

令和 6 年 5 月 21 日現在

機関番号：38005

研究種目：基盤研究(B)（一般）

研究期間：2020～2023

課題番号：20H04243

研究課題名（和文）高次元小標本データのための非線形選択的推論アルゴリズムの研究開発

研究課題名（英文）Research and development of nonlinear Selective Inference for high-dimensional and small number of samples data

研究代表者

山田 誠（Yamada, Makoto）

沖縄科学技術大学院大学・機械学習とデータ科学ユニット・准教授

研究者番号：00581323

交付決定額（研究期間全体）：（直接経費） 13,600,000円

研究成果の概要（和文）：本研究では、非線形選択的推論の高次元拡張に取り組んだ。2020年度は、HSIC LassoとSplit法を用いた統計的仮説検定方法を開発し、実データで有効性を示した。2021年度は、Polyhedral LemmaとKnockoff filterを適用したHSICに基づく手法を提案し、それぞれICML 2021とAISTATS 2022で報告した。2022年度は、最適輸送法に基づく新たな高次元データ解析手法を提案し、AISTATS 2022とTMLRで発表した。そして最終年度は、Distance Covariance Lasso法を提案し、選択的推論の理論的性質を示した。

研究成果の学術的意義や社会的意義

本研究は、非線形選択的推論を高次元データに適用する新たな手法を提案し、統計的仮説検定の検出力向上を目指した。さらに、木構造最適輸送に基づくBarycenterの推定手法やWasserstein距離の学習方法など、新たな高次元データ解析手法を開発した。つまり、我々は非線形データの特徴選択とスクリーニングの理論的基盤を確立したと言える。さらに今後、機械学習やバイオインフォマティクス分野での実用的な応用され、提案法による新規の科学的発見が期待できる。これらの成果は、学術的意義に加え、社会的にも広範な影響を与えると考える。

研究成果の概要（英文）：In this research, we worked on a high-dimensional extension of nonlinear selective inference. In FY2020, we developed a statistical hypothesis testing method using HSIC Lasso and the Split method, and demonstrated its effectiveness on real data. In FY2021, we proposed a method based on HSIC with Polyhedral Lemma and Knockoff filter, which were reported in ICML 2021 and AISTATS 2022, respectively. In the fiscal year 2022, we proposed a new high-dimensional data analysis method based on the optimal transport method, which was presented at AISTATS 2022 and TMLR, respectively. In the final year, we proposed the Distance Covariance Lasso method and showed the theoretical properties of selective inference.

研究分野：機械学習

キーワード：選択的推論 特徴選択

#### 1. 研究開始当初の背景

バイオロジーや医療分野においては、特徴数 (例: 遺伝子数) が標本数 (例: 患者数) よりも大きい高次元小標本データの解析が重要テーマの一つである。このような分野において、研究代表者らが独自に研究を進めているカーネル法に基づいた非線形選択的推論の枠組みを提案している。しかし、この方法は高次元小標本データを扱うことができない問題があった。そこで、本研究課題では、選択的推論 (Selective Inference) と呼ばれる特徴選択と統計的仮説検定を組み合わせた方法の研究開発を実施した。

#### 2. 研究の目的

研究代表者らが独自に研究を進めているカーネル法に基づいた非線形選択的推論の枠組みを高次元小標本データを扱えるように拡張することを目指す。さらに、提案アルゴリズムを急性骨髄性白血病やアトピー性皮膚炎といった現実の問題に適用しその有効性を示すことを目的とする。

#### 3. 研究の方法

**高次元非線形選択的推論手法:** 研究の目的の一つであった非線形特徴選択手法 HSIC Lasso の選択的推論手法を提案した [1]。本研究ではまず初めに Polyhedral Lemma (Lee et al. Annals of Stats 2016) を拡張し、共分散行列を標本から推定する場合においても検定ができることを理論的に示した。そしてその結果を Hilbert-Schmidt Independence Criterion (HSIC) Lasso に対して適用することで目的を達成した。さらに提案法をバイオデータ解析に適用し、重要な特徴を提案法を用いて見出すことができることを確認した。

また、Polyhedral Lemma に基づいた方法に加え、Knockoff filter に基づいた非線形特徴選択手法の枠組みも提案した [5]。本研究では、カーネル法と Knockoff filter を組み合わせた世界初の研究と言える。具体的には、HSIC および conditional Maximum Mean Discrepancy (cMMD) と Knockoff filter 法を組み合わせることで非線形な特徴選択が可能となった。実用的な貢献に加え、カーネル法を用いて特徴スクリーニングする際に、どれだけの特徴を選択すれば真の特徴をきちんと検出できるかの理論も与えた。HSIC Lasso に加えて、Distance Covariance Lasso (DC-Lasso) 法を提案しその選択的推論方法を提案しその理論的性質を示した。本研究は我々のチームが進めている情報理論に基づいた選択的推論法の新しい方法であり、今後研究が進むことで実的に有効な方法ができる可能性がある。さらに、これまでの HSIC Lasso の研究成果をまとめた論文を日本統計学会誌にて報告し、国内の統計分野にも提案法を紹介した [12]

**その他高次元データ解析:** 高次元非線形選択的推論の研究に加え、研究する上で得られた知見をもとに木構造最適輸送に基づいた Barycenter の推定手法の提案 [6] および木構造最適輸送距離の学習方法を提案した [8]。

#### 4. 研究成果

[1] Tobias Freidling, Benjamin Poignard, Héctor Climente-González, Makoto Yamada: Post-selection inference with HSIC-Lasso. ICML 2021: 3439-3448

- [2] Yuki Takezawa, Ryoma Sato, Makoto Yamada: Supervised Tree-Wasserstein Distance. ICML 2021: 10086-10095
- [3] Yanbin Liu, Makoto Yamada, Yao-Hung Hubert Tsai, Tam Le, Ruslan Salakhutdinov, Yi Yang: LSMI-Sinkhorn: Semi-supervised Mutual Information Estimation with Optimal Transport. ECML/PKDD (1) 2021: 655-670
- [4] Ryoma Sato, Makoto Yamada, Hisashi Kashima: Random Features Strengthen Graph Neural Networks. SDM 2021: 333-341
- [5] Benjamin Poignard, Peter J. Naylor, Hector Climente-Gonzalez, Makoto Yamada: Feature screening with kernel knockoffs. AISTATS 2022: 1935-1974
- [6] Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, Makoto Yamada: Fixed Support Tree-Sliced Wasserstein Barycenter. AISTATS 2022: 1120-1137
- [7] Mathis Petrovich, Chao Liang, Ryoma Sato, Yanbin Liu, Yao-Hung Hubert Tsai, Linchao Zhu, Yi Yang, Ruslan Salakhutdinov, Makoto Yamada: Feature-Robust Optimal Transport for High-Dimensional Data. ECML/PKDD (5) 2022: 291-307
- [8] Makoto Yamada, Yuki Takezawa, Ryoma Sato, Han Bao, Zornitsa Kozareva, Sujith Ravi: Approximating 1-Wasserstein Distance with Trees. TMLR 2022
- [9] Ryuichiro Hataya, Makoto Yamada: Nyström Method for Accurate and Scalable Implicit Differentiation. AISTATS 2023: 4643-4654
- [10] Cléa Laouar, Yuki Takezawa, Makoto Yamada: Large-scale similarity search with Optimal Transport. EMNLP 2023: 11920-11930
- [11] Weijie Liu, Jiahao Xie, Chao Zhang, Makoto Yamada, Nenggan Zheng, Hui Qian: Robust Graph Dictionary Learning. ICLR 2023
- [12] Makoto Yamada, Benjamin Poignard, Hiroaki Yamada, Tobias Freidling, Hilbert-Schmidt Independence Criterion Lasso 法に基づいた高次元非線形特徴選択 日本統計学会和文誌 2022年4月

## 5. 主な発表論文等

〔雑誌論文〕 計9件（うち査読付論文 9件/うち国際共著 5件/うちオープンアクセス 9件）

1. 著者名 Makoto Yamada, Yuki Takezawa, Ryoma Sato, Han Bao, Zornitsa Kozareva, Sujith Ravi	4. 巻 0
2. 論文標題 Approximating 1-Wasserstein Distance with Trees.	5. 発行年 2022年
3. 雑誌名 Transactions on Machine Learning Research	6. 最初と最後の頁 0
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1. 著者名 Tobias Freidling, Benjamin Poignard, Hector Climente-Gonzalez, Makoto Yamada	4. 巻 -
2. 論文標題 Post-selection inference with HSIC-Lasso	5. 発行年 2021年
3. 雑誌名 International Conference on Machine Learning (ICML)	6. 最初と最後の頁 3439-3448
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1. 著者名 Yuki Takezawa, Ryoma Sato, Makoto Yamada	4. 巻 -
2. 論文標題 Supervised Tree-Wasserstein Distance	5. 発行年 2021年
3. 雑誌名 International Conference on Machine Learning (ICML)	6. 最初と最後の頁 10086-10095
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

1. 著者名 Yanbin Liu, Makoto Yamada, Yao-Hung Hubert Tsai, Tam Le, Ruslan Salakhutdinov, Yi Yang	4. 巻 -
2. 論文標題 LSMI-Sinkhorn: Semi-supervised Mutual Information Estimation with Optimal Transport	5. 発行年 2021年
3. 雑誌名 European Conference on Machine Learning (ECML)	6. 最初と最後の頁 655-670
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1. 著者名 Ryoma Sato, Makoto Yamada, Hisashi Kashima	4. 巻 -
2. 論文標題 Random Features Strengthen Graph Neural Networks	5. 発行年 2021年
3. 雑誌名 SIAM data mining (SDM)	6. 最初と最後の頁 333-341
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, Makoto Yamada	4. 巻 -
2. 論文標題 Fixed Support Tree-Sliced Wasserstein Barycenter	5. 発行年 2022年
3. 雑誌名 International Conference on Artificial Intelligence and Statistics (AISTATS)	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 該当する

1. 著者名 Benjamin Poignard, Peter Naylor, Hector Climente, Makoto Yamada	4. 巻 -
2. 論文標題 Feature Screening with Kernel Knockoff	5. 発行年 2022年
3. 雑誌名 International Conference on Artificial Intelligence and Statistics (AISTATS)	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Hiroaki Yamada, Makoto Yamada	4. 巻 -
2. 論文標題 Dynamic Sasvi: Strong Safe Screening for Norm-Regularized Least Squares	5. 発行年 2021年
3. 雑誌名 NeurIPS	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Tam Le, Nhat Ho, Makoto Yamada	4. 巻 -
2. 論文標題 Flow-based Alignment Approaches for Probability Measures in Different Spaces	5. 発行年 2021年
3. 雑誌名 International Conference on Artificial Intelligence and Statistics (AISTATS)	6. 最初と最後の頁 3934-3942
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 該当する

[学会発表] 計4件 (うち招待講演 0件 / うち国際学会 4件)

1. 発表者名 Benjamin Poignard, Peter J. Naylor, Hector Climente-Gonzalez, Makoto Yamada
2. 発表標題 Feature screening with kernel knockoffs
3. 学会等名 AISTATS 2022 (国際学会)
4. 発表年 2022年

1. 発表者名 Yuki Takezawa, Ryoma Sato, Zornitsa Kozareva, Sujith Ravi, Makoto Yamada
2. 発表標題 Fixed Support Tree-Sliced Wasserstein Barycenter
3. 学会等名 AISTATS 2022 (国際学会)
4. 発表年 2022年

1. 発表者名 Ryoma Sato, Makoto Yamada, Hisashi Kashima
2. 発表標題 Re-evaluating Word Mover's Distance
3. 学会等名 ICML 2022 (国際学会)
4. 発表年 2022年

1. 発表者名 Mathis Petrovich, Chao Liang, Ryoma Sato, Yanbin Liu, Yao-Hung Hubert Tsai, Linchao Zhu, Yi Yang, Ruslan Salakhutdinov, Makoto Yamada
2. 発表標題 Feature-Robust Optimal Transport for High-Dimensional Data.
3. 学会等名 ECML 2022 (国際学会)
4. 発表年 2022年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	下平 英寿 (Shimodaira Hidetoshi) (00290867)	京都大学・情報学研究科・教授  (14301)	
研究分担者	POIGNARD BENJAMIN (Poignard Benjamin) (40845252)	大阪大学・大学院経済学研究科・講師  (14401)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オーストラリア	University of Technology Sydney			
米国	Carnegie Mellon University	Facebook AI Research	University of Texas Austin	他1機関
英国	University of Cambridge			