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

令和　５ 年　 ６ 月　 ２ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ０３６４２

研究課題名（和文）データに潜在する曲率情報に着目した統計解析手法の開発

研究課題名（英文）Development of statistical analysis methods focusing on curvature information latent in data

研究代表者
小林　景（Kobayashi, Kei）

慶應義塾大学・理工学部（矢上）・准教授

研究者番号：９０４６５９２２

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究では，データ空間の距離および曲率に着目し，α，β距離という２種類の距離変換を組み合わせた新しい統計解析手法を提案した．さらに，その手法を，年周期により生じる幾何学的特徴を活かした一般化分散の計算に応用し，降雨量データの新しい異常変動を捉えることに成功した．また，β距離変換の提案時にデータ解析に始めて導入した計量錐の理論を，グラフ埋め込みに応用し，ネットワークデータの階層構造を自動的に抽出する新しい手法を提案した．これらの結果を国際学会および学術論文にて発表した．

研究成果の学術的意義や社会的意義
本研究成果の学術的意義は，まずその新規性にある．従来のデータ解析は，与えられたデータ間の距離をそのまま用いるか，もしくは次元削減等により得られた特徴ベクトル間の距離を用いることがほとんどであったが，本研究では，データやそれが分布する空間の距離をうまく変換した上でデータ解析を行うという点が大きく異なる．これにより，曲率等の幾何学的な特徴量に着目したデータ解析手法の開発および改良が可能となった．また，これまで用いられたことがなかった計量錐をデータ解析に応用したことも今後の発展につながる大きな新規性を含んでおり，学術的意義の高い研究成果と言える．

研究成果の概要（英文）：In this study, we proposed a new statistical analysis method that combines two types of distance transformations, namely α and β distances, by focusing on the distance and curvature of the data space. Furthermore, we applied this method to calculate generalized variances that take advantage of geometric features caused by annual cycles, successfully capturing new abnormal variations in rainfall data. Additionally, we applied the theory of metric cones, which was introduced for the first time in data analysis during the proposal of β distance transformation, to graph embedding, proposing a new method that automatically extracts the hierarchical structure of network data. These results were presented at international conferences and published as academic papers.

研究分野：統計科学

キーワード：幾何学的データ解析　機械学習　データ埋め込み　距離変換　クラスタリング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1) 大規模データ解析においては，データからいかに解析に有効な特徴量を抽出ができるかが解析精度向上の大きな鍵となる．統計学においては，計算機の普及以前から，線形回帰分析，主成分分析などの線形的な解析手法が用いられており，その後に計算機が普及してからは，セミパラメトリック解析，ノンパラメトリック解析やカーネル法，ニューラルネットワーク，ブースティングなどの機械学習理論が盛んとなり今日に至っている．これらの解析手法は，基本的にはベクトルや関数空間などの線形空間の特徴を活かしたものである．

(2) その一方で，本研究が着目する測地距離空間や多様体上やその周辺にデータが分付する場合に，測地距離や曲率を用いて解析する分野も少数派ながら存在する．たとえば，球面統計学，方向統計学などは球面上のデータを扱い，また Shape Analysis と呼ばれる分野では，平行移動，回転や拡大縮小を無視した上での画像データの類似性を調べる Procrustes 解析とよばれる手法が用いられてきた．金字塔的な結果としては，Kendall(1990)による定曲率空間上での Fréchet 平均の一意性の理論があり，多様体の上での統計解析の理論の基礎が構築された．さらに近年になり，多様体状にデータが近似的に分布するという「多様体仮説」が高次元実データで実際に観測されるようになり，経験グラフ上の測地距離を用いてデータ解析を行う多様体学習や，Tree Space とよばれる多面体複体の CAT(0)性を用いた遺伝系統樹解析などが研究されている．

(3) 本研究は，測地距離や曲率に着目するという点ではこれらの先行研究に続く位置にあるが，単に与えられた距離空間について解析するのではなく，データの空間の距離を変換した上で解析を行うという新しい着想に基づく発展研究として位置づけられ開始された．

２．研究の目的

(1) 本研究では，データが潜在的に持つ幾何学的特徴量を新たな手法で抽出し，それらを利用して全く新しいデータ解析手法を開発することが目的である．幾何学を用いることにより，幾何学分野において長年蓄積されてきた深く膨大な知識を積極的に活かせることが本アプローチの大きな利点となる．

(2) また幾何学的な特徴量の中でも，特にデータ空間の測地線および曲率に着目しており，「データのもつ『曲率』はデータ解析で重要な役割を果たすのか？」という根本的な問題に挑戦し，答えを出すことが大きな目標でもあった．そのために，理論に基づき手法を提案するだけでなく，提案した解析手法を実際のデータに応用し有効性を確認する．

(3) さらに，曲率のデータ解析における重要性が確認できた後は，それを既存の統計解析手法や機械学習手法と組み合わせることにより，より広範囲への応用へと繋げるという目的もある．

３．研究の方法

(1)　London School of Economics の Henry P. Wynn との共同研究として，データ測地距離空間の CAT(k)曲率に着目したデータ解析手法の開発を行った．研究期間中に計２回，共同研究者の元を訪問し，その他メールやオンライン会議システムを用いて研究ディスカッションを重ね研究を進めた．2022 年度には London School of Economics に訪問研究員として４ヶ月間ほど滞在し，特に測地線の詳細な理論を構築するための研究を行った．

(2) 童祺俊（当時 Albert）とのエントロピー制約付きの Wasserstein 距離を研究する際には，主に対面およびオンラインでのセミナー，ミーティングを用いた．

(3) 竹原大翼（当時 Albert）との計量錐へのグラフ埋め込みによるネットワークデータの階層情報抽出に関する研究においては，主に対面およびオンラインでのセミナー，ミーティングを行った．また，Google Colaboratory や Amazon によるクラウドサーバ（AWS）を用いてデータ解析の計算を実行した．

(4) 熊本大学の折田充教授を中心とする外国語学習の研究者グループとともに，階層クラスタリングで用いられるデンドログラム間距離を用いて，英単語心内辞書の幾何学的構造についての研究を行った．特に Google Firebase を用いた英単語仕分け課題用のオンライン診断テストプログラムを，当時慶應の院生であった童祺俊，保母将希，田保健士郎らの協力のもと開発した．オンライン診断テストは，熊本大学および東京電機大学の学生に対して，英単語学習教材の効果を評価するために用いられた．

(5) その他，当時慶應義塾大学の学生であった鴨井遼，田保健士郎，保母将希，服部航大，山下

亮らとともに機械学習および統計解析分野の新しい諸手法についてセミナーやミーティングを重ねた．機械学習のための計算には，主に Google Colaboratory や Amazon のクラウドサーバ（AWS）を用いた．

４．研究成果

(1) Henry P. Wynn（London School of Economics）との共同研究では，α，β距離とよばれる新しく導入された距離変換を用いて，データのベクトルとしての距離を変換したうえでデータ解析を行う手法を開発した．α距離は局所的な測地距離変換であり，パラメータαを変化させることにより，空間の曲率を単調に変化させることが可能である．また，β距離は大域的な距離の変換であり，変換後は必ずしも測地距離にならないことから，そのままでは曲率を定義し評価することができない．一方で，β距離はデータ空間を計量錐とよばれる測地距離空間に埋め込んだ上での extrinsic 距離（外部を通ることを許容した最短経路長）として解釈することができ，またその計量錐の曲率はβにより単調に変化することが示される．このことを用いて，間接的にデータ空間自体の曲率の単調変化と似たことが実現可能となる．このように曲率の単調変化が保証されることから，２つのハイパーパラメータによるチューニングが理論的に正当化される．また，応用上どのようにこれらのハイパーパラメータを設定するかについても，いくつかの手法を提案した．以上のように開発された距離変換を応用して，イギリスの降雨量データの日ごとのばらつきの年次変化解析や地球上の人口分布データやその他のベンチマークデータに対するクラスタリングを行った．提案された手法，および理論と実験結果は論文[1]にまとめて発表された．

(2) 童祺俊（当時 Albert）との共同研究では，正規分布の場合のエントロピー制約付きの Wasserstein 距離を具体的に導出した．Wasserstein 距離は Kullback-Leibler ダイバージェンスなどの従来統計学で用いられていた距離やダイバージェンスと異なり，データの空間の距離情報を直接的に用いるため，機械学習等の勾配法によるアルゴリズムへの応用が注目されている．一方，ソフトウェアを用いた計算においては，エントロピー制約を付けることにより Sinkhorn アルゴリズムを用いることが可能となり，計算量は大幅に減少することが知られている．これまでは，こういった計算の簡便さのためだけに用いられてきたエントロピー制約付き Wasserstein 距離について，その理論を構築し，統計的有効性を実験的に示したという点で，本研究の成果は非常に重要である．また，（エントロピー制約付き）Wasserstein 距離の距離空間は曲率を持つ空間であり，その曲率がデータ解析に与える影響や有効性を評価するための基盤となる理論を構成することができた．本研究の成果は，論文[3]によって発表された．

(3) 竹原大翼（当時 Albert）との共同研究の成果として，事前学習でまずユークリッド空間等の測地距離空間にネットワークを埋め込んだ後，計量錐とよばれる別の測地距離空間に埋め込むことにより，階層構造を自然に抽出しつつ距離変換不変な手法を提案し，単語間ネットワーク WordNet 等に応用した．本研究以前も，空間の負曲率性を用いたグラフ埋め込手法として，Poincare 埋め込みや Lorentz モデルが提案されていたが，計量錐を用いたグラフ埋め込みは全く新しい手法であり，１次元のパラメータのみを学習すればよく，またハイパーパラメータによって計量錐の曲率を調整することにより，より精度良く階層構造を反映した埋め込みが可能であることを実験的に示した．また，その埋め込みの一意性に関する定理を証明し，提案手法の理論的妥当性も示すことができた．研究成果は論文[2]によって出版された．

(4) 折田充教授（熊本大学）らとの共同研究では，英単語を意味の近いグループ同士に仕分けするという課題の結果をもとに，被験者である学生の心内辞書デンドログラムを階層クラスタリングにより構成し，デンドログラム間の距離を定義することにより並べ替え検定を用いて被験者グループ間の心内辞書の差異の有無を調べた．特に，研究グループにより開発された学習教材を用いた学習の前後で仕分け結果がどの程度変化したかを，被験者の語彙レベルや品詞の影響に着目して調べた．研究成果は論文[4]をはじめ多数の論文や学会発表を通して報告された．

(5) 深層生成モデルの最終隠れ層の分布情報を用いた異常データ検出についての鴨井遼との共同研究の成果は論文にまとめられ SafeAI 2020 で報告された．また，田保健士郎との共同研究による日本語ＱＡデータの傾向の調査，保母将希との共同研究による繰り返しゲームの新しい強化学習手法の開発，服部航大との共同研究による多様体学習結果の精度の可視化手法の提案，山下亮との共同研究による新しいロバストダイバージェンスによるベイズ推定手法の提案の各成果については，国内学会において発表された．

[1] Kobayashi, K. and Wynn, H. (2020), Empirical geodesic graphs and CAT(k) metrics for data analysis, Statistics and Computing, 30(1), 1-18.
[2] Takehara, D. and Kobayashi, K. (2021): Enhancing Hierarchical Information by Using Metric Cones for Graph Embedding, arXiv:2102.08014 (学術誌 Mathematics に受諾済み)
[3] Tong, Q. and Kobayashi, K. (2021), Entropy-regularized optimal transport on multivariate normal and q-normal distributions, Entropy, 23(3), 302.

[4] 折田 充，小林 景，村里 泰昭，相澤 一美，レイヴィン リチャード，神本 忠光，吉井 誠 (2023)，英語心内辞書の再構築・変容を促進する語彙，学習プログラム―ネイティブ度診断テスト導入の効果―，ARELE，30，161-176．

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 折田 充，村里泰昭，小林 景，吉井 誠・Richard Lavin，相澤一美，神本忠光	４．巻 63.64
２．論文標題 語彙サイズの異なる大学生の英語心内辞書	５．発行年 2021年
３．雑誌名 熊本大学英語英文学	６．最初と最後の頁 207-222
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Qijun Tong, Kei Kobayashi	４．巻 23(3)
２．論文標題 Entropy-regularized optimal transport on multivariate normal and q-normal distributions	５．発行年 2021年
３．雑誌名 Entropy	６．最初と最後の頁 302
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/e23030302	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 折田 充，村里泰昭，小林 景，神本忠光，相澤一美，Richard Lavin，吉井 誠	４．巻 48
２．論文標題 心内辞書内の単語の結びつき方---英単語学習プログラムへの取り組み前、直後、１年半後---	５．発行年 2020年
３．雑誌名 KASELE BULLETIN	６．最初と最後の頁 9-18
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kei Kobayashi and Henry Wynn	４．巻 30(1)
２．論文標題 Empirical geodesic graphs and CAT(k) metrics for data analysis	５．発行年 2020年
３．雑誌名 Statistics and Computing	６．最初と最後の頁 1-30
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11222-019-09855-3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Ryo Kamoi and Kei Kobayashi	4．巻 -
2．論文標題 Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions	5．発行年 2020年
3．雑誌名 Proceedings of the Workshop on Artificial Intelligence Safety (SafeAI 2020)	6．最初と最後の頁 113-116
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 折田 充，小林 景，村里 泰昭，相澤 一美，レイヴィン リチャード，神本 忠光，吉井 誠	4．巻 34
2．論文標題 英語心内辞書の再構築・変容を促進する語彙学習プログラム―ネイティブ度診断テスト導入の効果―	5．発行年 2023年
3．雑誌名 ARELE	6．最初と最後の頁 161-176
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計11件（うち招待講演　1件／うち国際学会　3件）

1．発表者名 小林 景
2．発表標題 Schoenbergの理論とその相関行列変換への応用
3．学会等名 統計関連学会連合大会
4．発表年 2021年

1．発表者名 保母 将希，小林 景
2．発表標題 CFRによる不完全情報繰り返しゲームに対する近似的な均衡戦略
3．学会等名 第16回統計学会春季集会
4．発表年 2022年

１．発表者名
田保 健士郎，小林 景

２．発表標題
QAにおける評価用データセットの役割と日本語QAデータセットの必要性についての考察

３．学会等名
言語処理学会第28回年次大会ワークショップJED2022

４．発表年
2022年

１．発表者名
折田 充，村里 泰昭，小林 景，神本 忠光，相澤 一美， 吉井 誠，Richard Lavin

２．発表標題
英語心内辞書における名詞の結びつき --再構築・変容、精緻化--

３．学会等名
全国英語教育学会第46回長野研究大会

４．発表年
2021年

１．発表者名
Kei Kobayashi

２．発表標題
A new aspect of positive definiteness for correlation matrices

３．学会等名
統計関連学会連合大会2019

４．発表年
2019年

１．発表者名
Kei Kobayashi

２．発表標題
Statistical inference and data analysis on length metric spaces

３．学会等名
32nd European Meeting of Statisticians（国際学会）

４．発表年
2019年

１．発表者名
Ryo Kamoi and Kei Kobayashi

２．発表標題
Out-of-Distribution Detection with Likelihoods Assigned by Deep Generative Models Using Multimodal Prior Distributions

３．学会等名
The AAAI’s Workshop on Artificial Intelligence Safety (SafeAI 2020)（国際学会）

４．発表年
2020年

１．発表者名
服部航大，小林景

２．発表標題
データ空間の計量に着目した多様体学習の評価

３．学会等名
第17回統計学会春季集会

４．発表年
2023年

１．発表者名
山下亮，小林景

２．発表標題
ロバストダイバージェンスに基づく事後分布構成法の改良

３．学会等名
第17回統計学会春季集会

４．発表年
2023年

１．発表者名
折田 充，小林 景，村里 泰昭，吉井 誠，Richard Lavin，相澤 一美

２．発表標題
英語心内辞書の再構築・変容を促進する語彙学習プログラム ―ネイティブ度診断テスト導入の効果

３．学会等名
全国英語教育学会第47回北海道研究大会

４．発表年
2022年

１．発表者名 Kei Kobayashi, Henry P. Wynn
２．発表標題 Data analysis focusing on geodesic distance and curvature
３．学会等名 Algebraic Statistics 2022（招待講演）（国際学会）
４．発表年 2022年

〔図書〕　計1件

１．著者名 青木敏，伊藤陽一，岩崎学，紙屋英彦，黒住英司，小林景，佐井至道，清水泰隆，鈴木大慈，清智也，寒水孝司，竹村彰通，中西寛子，橋口博樹，原尚幸，日野英逸，姫野哲人，松浦峻，山田秀，汪金芳	４．発行年 2020年
２．出版社 学術図書出版社	５．総ページ数 330
３．書名 統計学実践ワークブック（８章「統計的推定の基礎」執筆）	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
英国	London School of Economics			