様式 C-7-2

# 自己評価報告書

平成２３年５月６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．研究計画の概要

本研究の目的は，大規模ランダム行列理論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の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．

２．研究の進捗状況

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，統計関連学会連合大会において紹介した．

また，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．現在までの達成度

②おおむね順調に進展している。

複数の国内，国際学会で発表し，また Nystrom 近似手法に関する論文作成中である．一方，途中で１年の長期海外出張を挟んだため，本来の研究計画から変更があった．しかし，出張中に（可換）代数統計学について本質的に新しい成果を出し，これを非可換な統計学に応用することが考えられるため，全体として進展はおおむね順調である．

４．今後の研究の推進方策
大規模ランダム行列理論を用いたNystrom近似手法についての論文を発表する．また，計算機代数を用いた推定理論やモデル選択，もしくはランダムな木構造や距離行列モデルに関する検定やモデル選択に，大規模ランダム行列理論を導入する．

５． 代表的な研究成果
（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）

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.

〔学会発表〕（計１３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.