様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月８日現在

機関番号：６２６０３
研究種目：若手研究（B）
研究期間：２００８　～　２０１１
課題番号：２０７００２５８
研究課題名（和文）　大規模ランダム行列を用いたモデル選択と機械学習理論

研究課題名（英文）　Model selection and machine learning theory via large-scale random matrices

研究代表者
小林　景（KOBAYASHI KEI）
統計数理研究所・数理・推論研究系・助教
研究者番号：９０４６５９２２

研究成果の概要（和文）：
カーネルマシンのモデル選択に必要な，カーネルグラム行列の Nyström 近似では次元削減を二段階で行い，それぞれに対して次元削減割合に対応するパラメータを設定する必要がある．ここで，近似が二段階であるため，計算量と近似誤差の間にはトレードオフの関係が成立し，次元削減パラメータの最適化の必要が生じる．そこで本研究では変分法等の計算物理学的な大規模ランダム行列解析の手法を用いて，分布に関する適当な仮定のもとでこのパラメータを最適化した．また，手書き文字データにおいてカーネルグラム行列の次元削減の最適化が有効に働くことを実験的に確かめた．また Nyström 法の近似誤差の PAC 学習的な上界を列の復元抽出の場合，列の非復元抽出の場合に分けて証明し，結果として一致性も証明した．さらに発展させ，Sparse greedy approximation や Incomplete Cholesky decomposition などの他のカーネルグラム行列の近似手法にも適用できることを示した．この研究と並行して，可換代数的な解析手法として計算機代数を漸近推定理論に応用し新しい推定量を提案した．また代数的なモデルの例として木構造を持つ心内辞書デンドログラムの新しい解析手法を提案した．

研究成果の概要（英文）：
Nyström approximation method for kernel gram matrices reduces the rank of each matrix in two steps. In order to approximate those matrices efficiently, it is important to set an adequate reduction rate for each step and handle the tradeoff between accuracy of the approximation and cost of the computation. In this research program, we used methods of computational physics for analyzing large-scale random matrices and optimized the reduction rates. We checked experimentally that the proposed method attains high accuracy even with very low computational cost for real data of hand-written characters. We derived an upper bound for approximation error of Nyström method and proved the statistical consistency. Moreover, the proposed method can be used not only for Nyström method but also for other approximation methods including the sparse greedy approximation and the incomplete Cholesky decomposition. In parallel with this research, we studied commutative algebraic statistics and proposed a novel statistical estimator by applying the computational algebra to the asymptotic estimation theory. In addition, we proposed a statistical method to analyze dendrograms of mental lexicon, which is an example of models holding an algebraic structure.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０年度	1,100,000	330,000	1,430,000
２１年度	900,000	270,000	1,170,000
２２年度	600,000	180,000	780,000
２３年度	600,000	180,000	780,000

総　計	3,200,000	960,000	4,160,000

研究分野：統計科学
科研費の分科・細目：情報学・統計科学
キーワード：カーネルグラム行列，カーネルマシン，モデル選択，大規模ランダム行列

１．研究開始当初の背景

パターン認識や機械学習で用いられるカーネル法では，データ数×データ数のサイズのカーネルグラム行列を計算する必要があり，数万～数十万のオーダーのデータ数を扱う場合，Nystrom 近似法などのカーネルグラム行列の近似計算は，不可欠である．一方，数学や物理学の分野で，大規模なランダム行列の固有値の漸近理論が研究が近年急速に進んできたが，未だにデータ解析への応用例は限られている．本研究では，大規模ランダム行列理論など代数的手法を用いることでその問題を解決することをめざした．

２．研究の目的

本研究の目的は，大規模ランダム行列理論や代数的手法を用いた大規模データのモデル選択手法の開発及び機械学習理論への応用である．近年機械学習分野で想定されることの多い巨大なデータを解析する際には，既存の統計学的手法を用いることはできない．その主たる問題点として (i) データの次元がサンプル数と同程度か，それより多いという p>>n 問題，(ii)計算量の問題，(iii)モデルの構造化の問題の三点があげられる．一方近年の確率，統計学および統計物理の両分野における大規模ランダム行列理論の発展はめざましい．本研究では，これら両分野の理論と手法を統一することにより，上に述べた大規模データの解析の問題点(i)～(iii)を解決することを目指す．

３．研究の方法

大規模ランダム行列理論の研究は大きく分けて以下の二つのアプローチがある．

（１）　確率論的アプローチ
（２）　計算物理学的アプローチ

これら二つの手法は同じ研究対象をもつにも関わらず，その研究者の分野も異なり，その研究の価値観も異なることから，交差することなく並行して発展してきた．両者にはともに長所と短所があり，どちらも応用における問題点をもつ．本研究の研究方針としては，これらふたつのアプローチを統計学，機械学習理論への適用，および現実の大規模データ解析への応用を目的として統合する．具体的には，実際に大規模データ解析のための，モデル選択規準を構成する．

また，大規模ランダム行列理論を統計学に応用する際に，まず可換部分である代数統計学を調べるため，その世界的権威である Henry Wynn 教授（ロンドン・スクール・オブ・エコノミクス）と共同研究を行う．その過程で，可換代数統計学の手法を非可換理論に応用する方法を調べる．

また，代数的な構造を持つ統計モデルの例として，木構造を持つ心内辞書デンドログラムの解析手法を開発する．本研究は折田充教授（熊本大学）との共同研究である．

４．研究成果

（１）大規模ランダム行列理論を用いたカーネルマシンのモデル選択手法

カーネルマシンのモデル選択に必要な，カーネルグラム行列の Nyström 近似では次元削減を二段階で行い，それぞれに対して次元削減割合に対応するパラメータを設定する必要がある．ここで，近似が二段階であるため，計算量と近似誤差の間にはトレードオフの関係が成立し，次元削減パラメータの最適化の必要が生じる．そこで本研究では変分法等の計算物理学的な大規模ランダム行列解析の手法を用いて，分布に関する適当な仮定のもとでこのパラメータを最適化した．また，手書き文字データにおいてカーネルグラム行列の次元削減の最適化が有効に働くことを実験的に確かめた．また Nyström 法の近似誤差の PAC 学習的な上界を列の復元抽出の場合，列の非復元抽出の場合に分けて証明し，結果として一致性も証明した．さらに発展させ，Sparse greedy approximation や Incomplete Cholesky decomposition などの他のカーネルグラム行列の近似手法にも適用できることを示し，学会において紹介した．

（２）計算機代数の漸近推定理論への応用

大規模ランダム行列理論の基礎とするため，まず可換部分である代数統計学を調べるため，ロンドン・スクール・オブ・エコノミクスの Henry Wynn 氏と共同研究を行った．主に可換代数学を用いる代数統計学を情報幾何学の問題の代数化，推定量の有効性の条件を与える微分幾何学的特徴量の代数的計算手法を提案した．次に，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単純な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，数値実験によって計算量の本質的な削減を確認できた．

（３）木構造を持つデータの新しい解析手法の提案

英語母語話者と日本人話者の単語心内辞書の違いを比較，検定する手法についても，並べ替え検定を用いる手法を提案した．また，単体的扇状の幾何学や代数的性質を用いた解析手法を提案し，それを用いて並べ替え検定の精度を評価，さらには新しい並べ替え検定手法を提案した．これは，心内辞書のみならず，デンドログラムをはじめ木構造を持つ一般のデータの解析手法として新しいものであり，今後の応用が期待される．

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６件）

Kobayashi, K. and Komaki, F. (2008), Bayesian shrinkage prediction for the regression problem, Journal of Multivariate Analysis, 99 (9), pp. 1888-1905.

折田充，小林景（2011），心内辞書内の意味的クラスタリング構造―L1 と L2 の違いの指標となり得る語類の特定―，熊本大学社会文化研究，第９号，pp. 19-37.

Orita, M. and Kobayashi, K. (2011), Effects of Intra-Lexical Features on the Completion Time of Sorting Tasks, International Journal of Social and Cultural Studies, Vol. 4, pp. 1-23.

小林景（2011），DeRobertis 分離度による全変動距離の上界，統計数理，59-2，pp. 321-328.

折田充，小林景(2011)，心内辞書内の意味的クラスタリング－高頻度英語形容詞における母語話者と第二言語話者の相違，九州英語教育学会紀要，Vol. 39，pp. 1-11.

折田充，小林景(2012)，母語話者と第二言語話者の心内辞書 －類語の混在する単語群における意味的クラスタリング構造－@@熊本大学社会文化研究，Vol. 10、pp. 17-30.

〔学会発表〕（計２１件）

Kobayashi, K.： A Bayesian prediction for the Normal distributions with changeable covariances, Joint Meeting of ISI, ISM and ISSAS, Taipei, 2008.06.20.

Kobayashi, K. and Komaki, F.： Minimaxity of Stein-type Bayesian prediction for normal regression problem, 7th World Congress in Probability and Statistics, Singapore, 2008.07.17.

小林景，大規模行列固有値を用いた Nystrom 近似法の改良，統計関連学会連合大会,慶応義塾大学，２００８年９月９日.

Kobayashi,K.： Shrinkage Bayesian prediction and its application to regression problems, Statistics Seminar, Queen Mary Univ. of London, 2009.03.04

小林景，折田充，日本人と英語母語話者との心内辞書構造の相違の統計的解析，統計関連学会連合大会，同志社大学，京田辺，2009 年 9 月 9 日

Orita, M. and Kobayashi, K.：Predictors of L1 and L2 differences in lexical organisation, The 6th Vocabulary Acquisition Research Group Conference, Tokyo, 2009.12.05.

Orita, M. and Kobayashi, K.： Effects of intra-lexical Features on the completion time of sorting tasks, 20th Vocabulary Acquisition Research Group Network Conference, Gregynog, 2010.3.17-20.

Kobayashi, K. and Wynn, H., Using algebraic method in information geometry, Information Geometry and its Applications III, Leipzig, 2010.8.2-5.

Kobayashi, K. and Orita, M., Difference in mental lexicon between native and non-native English speakers, 73rd Annual Meeting of the Institute of Mathematical Statistics, Gothenburg, 2010.8.13.

折田充，小林景，心内辞書内のネットワーク構造―Sorting tasks を用いた母語話者と第二言語話者の違いの解明，第 54 回熊本大学英文学会，熊本大学，2010 年 11 月 20 日

折田充，小林景， 心内辞書内の意味的クラ

スタリング―母語話者と第二言語話者の相違，第 39 回九州英語教育学会，鹿児島大学，2010 年 12 月 12 日

Orita, M. and Kobayashi, K., Semantic Clustering of High Frequency Nouns in L1 and L2 Mental Lexicons, Learners and Networks Conference 2011, Swansea University, 2011.3.18.

Kobayashi, K. and Wynn, H., Algebraic computations for asymptotically efficient estimators via information geometry, Workshop on Geometric and Algebraic Statistics 3, University of Warwick, 2011.4.7.

折田充，小林景，心内辞書内の意味的クラスタリング構造（３）－高頻度英語動詞における英語母語話者と日本人英語話者の相違，第 37 回全国英語教育学会山形研究大会 ，山形大学，2011 年 8 月 21 日

小林景，計算機代数学を用いた漸近有効推定量の構成，統計関連学会連合大会，九州大学，2011 年 9 月 5 日

折田充，小林景，母語の心内辞書と第二言語の心内辞書，第 40 回九州英語教育学会，宮崎県立看護大学，2011 年 12 月 10 日

小林景，計算機代数を用いた代数的統計モデルの推定理論，研究集会「数理統計学と代数統計の新たな展開」，つくば国際会議場　エポカルつくば，2012 年 1 月 20 日

Kobayashi, K., Asymptotic efficiency of statistical estimators via algebraic computation and information geometry, ISM-ISI-ISSAS Joint Conference, 統計数理研究，2012 年 2 月 3 日

小林景，計算機代数を用いた情報幾何学と漸近的推定理論，情報論的学習理論と機械学習研究会（IBISML），統計数理研究所，2012 年 3 月 13 日

Orita,M. and Kobayashi,K., Semantically Equivalent Lexical Items between L1 and L2 Mental Lexicons, 22nd Vocabulary Acquisition Research Group Network Conference, Swansea Univ.,2012 年 3 月 16 日

６．研究組織

(1)研究代表者

小林　景（KOBAYASHI KEI）

統計数理研究所・数理・推論研究系・助教

研究者番号：９０４６５９２２