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

平成　２９　年　　６　月　１４　日現在

機関番号：　８２６２６
研究種目：　若手研究(B)
研究期間：　2015～2016
課題番号：　１５Ｋ１６０５５

研究課題名（和文）ベイズ推論にもとづく全自動かつ高速なテンソル分解のモデル選択法

研究課題名（英文）Fully automatic and scalable Bayesian model selection method for tensor decomposition

研究代表者

林　浩平（Hayashi, Kohei）

国立研究開発法人産業技術総合研究所・人工知能研究センター・研究員

研究者番号：３０７０５０５９

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：テンソル分解によるデータ解析は様々な応用分野で需要が高まっているものの，正確な結果を得るためにはランクと呼ばれるパラメータを正しく設定する必要があり，これまでドメインエキスパートによる調整や長時間の試行錯誤が必要であった．本研究ではこの問題を解決すべく，簡単，高速，かつ信頼性が高いアルゴリズムを開発した．これにより，(1) ドメイン知識が不要で(2) 大規模データを高速に処理し(3) 性能が理論的に保証された全自動ランク選択が可能となった．将来的には，テンソル分解の応用範囲のさらなる拡大とそれにともなう新しい科学的発見や工学的応用が期待できる．

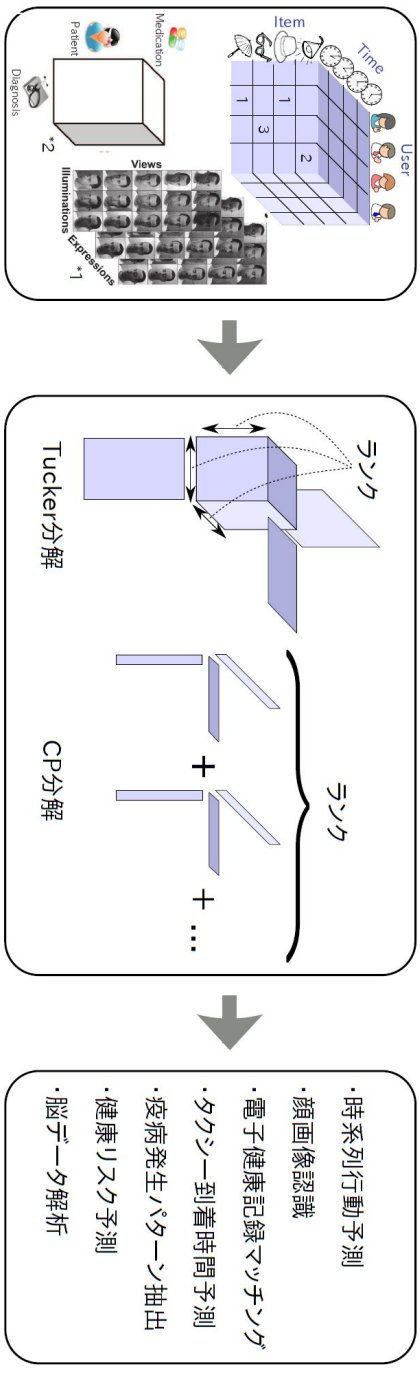
研究成果の概要（英文）：While data analysis with tensor decomposition is demanding in various application fields, in order to obtain accurate results, it is necessary to set a parameter called rank correctly, which has been adjusted by domain experts. In this study, we solve this problem by developing an algorithm that is simple, fast and highly reliable. The method has the following appealing points: (1) domain knowledge is unnecessary, (2) the algorithm is highly scalable, and (3) fully-automatic rank selection in which the performance is theoretically guaranteed is possible. Our result may yield further expansion of the application of tensor decomposition and new scientific discovery.

研究分野：　機械学習

キーワード：　テンソル分解　モデル選択　ベイズ学習　アルゴリズム

１．研究開始当初の背景

データがテンソルあるいは多次元配列としての構造を持つとき，これをデータテンソルと呼ぶ．例えば，「誰が」，「いつ」，「どの商品」を買ったかを記録した商品の購買履歴は，ユーザ× 時間× 商品の 3 項関係の集合であり，これは 3 階のテンソルで表現できる(図 1 上参照)．データテンソルとその解析は，その適用範囲の広さゆえ国際的に需要が高まっている．例えばデータマイニング分野のトップ会議である ACM SGKDD 2014 では 6 件もの発表がみられ，その応用もタクシー到着時間の予測や人の行動予測，またヘルスケア分野など多岐に渡る．また脳機能の解明といった科学分野でも利用が広がっている(図１下)．



*1 引用元: Morup, Applications of Tensor Decomposition in Data Mining and Machine Learning, NIPS workshop TKML2010.
*2 引用元: Ho et al., Limestone: High-throughput candidate phenotype generation via tensor factorization, J Biomed Inform, to appear

図１

データテンソルの解析手法としてはテンソル分解が広く使われている．テンソル分解はデータテンソルに内在する特徴的なパターンを低ランクテンソルとして抽出する．このパターンから，データの高度な解釈(例：購買履歴における若者ユーザ群の特定や彼らが好む商品カテゴリの抽出) や未知の値の予測(例：ユーザにとってある未知の商品をどれくらい好みそうか予測し，買ってくれそうなものを推薦) が可能となる．低ランクテンソルの構成法には様々なバリエーションが存在するが，中でも正準多項（Canonical Polyadic, CP） 分解とタッカー（Tucker）分解が基本的な方法として知られている（図１中央）．

定式化の上ではテンソル分解はランク–すなわちパターンの数–を既知として解くが，実用上はランクの選択が重要な課題となる．ランクを過大に設定するとデータへの過適合による予測精度の低下やパターン数増加による解釈性の低下を招くが，同時に過小なランクではデータを十分に表現しきれず，同様の性能劣化を招く(図２参照)．現状ではエキスパートや試行錯誤による調整が主に行われているが，人手と時間を必要とするため大規模データに適用できない，また結果に客観性がないといった問題点を抱えていた．

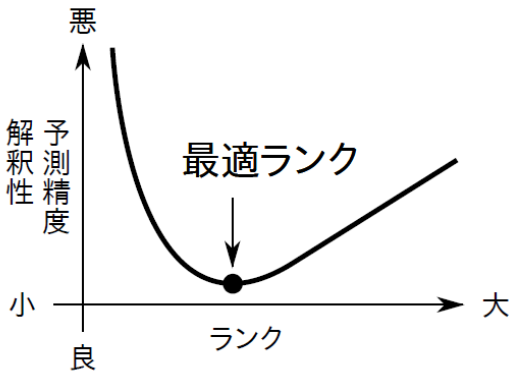


図２

一般にランク選択のような問題はベイズ統計学ではモデル選択と呼ばれ，周辺尤度関数やその最適解周りでの近似であるベイズ情報量規準(BIC) によって決定される．しかし，周辺尤度関数は基本的に計算困難であり，またテンソル分解は特異モデルと呼ばれる最適解が一意でないクラスに属すため，BIC の近似は成り立たない．

２．研究の目的

テンソル分解におけるランク選択の問題を解決するためのモデル選択法を提案する．

３．研究の方法

ランク選択の問題を解く方法として，因子化漸近ベイズ推論 (Factorized Asymptotic

Bayesian, FAB)法とよばれる手法を用いる．これはベイズ推論の一種であるが BIC のような近似に関する問題をうまく回避しており，テンソル分解にも適用可能であることが示唆されていた．

４．研究成果

テンソル分解の特殊ケースである主成分分析（PCA）に対して FAB 法を導出し，理論的にも実験的にも最適なランクが選択できることを示した[9]．またランクが与えられた元でテンソル分解を高速に計算するアルゴリズムを提案した[5]．さらに与えられたテンソルが画像のように近接する要素が似ている，すなわちなめらかな構造を持つときに適した新しいテンソル分解法およびその理論解析を行った[3]．

しかしながら，一般のテンソル分解への FAB 法の導出には理論的な困難があることを発見し，そのため方向転換を行った．

具体的にはベイズ的手法にこだわらず，他の指標での最適ランク探索の問題を考えた．結果，テンソル分解の訓練誤差を最小化する問題を考えた．訓練誤差は分解の過剰適合を回避できないためランク選択の基準としてはややナイーブなものであるが，最初の一歩としては問題ないものと考えた．

この結果，テンソル分解の訓練誤差を，入力テンソルのサイズによらず定数時間で最小化するアルゴリズムを開発した．この手法の開発のためにベースとして用いたのは二次形式の定数時間最小化[2]である．この技術は，任意の二次形式でかける方程式を入力サイズによらずに定数時間で最小化することができる．コアとなるアイデアは値が一様にばらつく行列で成り立つある種の正則性であり，それを用いることでたかだか定数個の行列の要素のみから解を高い確率で近似できる．このアイデアをテンソル分解に拡張し，同様の近似保障が成り立つアルゴリズムを提案した．実データを用いた実験結果から提案アルゴリズムは分解誤差を非常に高速かつ高精度で近似できることを確認した．この研究成果は国際会議に投稿済みである．

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　９件）

以下はすべて査読あり

[1] Yohei Kondo, Kohei Hayashi, Shin-ichi Maeda. Sparse Bayesian linear regression with latent masking variables. Neurocomputing, 2017 (In press)

[2] Kohei Hayashi, Yuichi Yoshida. Minimizing Quadratic Functions in Constant Time. In proceedings of NIPS, 2016 (pp. 2217--2225)

[3] Masaaki Imaizumi, Kohei Hayashi. Doubly Decomposing Nonparametric Tensor Regression. In proceedings of ICML, 2016 (pp. 727-736)

[4] Takuya Konishi, Tomoharu Iwata, Kohei Hayashi, Ken-ichi Kawarabayashi. Identifying Key Observers to Find Popular Information in Advance. In proceedings of IJCAI, 2016 (pp. 3761-3767)

[5] Takanori Maehara, Kohei Hayashi, Ken-ichi Kawarabayashi. Expected Tensor Decomposition with Stochastic Gradient Descent. In proceedings of AAAI, 2016 (pp. 1919-1925)

[6] Takuya Konishi, Takuya Ohwa, Sumio Fujita, Kazushi Ikeda, Kohei Hayashi. Extracting Search Query Patterns via the Pairwise Coupled Topic Model. In proceedings of WSDM, 2016 (pp. 655-664)

[7] Yohei Kondo, Kohei Hayashi, Shin-ichi Maeda. Bayesian Masking: Sparse Bayesian Estimation with Weaker Shrinkage Bias. In proceedings of ACML, 2015 (pp. 49-64)

[8] Kohei Hayashi, Takanori Maehara, Masashi Toyoda, Ken-ichi Kawarabayashi. Real-time Top-R Topic Detection on Twitter with Topic Hijack Filtering. In proceedings of KDD, 2015 (pp. 417-426)

[9] Kohei Hayashi, Shin-ichi Maeda, Ryohei Fujimaki. Rebuilding Factorized Information Criterion: Asymptotically Accurate Marginal Likelihood. In proceedings of ICML, 2015 (pp. 1358-1366)

〔学会発表〕（計　０件）

〔図書〕（計　１件）

石黒 勝彦，林 浩平．関係データ学習 (機械学習プロフェッショナルシリーズ)．講談社，2016．192

〔産業財産権〕

○出願状況（計　０件）

○取得状況（計　０件）

〔その他〕

ホームページ等

https://sites.google.com/site/koheihaya

shi84/

６．研究組織

(1)研究代表者

林　浩平（Kohei Hayashi）
産業技術総合研究所・人工知能研究センター・研究員
研究者番号：30705059