計解析」，（D）「データ空間上の情報幾何学の理論と応用」については，当初期待していたものと比べると，理論の想定以上の深遠さや解析データの扱いの困難さから，限定的なものとなった．他方で，当初研究計画に入れていなかったデータの視覚化については，さきがけ研究アドバイザーの先生方からのアドバイスもあり，新たな手法を提案し，その解析プログラムを公開することができた．また，研究の実施体制としては研究補助者を各年１人～２人雇い，プログラミングやデータの整理を中心に研究をサポートしてもらったことが，研究の遂行や成果公開用のホームページの作成に大きな助けとなった．今回扱ったデータのうち，特に気候データは大規模なデータであり，また今回の研究では計算量が大きい統計手法も用いたため，購入した計算機サーバや Mac Pro はデータ解析の大きな助けとなり，有効に研究費を使用することができた．今後は，本研究で開発された手法を，論文や成果公開用ホームページにてより広く知ってもらい，そのフィードバックを参考に改良を加えていきたい．本研究を通して，やがて「データ空間の統計学」という一分野となり，社会問題の解決に大きく貢献できるポテンシャルがあることを確信できた．また，さきがけ研究を通して他分野の数学者とディスカッションをし，いくつかの問題が本質的に解決し，さらに新しい課題を見つけられたことは，本研究者の大きな収穫であった．

（2）研究総括評価（本研究課題について、研究期間中に実施された、年2回の領域会議での評価フィードバックを踏まえつつ、以下の通り、事後評価を行った）。

（研究総括）

データ空間がユークリッド空間の中の一部分に局在していることに着目し，その幾何学的情報を利用してデータ解析の精度を向上させることをめざして研究を行い，データの曲率の制御による判別分析やクラスタリングなどの統計解析の精度向上の方法の提案，その気象データへの応用における従来の統計では見られない現象の検出，ハブの削減による近似精度の維持など，理論と応用の両面にわたる多くの優れた成果を挙げた．また，これらを気象データや心内辞書データに適用して理論の有効性を検証し，統計学に幾何学の手法を取り入れて，汎用性や精度の高い解析法を構築することに成功した．

海外研究者との共同研究や領域内外の研究者との連携も積極的に行い，また，研究の成果を取り入れて，データ解析に活用するためのソフトウェアの開発を行い，それをホームページで公開したこと，領域の趣旨にも合致し，本さきがけ研究の成果の社会への普及という点でも高く評価される．

５．　主な研究成果リスト

（1）論文（原著論文）発表

1. Kobayashi, K., Orita, M. and Wynn, H. (2015) Statistical analysis via the curvature of data space, Bayeasian Inference and Maximum Entropy Methods in Science and Engineering

(MAXENT) 2014, AIP Conf. Proc. 1641, 97, pp. 97-104
2. Hara, K., Suzuki, I., Kobayashi, K. and Fukumizu, K. (2015), Reducing Hubness: A Cause of Vulnerability in Recommender Systems, In proceedings of the 38th Annual ACM SIGIR Conference, pp. 815-818.
3. 折田充, 小林景, 村里泰昭, 神本忠光, 吉井誠, Richard S. Lavin, 相澤一美 (2015), 自律的語彙学習が英語心内辞書構造に与える影響, 九州英語教育学会紀要, 43, pp.1-10.
4. Hara, K., Suzuki, I., Shimbo, M., Kobayashi, K., Fukumizu, K. and Radovanovi, M. (2015) Localized Centering: Reducing Hubness in Large-Sample Data, Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI), pp. 2645-2651.
5. Hara, K., Suzuki, I., Kobayashi, K., Fukumizu, K. and Radovanovi, M.(2016), Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness, Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI), pp.1659-1665

（2）特許出願

なし.

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

【主な学会発表】

Kobayashi, K. and Wynn, H.: Intrinsic and extrinsic means and curvature of metric cones, Algebraic Statistics 2015, Genoa, 2015.6.9

Kobayashi, K.: Generating statistically efficient estimators via computational algebra, Application of Algebraic Methods to Statistics, RIMS, Kyoto University, 2016.6.23 (招待講演)

Kobayashi, K.: Data analysis using curvature of data spaces and their metric cones, The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting(IMS-APRM2016), The Chinese University of Hong Kong, 2016.6.28.

Kobayashi, K. and Wynn, H.: A new aspect of geometrical data analysis using curvature of the data space and the empirical graph, Royal Statistical Society Conference 2017, Glasgow, 2017.9.5.

【著作物】

現代統計学, 日本評論社, (第１１章「統計における最適化」執筆)