様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　１　日現在

機関番号：　６２６０３
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７００２８８

研究課題名（和文）代数的性質を用いた新しい統計解析手法の開発

研究課題名（英文）Novel tools of statistical analysis via their algebraic properties

研究代表者
小林　景（Kobayashi, Kei）

統計数理研究所・大学共同利用機関等の部局等・助教

研究者番号：９０４６５９２２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：代数的な統計モデルに対する，全く新しい代数的な推定量を提案し，その推定法手式とその解（推定値）を計算する方法を実際に示した．この推定量は漸近有効性と尤度方程式の低次性の両方を実現する．これは，「統計学的な有効性と代数計算の簡便性のトレードオフの解析」という全く新しい研究の可能性を示したという点でも重要である．一方，デンドログラムデータの幾何学的な特徴を用いた新しい検定手法を提案し，その理論的妥当性を示した．また，それを用いて，被験者群間の英単語心内辞書の差異の有無を検定した．本研究は，複数の国際会議や論文で発表された．

研究成果の概要（英文）：A class of novel algebraic estimators for algebraic models is proposed. We proposed an explicit method to compute the estimators and estimates by using computational algebraic methods. The estimators are defined by some lower-degree polynomial equations by holding some asymptotic statistical efficiency. This allows fast computation of the estimates. This research should be the first result evaluating and controlling trade-off between algebraic computation and statistical efficiency. Another result is a new permutation test for dendrograms by using their geometrical and algebraic structures. We tested difference between mental lexicons between groups of examinees by using the proposed method. The results of this research project have been published in several international conferences and journals.

研究分野：　統計科学

キーワード：　代数統計学　漸近統計学　情報幾何学　高次元データ解析

１．研究開始当初の背景
これまでの統計学や機械学習の分野では，モデルや推定量，またその評価基準となる関数が代数的な（多項式で表される）場合であっても，その特徴を用いずに推定値やリスクなどを計算していた．　一方近年，グレブナー基底等を用いた計算機代数ソフトウェアの性能が向上し，「代数的だからこそ有効に計算できる」手法が実用段階に入った．本研究では，こういった新たな手法を用いることにより飛躍的に計算効率が改善されるような，統計的解析手法を提案する．
また，大規模ランダム行列理論は，固有値分布の極値理論である Tracy-Widom 則の条件の緩和など，近年目覚ましい発展を見せている．その一方，その統計学への応用は限定的である．本研究では，大規模ランダム行列理論の結果やその手法を用いた，非可換代数の構造を持つデータ（行列データ）やモデルの解析を行う．

２．研究の目的
本研究の主たる目的は，代数的手法を用いて新しい統計的データ解析法を開発することである．本研究の目的としては，大きく分けて（１）計算機代数を用いた統計的解析手法の開発及び（２）大規模ランダム行列データの解析手法の開発がある．代数学の分類としては，（１）は主に可換代数を用い，（２）は主に非可換代数を用いるが，理論および応用で両者は共に関係し合い，この２つの研究を同時に進めることによってそれぞれのブレークスルーとなるような発展をめざすことも本研究の特徴のひとつである．
また，代数統計の一分野としても注目されているデータ空間の曲率や CAT(k)特性に注目したデータ解析についても，新しい統計解析手法の提案や理論評価を行う．これにより，情報幾何学などの既存の方法とは異なるアプローチで，幾何学分野における理論を統計解析に応用することが可能となる．

３．研究の方法
（１）代数的統計モデルに対して，統計的な妥当性を持ち，かつ代数計算により計算可能な推定量を自動構成する方法について，代数統計学の第一人者である Henry P. Wynn（ロンドン・スクール・オブ・エコノミクス）と共同研究を行い，これまでの代数統計学の枠組みにとらわれない新しい手法を提案する．そのため，研究代表者のイギリスへの渡航や，共同研究者の日本への招待を頻繁に行い，ディスカッションを重ねた．また，Maple などの計算機代数用ソフトウェアを用いて，具体的に推定量を計算するプログラムを構成した．また，Genova，大阪，京都で開催された代数統計に関する国際学会や，Ambois で開催された情報幾何学の国際学会で，それぞれ代数統計や情報幾何学の専門家と成果についてディスカッションを行った．
（２）幾何学的な特徴量を用いた統計解析についても，Henry P. Wynn との共同研究を中心に研究を進めた．また，木グラフの構造を持つようなデータに関して，木空間とよばれる測地空間上の測地線を用いる手法の英単語心内辞書解析への応用について，折田充（熊本大学）を始めとする外国語学習の研究者グループとの共同研究を進めた．
（３）大規模ランダム行列理論の生産管理，特にフローショップ型スケジューリングに関する問題への応用に関しては，新里隆（一橋大学）や郭偉宏（東京都立大学）との共同研究を通して，これまでとは異なる分野への大規模ランダム行列理論の応用を行った．また，パーコレーションなどの統計物理学の理論も用いた．

４．研究成果
（１）代数的な統計モデルに対する，全く新しい代数的な推定量を提案し，その推定法手式とその解（推定値）を計算する方法を実際に示した．具体的には，モデルと推定量が代数的に（多項式を用いて）定義される場合を考える．この仮定は一見強く思われるが，ポアソン分布や多項分布などの分割表モデルや，ガウス分布で共分散行列に多項式制約がつく場合など適用範囲は広い．
この場合において，まずフィッシャー情報計量やアファイン接続，埋め込み曲率などの情報幾何学的量を代数的に計算し，二次漸近有効性の十分条件をこれらを用いて表す．この十分条件を満たす代数的な推定量のクラスは，二次漸近有効な推定量全体の中で充分豊かであり，また局所的に推定値の一意的な存在を示すことができる．次に，この二次漸近有効な推定量のクラスの推定方程式は代数的に特定の形をしていることから，その中に２次以下の連立多項式方程式で表されるようなものが存在することが示される．また，尤度方程式からグレブナー基底による剰余を行うことにより，その連立多項式方程式を導出することができる．
多項式の次数が下がると，ホモトピー連続化法などの数値計算手法を用いた推定値の計算の計算量を本質的に削減できるという利点がある．実際，いくつかの簡単な問題に対して，数値実験によって計算量の本質的な削減を確認できた．
本研究は，代数統計学のこれまでの主流であった，既存の推定量，特に最尤推定量の推定方程式の次数（ML-degree）を各モデルごとに評価するという手法とは全く異なるアプローチであり，各統計モデルに依存して新しい推定量を「作る」という点が新しい．これは，「統計学的な有効性と代数計算の簡便性のトレードオフの解析」という全く新しい研究の可能性を示したという点でも重要である．本研究は，複数の国際会議や論文で発表され，その新規性は査読者に高く評価された．

（２）デンドログラムデータの幾何学的な特徴を用いた新しい検定手法を提案し，その理論的妥当性を示した．また，それを用いて，被験者群間の英単語心内辞書の差異の有無を検定した．

具体的には，並べ替え検定と呼ばれる統計的手法を用い，「２つのグループの心内辞書に差異は無い」という仮説のもとでダミーのデンドログラムを多数生成させ，それらと実際に得られたデンドログラムを比較することにより，定量的に差異を評価する手法を提案した．ただし，並べ替え検定で通常扱われる統計データは，通常はベクトルなどの単純な構造を持っているが，デンドログラムの集合の幾何学は，単体的扇と呼ばれる構造を持っていることが知られている．一般的な単体的扇は，複雑な構造のため解析が困難だが，デンドログラムの集合は，その中においてCAT(0)とよばれる数学的に良好な性質を持っている．そこで，測地線を一意に定義でき，またそれを効率的に計算できるという利点を用いて，測地距離を用いた新たな並べ替え検定も提案した．

ここで提案した手法は全く新しいものなので，その正当性を評価する必要がある．そこで，デンドログラムの構成法として，局所線形性という性質をみたすものを用いれば，提案した並べ替え検定は一致性を持つことを証明した．特に，Lance-Williams 法とよばれるクラスター解析のデンドログラム構成法のクラスの中で，群平均法がこの局所線形性と，射影性という好ましい特徴をともに持つということを示した． この理論解析において，デンドログラム空間の CAT(0)性を用いた．デンドログラム空間は木空間の部分集合であり，木空間は CAT(0)性を持つことは知られていたが，その性質は必ずしも部分集合に引き継がれないことから，この事実は自明ではなく，証明が必要であった．

心内辞書解析としては，冒頭で述べた実験データのみでなく，その後も英語初級者を被験者にした場合や，日本語訳語を用いた場合など，様々な場合について同様な実験と解析を行なった．本研究で得られた心内辞書の差異の解析結果を用いて，学習効率の高い英語学習教材の開発も期待できる．実際，折田先生を中心とした英語学習研究者のグループと共に，教材開発をめざした共同研究も進行中である．

本研究の成果は，多数の国内，国際会議や論文で発表された．

（３）生産管理，特にフローショップスケジューリングについて，大規模ランダム行列の理論を用いる総加工時間評価手法を提案した．

具体的には，統計物理や確率論で研究されているサイトパーコレーションの理論を用いてフローショップの総処理時間（メイクスパン）の漸近的な分布を評価した．特に，フォワード法，バックワード法のメイクスパン関数は漸近的に，shape function とよばれる関数と一致し，ハイブリッド法とよばれるスケジューリング手法では，二つの shape function を足し合わせたものとなった．また，shape function からの誤差は漸近的に Tracy-Widom 則とよばれる大規模ランダム行列の固有値の極値分布として知られる確率分布に従うことが示される．本研究は，大規模ランダム行列理論やパーコレーションの新しい応用分野を開拓したという意義がある．また，本成果は国際会議で発表され，論文執筆中である．

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 11 件）

① 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)，日本人英語学習者の英語心内辞書の変容，熊本大学社会文化研究，13，15-30．

② 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，43，1-10．

③ 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.

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 Kobayashi, K., Orita, M. and Wynn, H. (2015) Statistical analysis via the curvature of data space, BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING (MAXENT 2014), AIP Conf. Proc. 1641, 97 (2015), pp. 97-104, doi:10.1063/1.4905968.

⑥ 折田充，小林景，村里泰昭，相澤一美，吉井誠，Lavin, R. (2014) 英語熟達度と心内辞書内の意味的クラスタリング構造の関係，九州英語教育学会紀要，42，1-10．

⑦ Kobayashi, K. and Wynn, H. (2014) Computational algebraic methods in efficient estimation, Geometric Theory of Information (Signals and Communication Technology), 119-140, doi:10.1007/978-3-319-05317-2_6.

⑧ 折田充, 小林景, 村里泰昭, Lavin, R., 吉井誠, 相澤一美, 神本忠光 (2014) 英語母語話者と日本人英語学習者の心内辞書における語彙項目間類似度の比較, 熊本大学社会文化研究, 12., pp. 11-24.

⑨ Kobayashi, K. and Wynn, H. (2013) Asymptotically Efficient Estimators for Algebraic Statistical Manifolds, Geometric Science of Information: Lecture Notes in Computer Science, 8085, pp. 721-728.

⑩ 折田充, 小林景, 村里泰昭, 神本忠光, 吉井誠, Lavin, R. (2013) 語彙サイズと心内辞書内の意味的クラスタリング構造の関係, KASELE Bulletin, 41, pp. 1-10.

⑪ 折田充, 小林景 (2013) 日本語の心内辞書と英語の心内辞書－日本人英語学習者における日英語間で対応する訳語関係にある高頻度形容詞群の意味的クラスタリング構造, 熊本大学社会文化, 11. pp. 21-34.

〔学会発表〕(計 28 件)

① Hara, K., Suzuki, I., Kobayashi, K. and Fukumizu, K. and Radovanovic, M.: Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness, 30th AAAI Conference on Artificial Intelligence (AAAI), Phoenix, USA, 2016.2.15(poster)

② Kobayashi, K. and Orita, M.: Geometry of dendrogram space and its application to mental lexicon analysis, 6th International Conference on Applied Physics and Mathematics (ICAPM 2016), Singapore, 2016.1.14.

③ Hara, K., Suzuki, I., Kobayashi, K. and Fukumizu, K. and Radovanovic, M.: Reducing Hubness for Kernel Regression, SISAP2015, Glasgow, 2015.10.12. (poster)

④ Shinzato, T., Kaku, I. and Kobayashi, K.: A Discussion on Universality of Makespan in Flow Shop Scheduling Problem, 2015 Asian Conference of Management Science & Applications, Dalian, China, 2015.9.13.

⑤ 小林景 (2015) 高次元データにおける近傍構造の指標と統計的解析, 統計関連学会連合大会, 岡山大学, 2015.09.07.

⑥ Hara, K., Suzuki, I., Kobayashi, K. and Fukumizu, K.: Reducing Hubness: A Cause of Vulnerability in Recommender Systems, In proceedings of the 38th Annual ACM SIGIR Conference, pp. 815-818, Santiago de Chile,2015.8.11.(poster)

⑦ Kobayashi, K. and Wynn, H.: Intrinsic and extrinsic means and curvature of metric cones, Algebraic Statistics 2015, Genoa, 2015.6.9 (poster)

⑧ Kobayashi, K.: Geodesic distances on data spaces: their computation and modification, ISI-ISM-ISSAS joint Conference 2015, Tokyo, 2015.4.2

⑨ 小林景 (2015) データ空間の測地距離と曲率の計算統計, 大規模統計モデリングと計算統計,東大駒場, 2015.02.06.

⑩ Hara, K., Suzuki, I., Shimbo, M., Kobayashi, K., Fukumizu, K. and Radovanovic, M.: Localized Centering: Reducing Hubness in Large-Sample Data, The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), Austin, Texias, 2015.1.29.(poster)

⑪ 折田充, 小林景, 村里泰昭, 神本忠光, 吉井誠, Lavin, R., 相澤一美 (2014) 自律的語彙学習が英語心内辞書構造に与える影響, 第 43 回九州英語教育学会大分研究大会, 大分, 2014.12.06.

⑫ 小林景 (2014) データ空間の曲率を用いた統計解析手法の改良, 京都大学数理解析研究所共同研究　統計多様体の諸分野への応用, 京大数理解析研究所, 2014.11.20.

⑬ 小林景, Wynn, H. (2014) データ空間の曲率と距離変形を用いた解析手法, 統計関連学会連合大会, 東京大学, 2014.09.14.

⑭ Kobayashi, K., Orita, M. and Wynn, H.: Statistical analysis via the curvature of data space, MaxEnt 2014, Amboise, 2014.9.22.

⑮ 折田充, 小林景, 村里泰昭, 神本忠光, 吉井誠, Lavin, R., 相澤一美 (2014) 日本人大学生の英語心内辞書の変容, 全国英語教育学会　第４０回　徳島研究大会, 徳島, 2014.08.09.

⑯ Kobayashi, K. and Wynn, H.: The empirical geodesic graphs and their deformation for data analysis, ASC-IMS 2014, Australian Technology Park, Sydney, 2014.7.9.

⑰ Kobayashi, K.: Curvature of empirical metrics on a data space and its deformation, Workshop on Mathematical Approaches to Large-Dimensional Data Analysis, ISM, Tokyo, 2014.3.14.

⑱ Kobayashi, K.: Hypothesis Testing for the Difference of Dendrograms,

ISI-ISM-ISSAS Joint Conference 2014, ISI, Delhi, 2014.2.20.

⑲ Kobayashi, K. and Wynn, H.: The empirical geodesic graphs and a deformation of their metric, Computational Algebraic Statistics, Theories and Applications (CASTA 2014), Kyoto, 2014.1.23.

⑳ 小林景，折田充 (2013) 木構造およびクラスター構造をもつデータの測地的解析手法，p.313，統計関連学会連合大会，大阪大学，2013.9.11.

21 Kobayashi,K. and Wynn,H.: Asymptotically Efficient Estimators for Algebraic Statistical Manifolds, First International Conference on Geometric Science of Information 2013, Ecole des Mines de Paris, 2013.8.28.

22 Kobayashi,K.: An algebraic computation method for asymptotically efficient estimators, Joint Meeting of the IASC Satellite Conference and the 8th Conference of the Asian Regional Section of the IASC, Seoul, 2013.8.23 (invited Talk).

23 折田充，小林景，村里泰昭，相澤一美，吉井誠，Lavin, R. (2013) 英語熟達度と心内辞書内の意味的クラスタリング構造の関係，第 39 回全国英語教育学会北海道研究大会，北星学園大学，2013.8.11.

24 Kobayashi, K.: The best upper bound on total variation distance by DeRobertis separation, 8th World Congress in Probability and Statistics, Istanbul, 2013.7.9.

25 折田充，小林景，村里泰昭，神本忠光，吉井誠，Lavin, R. (2012) 語彙サイズと心内辞書内の意味的クラスタリング構造の関係，九州英語教育学会.

26 折田充，小林景 (2012) 母語の心内辞書と第二言語の心内辞書（3）―日英 語間で訳語関係にある高頻度形容詞群の意味的クラスタリング構造，第 38 回全国英語教育学会愛知研究大会.

27 小林景，折田充 (2012) 英語心内辞書の木構造データ解析の新手法，日本行動計量学会，pp.101-104，新潟県立大学，2012.9.14.

28 Kobayashi, K. and Wynn, H.: Asymptotic estimation theory via algebraic computation, The 2nd Institute of Mathematical Statistics Aisa Pacific Rim Meeting (IMS-APRM), Tsukuba, 2012.7.3.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
小林　景　（KOBAYASHI, Kei）
情報・システム研究機構　統計数理研究所・数理・推論研究系・助教
研究者番号：９０４６５９２２

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：