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

令和　２ 年　 ６ 月　 ９ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３０５

研究課題名（和文）非線形性に基づく大規模因果推論原理・手法の研究

研究課題名（英文）Research on Principle and Methods of Large-scale Causal Infrerence Based on Nonlinearity

研究代表者

鷲尾　隆（Washio, Takashi）

大阪大学・産業科学研究所・教授

研究者番号：００１９２８１５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：ビッグデータから統計的因果推論により大規模な対象のメカニズムを把握するニーズが増大している。しかし、線形で非ガウスノイズを有する大規模系でしか、実用的解析原理・手法は知られていない。この課題に対し本提案研究では、（１）非線形系の多数の観測変数間の因果関係を高精度推定する新原理の確立、（２）新原理の拡張による高精度、高速な大規模系の統計的因果推論手法の開発、（３）大規模人工データによる基本性能検証、（４）実データによる実際的性能検証を行い、広範な大規模非線形系に関する実用的原理・手法を開発した。さらに成果を主要国際会議や主要国際ジャーナルで発表し、統計的因果推論のブレークスルー手法を広めた。

研究成果の学術的意義や社会的意義
開発した因果推論手法は対象系の非線形性に基づき、データの非線形回帰残差の大小のみで変数間の因果関係を推定できる。これは変数とノイズの独立性の推定に基づく従来の因果推論の枠組みと全く異なる。この新原理により、ノイズの性質や変数とノイズの独立性、交絡変数の有無に係わらず因果関係を一意かつ高速に推定できる。本提案原理は学術的に独創的かつ基礎的であり、本分野の世界的研究動向に新しい方向性を与えている。実世界の殆どの対象系は何等かの非線形性を有し、本開発手法は実用的にも広範な対象のメカニズム解析に適用可能である。今後、物理学、化学、生物学、各種産業の現象解析や設計にて重要な役割を担うと期待される。

研究成果の概要（英文）：There is an increasing need to understand the mechanism of large-scale systems by analyzing big data by statistical causal inference. However, its practical principles and methods have been established only for large-scale systems that are linear and have non-Gaussian noise. This research achieved (1) establishment of a new principle for estimating the causal relationship between many observation variables in a non-linear system with high accuracy, (2) development of statistical causal inference methods for large-scale systems by further extending the new principle, (3) basic performance verification using large-scale artificial data, and (4) practical performance verification using real-world data. We developed practical principles and methods for a wide range of large-scale nonlinear systems though these studies. Furthermore, we presented these results in major international conferences and international journals, and spread the breakthrough method of statistical causal reasoning.

研究分野：機械学習，データマイニング

キーワード：統計的因果推論　因果解析　機械学習　データマイニング　回帰解析　非線形性

１．研究開始当初の背景

IoT 時代を迎え、複雑な対象に関する膨大な観測変数のデータ収集が容易になり、対象メカニズム理解のため、多数の変数値の因果的依存関係を推定する機械学習研究が重要性を増している。

観測データから変数間の因果的依存関係を推定する方法論は統計的因果推論と呼ばれる。20 世紀には線形構造方程式やベイジアンネットなどについて、線形でガウスノイズを有する対象系の研究が多く行われ、限定的条件でしか因果関係を推定できないことが分かった[1]。そのため、これらの研究による因果推論手法では、大規模対象系を必ずしも十分な精度で解析できない[2]。

これに対し 2003 年から 2006 年にかけて、図１(a)に示すように対象系が線形でも非ガウスノイズを有するなら、変数とノイズの独立性を調べることで任意の変数間因果関係を一意に推定可能なことを、清水, 狩野, Hyvarinen, Hoyer 等が明らかにした[3,4]。さらに 2009 年から 2011 年には本研究提案者の研究グループを中心に清水, Hyvarinen, 河原, 鷲尾等が、線形系に関して変数とノイズの独立性を損なう因果的上流の共通変数（交絡変数）の影響を取り除く原理を確立し、大規模で線形な対象系を表す多数の変数間の因果関係を高精度推定可能な手法に拡張した[5,6]。これは線形・非ガウスな系に限定されるものの、大規模対象系に関して一意な統計的因果推論の解を理論的に保証し、さらに実データ適用可能な解析手法を提供する世界初の研究成果であった。

一方、欧米を中心に非線形系について、変数とノイズの独立性を仮定する因果推論研究が進められている[7,8]。しかし非線形系では、上流の変数やノイズの独立性が下流で失われ易く、交絡変数の影響の除去が困難であり、大規模対象系で一意な解を得る実用的手法は提案されていない。

これに対し、本研究提案者の研究グループは、非線形系では非線形回帰の残差が系内の因果の向きで異なることを発見し、理論的解析と実データ検証を進めた[9]。それによれば、ノイズがガウスか非ガウスか、変数とノイズが独立かなどに係わらず、変数間の因果の向き沿った非線形回帰残差が逆向きの残差より小さい。すなわち我々の理論では、図１(b)に示すように対象系が非線形ならば、ノイズの性質や独立性、交絡変数の有無などに依らず、非線形回帰残差から変数間の因果関係を一意に推定できる。さらに、このような非線形回帰は高速に計算可能である。

以上の着想により、上記欧米研究の困難に直面することなく、大規模対象系の多数の観測変数間の因果関係を、高精度、高速に推定可能な原理と手法を確立できると予想された。このように汎用でスケーラブルな手法の研究は世界的にも行われておらず、本提案研究の成果は統計的因果推論の学術研究、実用化研究において、大きなブレークスルーをもたらし、日本の当該研究分野での世界的地位を確固たるものにすると予想された。

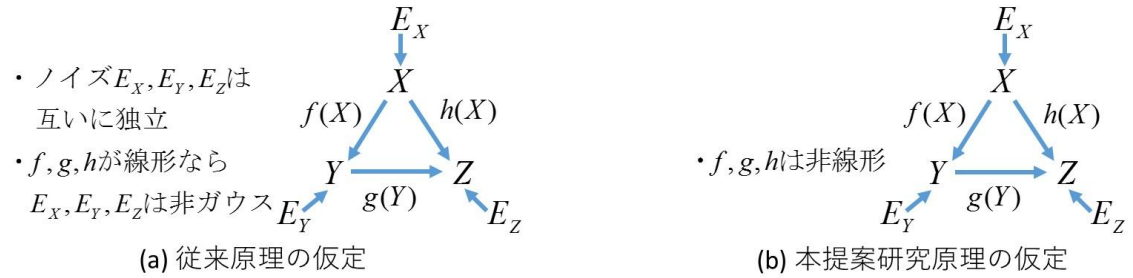

図１　従来原理と本提案研究原理における仮定の違い

２．研究の目的

以上の背景と着想に基づき、本提案研究では、（１）非線形な対象系の多数の観測変数間の因果関係を高精度推定する全く新しい統計的因果推論原理の確立、（２）新原理の拡張による高精度かつ高速な大規模対象系の統計的因果推論手法の開発、（３）手法の大規模人工データへの適用による基本性能の検証、（４）同じく実データ適用による実際的性能検証、（５）国際ジャーナルや主要国際会議での成果発表に並行した手法実装プログラムの公開と世界的普及を目的とした。

３．研究の方法

大規模対象系の因果推論ニーズの高まりと本提案の理論的糸口を背景とし、３年間で集中的成果創出を狙った。研究代表者鷲尾が全体を統括・主宰し、統計的因果推論、機械学習最適化・高速化のそれぞれ世界的研究者である同研究室の清水、河原が連携研究を行う。さらに博士後期課程学生 P.Blobaum などが計算実験・検証を支え、当該分野で世界的に最強の研究者集団で取り組んだ。平成２９年度は、因果推論の新原理確立を中心に、その大規模化アルゴリズム開発にも踏み込んだ。平成３０年度と最終年度で、大規模化アルゴリズムの人工及び実データによる計算実験・検証を通じた性能改良を繰り返し、実用的手法と実装プログラムを得た。そして、主要な国際ジャーナルや国際会議での成果発表を行い、世界的ブレークスルーとして普及を図ることを目指した。

以下、研究目的に掲げた研究項目（１）～（５）の内容を記す。

（１）非線形系の多数の観測変数間因果関係を高精度推定する新しい統計的因果推論原理の探求

変数間の因果の向きと非線形回帰残差の大小から、非線形対象系のノイズの性質や独立性、交絡変数の有無などに依らずに、因果関係を推定する原理を探求した。当初より理論的糸口が得られており[9]、それを因果推論手法として展開する理論的研究と、因果推論に適した非線形回帰の理論的原理を研究した。正確な理論的探求を効率的に進めるため、少数変数からなる小規模データに理論を適用したテスト解析を行い、その結果を確認しながら検討を進めた。

（２）新原理の拡張による高精度かつ高速な大規模対象系の統計的因果推論手法の開発

上記（１）で得られる理論的原理を、多数の変数間の因果関係推定を行うための原理へと拡張し、それを組み入れた大規模問題に適した効率的アルゴリズムの開発を行った。非線形回帰のアルゴリズムや多変数間因果関係の探索アルゴリズムなどを、本研究提案者等の過去の研究成果[5,6]や最適化理論、計算理論を踏まえて検討し開発した。

（３）手法の大規模人工データへの適用による基本性能の検証

上記（２）で得られた因果推論手法を大規模な人工データへ適用して推定精度や計算効率を評価し、その性質や問題点の洗い出しを行った。そして、明らかになった改善点を、（１）（２）の作業内容にフィードバックした。

（４）手法の実データ適用による実際的性能の検証

上記（２）（３）である程度完成した因果推論手法を、ベンチマーク用に公開されたレポジトリ実データへ適用し、さらなる性質や問題点の洗い出しを行った。そして、それまで明らかになった手法の性質に照らして一層の改善方法を検討し、（１）（２）の作業にフィードバックし研究のブラシュアップを行った。

（５）国際ジャーナルや主要国際会議での成果発表と並行した実装プログラム公開と世界的普及

本研究提案内容が実現すれば、統計的因果推論の学術・実用両面で大きなインパクトがあるため、以上（１）～（４）によって得られた研究成果を、機械学習や統計分野の主要な国際ジャーナルや国際会議で積極的に発表すること目指した。並行して、一定段階まで完成した手法・アルゴリズムの実装プログラムを GitHub など機械学習、統計分野で世界標準のプラットフォームで公開し、世界的な普及を図った。

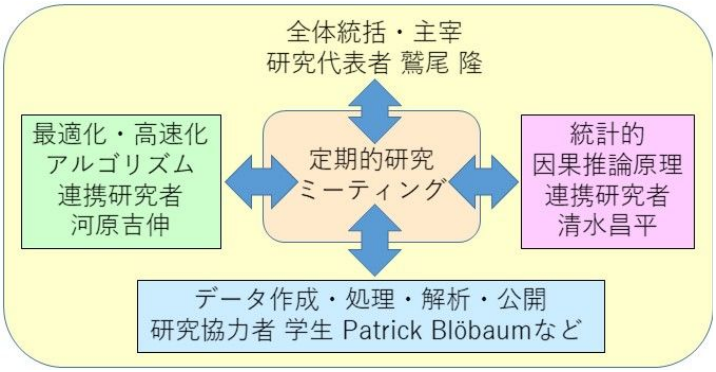

図２　研究体制と参加者の役割分担

本提案の研究では、統計的因果推論に関する理論的・実践的専門家と、それを大規模対象問題に適用する上で必要な最適化・高速化の理論・アルゴリズムの専門家、そしてこれらを十分に理解した上でデータ作成・処理・解析を行う人材、さらに全体を理解し統括する主宰者が必要である。幸いにも、本研究提案者の研究室は、機械学習に関する世界的に優れた実績を有し、これらすべての人材が揃っていた。そこで、図２に示す研究体制を取った。研究代表者の鷲尾が研究全体を統括・主宰し、統計的因果推論研究において世界的実績を持つ清水特任准教授と機械学習の最適化・高速化理論研究において同じく世界的実績を持つ河原准教授が理論構築の主力となった。そして、博士後期課程学生で統計的因果推論の手法開発・解析に取り組む P.Blobaum などが、データによる解析・検証・プログラム公開を担った。同一研究室所属のため常に密接な研究議論を行い、かつ全体作業内容や進捗の確認・調整のために定期的研究ミーティングも行った。世界的にも本提案研究テーマを遂行しうる質の高い人材が１つの研究室に集約しているところは殆どなく、少数かつ適任の精鋭で統計的因果推論研究のブレークスルーが実現できた。

４．研究成果

統計的因果推論の基本問題は、２つの変数 X と Y の統計的観測結果データが与えられた場合に、因果関係を何れの変数値かもう１つの変数値を決めているかという決定関係として捉え、原因変数と結果変数を同定することである。X と Y の関係は統計的であり、ノイズによる揺らぎを含むことを前提とする。

我々の提案原理は、X で Y を回帰した際の回帰誤差の MSE と Y で X を回帰した際の回帰誤差の MSE の単純な比較に基づく[10,11,12]。ここで、原因変数と結果変数をそれぞれ C,E∈{X、Y}と表す。結果変数 E の値は原因変数 C の値から以下の式で表せれる過程によって生成されると仮定する。ここで N はガウス、非ガウス何れでも構わないランダムなノイズである。

$$E=\phi(C)+N$$

X,Y の何れが C あるいは E であるかは予め不明である。ここでφは逆関数を有する厳密に単調増加関数であって 2 回微分可能であるとする。さらに、関数φと N の確率分布関数は独立に与えられ、 一般性を失うことなく C の値域は有限であるとし、N は期待値がゼロであるとする。

以上より、φは C から E を回帰するときの最小二乗誤差を最小化する関数であり、すなわちφは C がある値 c を取るときの E の値の条件付き期待値φ(c)= E[E|c]である。また、ψを E から C を回帰するときの最小二乗誤差を最小化する関数が存在し、すなわちψは E がある値 e を取るときの C の値の条件付き期待値ψ(e)=E[C|e]である。

以上の問題設定において、我々は以下の不等式が常に成立することを数学的に証明した。

$$E[(E-\phi(C))^2] < E[(C-\psi(E))^2]$$

これは特に C と E の値域がぞれぞれ[0,1]の間に規格化されている場合には、次のように書き直すことができる。

$$E[Var[E|C]] < E[Var[C|E]]$$

この不等式は C,E が線形関係である場合には等号となり成立しないが、任意の非線形関係である場合には必ず成立する。従って、どちらが原因変数 C か結果変数 E であるかが不明であるが関係が非線形であることが知られている 2 変数 X,Y のデータが与えられた時に、X から Y への回帰誤差と Y から X への回帰誤差を比較して、誤差が小さい回帰向きに従って何れが原因変数と結果変数であるかを決定すれば良い。

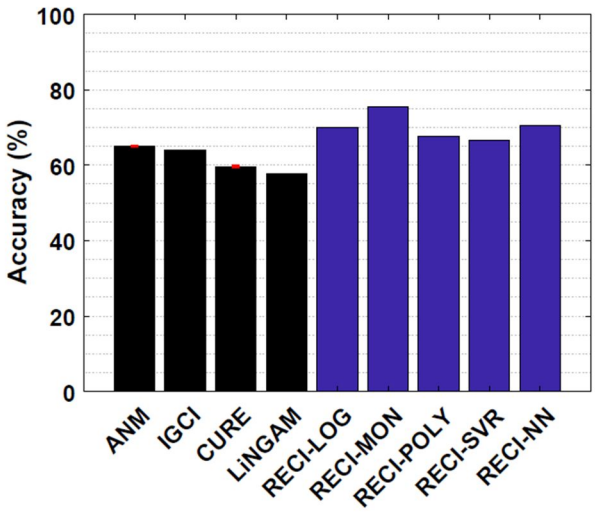

図３　原因結果変数ペア(CEP)ベンチマークデータセットによる従来手法と提案手法の精度比較

我々の主要国際会議発表文献[11]では以上の原理を提案した上で、人工的シミュレーションで作成した様々なデータと一部実世界データについて理論通りに因果推論が可能であることを示した。図３は、統計的因果推論研究の分野で標準的に用いられている原因結果変数ペア(CEP)ベンチマークデータセットに関して、横軸に異なる手法を取り縦軸に原因・結果変数同定の正答率を取って、棒グラフで比較した結果を示す。黒い棒の ANM, IGCI, CURE, LiNGAM は従来の統計的因果推論手法による結果であり、青い棒は提案手法による結果である。我々の提案手法は 5 種類の異なる回帰モデルを用いた結果を提示している。何れの回帰モデルを用いても、我々の手法が

従来手法よりも良い精度を示している。我々の主要ジャーナル論文文献[12]では、上記の理論に関するさらに厳密な理論解析と証明を与え、より広範な人工的シミュレーションデータと実世界データを用いて検証を行い、上述の結果を確認した。

特に文献[11,12]の国際会議とジャーナルは機械学習研究分野で多くの研究者に読まれる情報源であり、それらでの発表、採択により世界の統計的因果推論研究者にインパクトを与えることができた。

図３　原因結果変数ペア(CEP)ベンチマークデータセットによる従来手法と提案手法の精度比較

参考文献

[1] Pearl J., Causality: Models, Reasoning and Inference, 2nd Ed., Cambridge University Press, 2009

[2] Imoto S., Tamada Y., Araki H., Yasuda K., Print C.G., D. Charnock-Jones S., Sanders D., Savoie C.J., Tashiro K., Kuhara S. and Miyano S., Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles, Pacific Symposium on Biocomputing, Vol.11, pp.559-571, 2006

[3] Shimizu S., Hyvarinen A., Hoyer P.O. and Kano Y., Finding a causal ordering via independent component analysis, Computational Statistics & Data Analysis, Vol.50, No.11, pp.3278-3293, 2006

[4] Shimizu S., Hoyer P.O., Hyvarinen A. and Kerminen A., A linear non-Gaussian acyclic model for causal discovery, J. Machine Learning Research, Vol.7, pp.2003-2030, 2006

[5] Shimizu S., Hyvarinen A., Kawahara Y. and Washio T., A direct method for estimating a causal ordering in a linear non-Gaussian acyclic model, In Proc. 25th Conf. on Uncertainty in Artifficial Intelligence (UAI), pp.506-513, 2009

[6] Shimizu S., Inazumi T., Sogawa Y., Hyvarinen A., Kawahara Y., Washio T., Hoyer P.O. and Bollen K., DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. J. Machine Learning Research, Vo.12, pp.1225-1248, 2011

[7] Hoyer P.O., Janzing D., Mooij J., Peters J., Scholkopf B., Nonlinear causal discovery with additive noise models, In Proc. 21st Int. Conf. on Neural Information Processing Systems (NIPS), pp.689-696, 2008

[8] Zhang K. and Hyvarinen A., On the Identifiability of the Post-Nonlinear Causal Model, In Proc. 25th Conf. on Uncertainty in Artificial Intelligence (UAI), pp.647-655, 2009

[9] Blobaum P., Washio T. and Shimizu S., Error Asymmetry in Causal and Anticausal Regression, Behaviormetrika, Vol.44, No.2, pp 491-512, 2017

[10] Patrick Bloebaum, Takashi Washio, Shohei Shimizu, A Novel Principle for Causal Inference in Data with Small Error Variance, Proc. of European Symposium on Artificial Neural Networks, ISBN: 978-287587039-1, 347-352, 2017

[11] Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, Bernhard Schoelkopf, Cause-Effect Inference by Comparing Regression Errors, Proc. AISTATS2018:The 21st International Conference on Artificial Intelligence and Statistics, Paper No.298, 2018

[12] Patrick Blobaum, Dominik Janzing, Takashi Washio, Shohei Shimizu and Bernhard Scholkopf, Analysis of cause-effect inference by comparing regression errors, PeerJ Comput. Sci, Vol.5, e169, 2019

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　2件／うちオープンアクセス　2件）

１．著者名 Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, Bernhard Schoelkopf	４．巻 84
２．論文標題 Cause-Effect Inference by Comparing Regression Errors	５．発行年 2018年
３．雑誌名 Proc. of Machine Learning Research	６．最初と最後の頁 900-909
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 Patrick Blobaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, Bernhard Scholkopf	４．巻 5
２．論文標題 Analysis of cause-effect inference by comparing regression errors	５．発行年 2019年
３．雑誌名 PeerJ Computer Science	６．最初と最後の頁 e169
掲載論文のDOI（デジタルオブジェクト識別子） doi.org/10.7717/peerj-cs.169	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 Patrick Blobaum, Takashi Washio and Shohei Shimizu	４．巻 44
２．論文標題 Error Asymmetry in Causal and Anticausal Regression	５．発行年 2017年
３．雑誌名 Behaviormetrika	６．最初と最後の頁 491-512
掲載論文のDOI（デジタルオブジェクト識別子） DOI: 10.1007/s41237-017-0022-z	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Takashi Washio, Gaku Imamura and Genki Yoshikawa
２．発表標題 Machine Learning Independent of Population Distributions for Measurement
３．学会等名 DSAA2017: 4th IEEE International Conference on Data Science and Advanced Analytics（国際学会）
４．発表年 2017年

１．発表者名 Patrick Bloebaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, Bernhard Schoelkopf
２．発表標題 Cause-Effect Inference by Comparing Regression Errors
３．学会等名 , AISTATS2018:The 21st International Conference on Artificial Intelligence and Statistics, Paper No.298（国際学会）
４．発表年 2018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