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　　１　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３００１３

研究課題名（和文）少数の観測信号にも適用可能な頑健性の高い独立成分分析手法の開発と応用

研究課題名（英文）Research on robust independent component analysis applicable to only a few observed signals

研究代表者

松田　源立（MATSUDA, Yoshitatsu）

東京大学・大学院総合文化研究科・学術研究員

研究者番号：４０４３３７００

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：独立成分分析（Independent Component Analysis, ICA）は、混合された信号のみから源信号を推定する手法である。ICAは機械学習や信号処理の分野において広く使われており、重なった画像の分離、音源推定、画像からの特徴抽出等の様々な応用に利用されている。本研究では、ICAを改良し、推定された源信号の一意な順序付けを可能とした。更に、ガウシアンノイズの分離を可能とし、これにより、ノイズを取り除いた源信号の真の個数の推定を行うことができるようになった。併せて、ICAを含む機械学習的な手法を、ウェブデータ分析、自然言語処理、教育工学の応用分野に適用した。

研究成果の概要（英文）：Independent Component Analysis (ICA) is a method estimating the sources from only the mixed signals. ICA is widely used in machine learning and signal processing and is known to be useful in many applications such as image separation, sound separation, and feature extraction from images. In this research, we improved ICA and enabled it to order the sources uniquely. Moreover, we enabled ICA to separate the Gaussian noises so that we can estimate the number of the non-Gaussian sources. In addition, we applied some machine learning methods including ICA to the applications of web data analysis, natural language processing, and educational engineering.

研究分野：　機械学習

キーワード：　独立成分分析　機械学習　信号処理　統計科学　ウェブデータ解析　自然言語処理　教育工学

**１．研究開始当初の背景**

(1) ブラインド信号源分離(Blind Source Separation, 以下 BSS) は、信号処理における重要な問題の一つである。線形の場合、BSS のモデルは以下の式で定式化される。

$$X = AS$$

ここで、X は観測信号、S は源信号、A は混合行列である。X のみが与えられている状態で、未知の A を推定するのが BSS の課題である(A が推定されれば S は決定される)。観測信号 X の与え方により、BSS は、音源分離や画像分離等の幅広い応用事例に適用可能である。既知のサンプル X の成分数よりも未知のパラメータ A および S の成分数の方が大きいため、BSS は不良設定問題であり、源信号 S や混合行列 A に対して何らかの仮定を置かなければ解決不能である。独立成分分析(Independent Component Analysis, 以下 ICA) は、広く使われている BSS の解法の一つである。ICA の主要な仮定は以下の二つである。① 源信号 S は互いに独立である。② 源信号は任意の非ガウス分布に従って発生する。これらの仮定をもとに、尖度等の高次統計量を利用することによって、ICA は混合行列 A を推定する。BSS には様々な解法が存在するが、源信号の性質について、音波に限定するなどの、強い仮定をおくものが多い。一方、ICA の仮定は比較的弱いため、BSS の多くの事例で一般的に利用可能であり、これが大きな優位性である。実際、上述の音源分離や画像分離に加えて、脳波等の生体信号分析、気象衛星観測データ等の大規模データ分析、あるいは機械学習の様々な応用分野において ICA が有効であることが知られている。

(2) ICA を含む様々な機械学習的な手法は広い応用分野に適用されつつある。例えば、ウェブデータ分析、自然言語処理、教育工学等が具体的な分野として挙げられ、知識発見、精度向上等の成果が上がりつつある。

**２．研究の目的**

(1) 本研究の第一の目的は、ICA の頑健性を向上させて、様々な源信号に対して柔軟に適応可能な手法を構築することである。従来の ICA では、源信号に関して近似的に事前分布を与える方法が主流であるが、源信号の分布が多様な場合や、多数のノイズが混合する場合には頑健性が低い。事前分布に関して制約の弱い手法もいくつか存在するが、計算量の増加等の欠点がある。本研究では、一般的な弱い仮定のみ利用しつつ、計算量も従来法並みに効率的となる ICA 手法を提案することを目的とする。更に、開発した手法を、画像分離等の応用事例に適用することも目的とする。

(2) 本研究の第二の目的は、ウェブデータ分析、自然言語処理、教育工学の応用分野において、ICA を含む様々な機械学習的な手法を活用し、精度向上や知識発見等を実現することである。

**３．研究の方法**

(1) ICA の目的関数の導出において、源信号の分布そのものを何らかの確率分布で近似するのではなく、源信号を二次多項式の特徴空間に射影し、その特徴空間内でのガウス近似を行った。そしてガウス分布の尤度を目的関数として利用した。これにより、弱い仮定で多様な分布に適応可能かつ単純で最適化しやすい目的関数を導出した。更に、その目的関数の収束性、局所最適性、大域最適性に関する数学的な証明を与えた。また、特徴空間上で、Fisher 情報量等の統計指標も導出した。それらの知見に基づき、新たな ICA 手法を提案した。

(2) ウェブデータ分析に関しては、ガウス分布近似によりユーザの時空間分布を推定し、推定された分布に ICA やクラスタリング等の機械学習手法を適用することで、新たな知見の発見やユーザの挙動予測等を行った。

(3) 自然言語処理に関しては、ニューラルネットワークモデルに基づく Skip-gram を文脈限定型に拡張することによる同義語獲得の改良等を行った。

(4) 教育工学に関しては、単純化教師有り Latent Dirichlet Allocation (以下 LDA)を提案し、シラバス及びカリキュラム分析を行い、新たな知見の発見やこれまでの分析の検証等を行った。

**４．研究成果**

(1) 二次多項式特徴空間におけるガウス分布近似に基づき ICA の新しい目的関数(Adaptive ICA Function, 以下 AIF)を提案した。更に、数学的な分析により、AIF の収束性、局所最適性、大域最適性の成立する条件を導出し、証明を与えた。更に、確率勾配法、交互最適化法、不動点法等の様々な最適化手法を適用し、実用的な AIF の最適化アルゴリズムを構築した。そして、画像分離の問題について AIF を適用して数値実験を行い、その有効性を実証した。

(2) AIF の数学的解析により、AIF の最大化により源信号を推定した場合、各源信号を、非ガウス性の大きさにより一律に順位付けることが可能であることを証明した。これにより、従来の ICA における大きな問題であった信号置換の不定性を解決した。すなわち、従来の ICA では抽出信号の並び替えは ICA 以外の方法で別途行う必要があったが、AIF により、直接順序の確定した状態で抽出できることを示した。

(3) AIF の特徴空間上でのガウス分布近似に対し、Fisher 情報量等の統計的技法を適用することにより、非ガウス性を判定する閾値を導出した。従来の ICA では、抽出信号がガウス分布に従うホワイトノイズか、非ガウス信号かを決定する理論的かつ汎用的な閾値は存在せず、非ガウス信号の個数を推定することは困難であった。今回、AIF に基づく提案手法により、非ガウス信号の個数を推定できるようになった。これにより、ホワイトノイズの影響が除去可能となった。また、非ガウス信号のみの推定を行うことで、計算時間の短縮も可能であることを示した。

(4) ウェブデータ分析において、二次元マップ上の GPS 情報を利用した SNS である Foursquare のログを収集し、そこから適切なハイパーパラメータを推定し、ユーザの時空間分布を導出した。導出された分布に、ICA や PCA、クラスタリングといった機械学習的手法を適用し、周期的あるいは突発的な様々な時空間パターンを発見した。

(5) 自然言語処理では、日本語に固有の文脈情報を取り込んだ単語の分散表現を利用することで、日本語の自動同義語獲得の精度向上に成功した。

(6) 教育工学においては、LDA をベースとした手法を標準的なコンピュータ科学教育カリキュラムに適用し、任意のシラバスのトピックを推定できる手法を構築した。さらに、その手法により、実際の大学のシラバスやカリキュラムを調査し、実用的な知見を獲得した。また、複数のカリキュラムを二次元上で可視化して比較できるシステムを開発した。

(7) その他応用においては、ウェブデータによるユーザの行動予測の精度向上、日英自動翻訳の精度向上、英語教育における標準辞書の解析等の研究を行った。

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 12 件）
すべて査読あり


1. 城光英彰, 松田源立, 山口和紀 (2017). 文脈限定 Skip-gram による同義語獲得, 自然言語処理, Vol. 24, No. 2, pp. 187-204. http://anlp.jp/abst/vol24/no2.html

2. Yoshitatsu Matsuda and Kazunori Yamaguchi (2017). Efficient Optimization of the Adaptive ICA Function with Estimating the Number of Non-Gaussian Sources. LVA/ICA2017, Grenoble, France, LNCS 10169, Springer-Verlag, 469-478. DOI: 10.1007/978-3-319-53547-0_44

3. Ken-ichiro Nishioka, Yoshitatsu Matsuda, and Kazunori Yamaguchi (2016). User Location Prediction by Diffusion-type Estimation using Location-based SNS Check-in Data. Proceedings of iThings, Chengdu, China. DOI:10.1109/iThings-GreenCom-CPSCom-SmartData.2016.30

4. Yoshitatsu Matsuda and Kazunori Yamaguchi (2016). Gram-Schmidt Orthonormalization to the Adaptive ICA Function for Fixing the Permutation Ambiguity. ICONIP2016, Kyoto, Japan, LNCS 9948, Springer-Verlag, 152-159. DOI: 10.1007/978-3-319-46672-9_18

5. Yoshitatsu Matsuda and Kazunori Yamaguchi (2016). Adaptive Objective Function of ICA by Gaussian Approximation in Second-order Polynomial Feature Space. Proceedings of IJCNN2016, Vancouver, Canada, IEEE, 2382-2389. DOI: 10.1109/IJCNN.2016.7727495

6. Yoshitatsu Matsuda, Kazunori Yamaguchi, and Ken-ichiro Nishioka (2015). Discovery of Regular and Irregular Spatio-Temporal Patterns from Location-Based SNS by Diffusion-Type Estimation. IEICE TRANSACTIONS on Information and Systems, E98-D (9), 1675-1682. DOI: 10.1587/transinf.2015EDP7095

7. Yoshitatsu Matsuda and Kazunori Yamaguchi (2015). Objective Function of ICA with Smooth Estimation of Kurtosis. ICONIP2015, Istanbul, Turkey, LNCS 9491, Springer-Verlag, 164-171. DOI: 10.1007/978-3-319-26555-1_19

8. Ken-ichiro Nishioka, Yoshitatsu Matsuda, and Kazunori Yamaguchi (2015). Discovery of Localized Spatio-temporal Patterns from Location-based SNS by Clustering Users. Proceedings of IJCNN2015, Killarney, Ireland, IEEE, 1-8. DOI: 10.1109/IJCNN.2015.7280597

9. Takayuki Sekiya, Yoshitatsu Matsuda, and Kazunori Yamaguchi (2015). Curriculum Analysis of CS Departments Based on CS2013 by Simplified, Supervised LDA. Proceedings of LAK2015, New York, USA, ACM, 330-339. DOI: 10.1145/2723576.2723594

10. Takayuki Sekiya, Yoshitatsu Matsuda, and Kazunori Yamaguchi (2014). Mapping

Analysis of CS2013 by Supervised LDA and Isomap. Proceedings of TALE2014, Wellington, New Zealand, IEEE, 33-40. DOI: 10.1109/TALE.2014.7062644

11. Yoshitatsu Matsuda, Kazunori Yamaguchi, and Ken-ichiro Nishioka (2014). Discovery of Spatio-temporal Patterns from Foursquare by Diffusion-type Estimation and ICA. ICANN2014, Hamburg, Germany, LNCS 8681, Springer-Verlag, 765-772. DOI: 10.1007/978-3-319-11179-7_96

12. Ikumi Horie, Kazunori Yamaguchi, Kenji Kashiwabara, and Yoshitatsu Matsuda (2014). Improvement of Difficulty Estimation of Personalized Teaching Material Generator by JACET. ITHET2014, York, England, IEEE, 1-8. DOI: 10.1109/ITHET.2014.7155695

〔学会発表〕（計 5 件）

1. 伊部早紀，松田源立，山口和紀（2017 年 3 月 15 日）. 格助詞と冠詞に着目した日英 tree-to-string 翻訳における単語アライメント表の改良，言語処理学会第 23 回年次大会，筑波大学，茨城県つくば市

2. 堀江郁美，松田源立（2017 年 3 月 12 日）. 語彙をベースとした初心者用英語多読学習支援システム改良のためのユーザ分析，研究報告コンピュータと教育（CE），2017-CE-139(19)，1-8，津田塾大学，東京都小平市

3. 城光英彰，松田源立，山口和紀（2016 年 3 月 9 日）. 文脈限定 Skip-gram による同義語獲得に関する研究，言語処理学会第 22 回年次大会，東北大学，宮城県仙台市

4. 松田源立，山口和紀（2015 年 11 月 27 日）. 2 次元特徴空間でのガウス分布近似を利用した適応的独立成分分析，IBIS2015，信学技報，115（323），285-292，つくば国際会議場，茨城県つくば市

5. 城光英彰，松田源立，山口和紀（2015 年 7 月 11 日）. 同義語判定問題を用いた語義ベクトルの評価の検討--Skip-gram モデルで獲得した語義ベクトルを例として--，インタラクティブ情報アクセスと可視化マイニング第 10 回研究会，東京大学駒場キャンパス，東京都目黒区

〔その他〕
ホームページ等
https://sites.google.com/site/yoshitatsumatsuda/publications

６．研究組織
(1)研究代表者
松田　源立　（MATSUDA, Yoshitatsu）
東京大学・大学院総合文化研究科・学術研究員
研究者番号：40433700