氏　　名　原田 和治

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2320 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科 統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Statistical estimation for causal relationships under sparsity and contamination

論 文 審 査 委 員　主　査　二宮 嘉行
統計科学専攻 教授
藤澤 洋徳
統計科学専攻 教授
逸見 昌之
統計科学専攻 准教授
清水 昌平
滋賀大学 データサイエンス学部 教授
田栗 正隆
横浜市立大学 データサイエンス学部 教授

（様式３）

博士論文の要旨

氏　　名　原田　和治

論文題目　Statistical estimation for causal relationships under sparsity and contamination

There are two main goals in the quest for causality using statistics. The first goal is to infer the causal structure of the system of interest from data when the structure itself is partially or globally unknown. We assume that the system can be represented by a directed graph and formulate the problem as an estimation of the directed graph. In particular, recently, a number of models and estimation algorithms have been proposed that can identify the complete structure of the graph. The second goal, on the other hand, is to estimate the magnitude of the causal relationship between specific variables under the given causal structure. The framework for this case is called statistical causal inference. In particular, the causal effect of a variable treatment on the target variable has significant real-world implications in policy making and drug development, for example. For both causal discovery and causal inference, it is necessary to use statistical inference based on available data. There is no difference from ordinary statistical inference on this point. This means that various difficulties of the data, such as sparsity and outliers, affect the efficiency and accuracy of the estimation. Furthermore, the combination of causal and data difficulties sometimes evokes additional difficulties, so it is not sufficient to deal with these difficulties separately. We are interested in this type of problems. In this thesis, we discuss the sparsity in causal discovery and robustness to outliers in causal inference. Our study reveals that statistical methods for causality that deal with sparsity and outliers require nontrivial attentions, which is unique to causal estimation.

The first work is to deal with sparsity in statistical causal discovery. While there are several identifiable models for causal discovery, we focus on the linear non-Gaussian acyclic model (LiNGAM), which can be formulated as an independent component analysis (ICA) problem. ICA is well known in the field of signal processing. The linearity of LiNGAM enables an analyst to draw practical implications easier than other complicated nonlinear models. LiNGAM can also be seen as a linear structural equation, and its coefficient matrix has a sparse structure with at least half of its elements being zero because of acyclicity. Besides, it is natural to think that not all variable pairs have direct causal relationships, especially under high dimensional settings. This allows us to suppose the coefficient matrix of LiNGAM is much sparser. For LiNGAM, various estimation methods have been developed. However, the existing

methods are not efficient for some reasons: (i) the sparse structure is not always incorporated in causal order estimation, and (ii) the information of higher-order moments of the error terms is not used in parameter estimation. To address these issues, we propose a new estimation method for a linear DAG model with non-Gaussian noise. The proposed method is based on a single statistical criterion that includes the log-likelihood of independent component analysis (ICA) and two penalty terms. The two penalties are related to the sparsity and the prerequisite for consistency, respectively. This criterion enables us to leverage the sparse structure and the information of higher-order moments throughout the estimation. For stable and efficient optimization, we propose some devices, such as a modified natural gradient. Numerical experiments show that the proposed method outperforms the existing methods.

The second work is the estimation of causal effects when the target variable is contaminated with outliers. Estimators for causal quantities sometimes suffer from outliers. We investigate the outlier-resistant estimation of the average treatment effect (ATE) under challenging but realistic settings with contamination. We assume that the ratio of outliers is not necessarily small and that it can depend on covariates, namely, heterogeneous. We propose three types of estimators of the ATE, which combines the well-known inverse probability weighting (IPW)/doubly robust (DR) estimators with the density power weight. Under heterogeneous contamination, our methods can reduce the bias caused by outliers. In particular, under homogeneous contamination, our estimators are almost consistent with the true ATE. An influence-function-based analysis indicates that the adverse effect of outliers is negligible if the ratio of outliers is small even under heterogeneous contamination. We also derived the asymptotic properties of our estimators. We evaluated the performance of our estimators through Monte-Carlo simulations and real data analysis. The comparative methods, which estimate the median of the potential outcome, do not have enough outlier resistance. In the experiments, our methods outperformed the comparative methods.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　原田　和治

Title
論文題目　Statistical estimation for causal relationships under sparsity and contamination

[論文の概要]

提出された論文は，スパース構造を想定した場合の因果探索と，因果推論における平均処置効果のロバスト推定に関わる研究を扱っている．英文で書かれており，全５章と引用文献で計 106 頁からなる．

第 1 章は，本論文の序章である．統計科学を用いた因果関係の探求には大きく分けて 2 つある．1 つ目は対象となるシステムの因果構造をデータから同定することを目指す因果探索である．2 つ目はシステムの因果構造が与えられた下で因果関係の大きさを推定することである．本論文では特に次のような問題を考えている．1 つ目に関しては，構造因果モデルにおいてスパース構造を想定して因果構造を同定する問題を考えている．この問題は第 3 章で扱っている.2 つ目に関しては,逆確率重み付き推定や二重頑健推定において，平均処置効果をロバスト推定する問題を考えている．この問題は第 4 章で扱っている.

第 2 章は，第 3 章と第 4 章の準備を行っている．まずは因果で重要な交絡やバイアスに関して説明している.次に第 3 章に関連して,構造因果モデルと因果探索に関して紹介し，識別性をもつ構造因果モデルの一つである線形非ガウス非巡回モデル(LiNGAM)について詳しく説明している．特に，LiNGAM の同定には，独立成分分析の利用が重要であり，その関係についても解説されている．さらに第 4 章に関連して，因果モデルの一つである潜在アウトカムモデルや平均処置効果の推定問題を紹介し，逆確率重み付き推定や二重頑健推定に基づいた平均処置効果推定に関して説明している.

第 3 章は，LiNGAM においてスパース構造を想定し，因果構造を同定する問題を取り扱っている． LiNGAM は線形構造方程式に基づいており，非巡回性から係数行列は少なくとも半分の要素がゼロになるというスパース構造が内在している．また，高次元では，直接的な因果関係はスパースになると想定することもできる．これらを考慮して係数行列はスパースであると想定し，独立成分分析で使われる尤度にスパース罰則を導入した罰則付き尤度に基づく手法を提案している．過去に類似手法は提案されているが，重要な一致性がきちんと考慮されていなかった.そこで,一致性を復元する目的の罰則も導入している．しかし，そのタイプの罰則とスパース罰則を同時に直接扱うと，妥当なスパース構造が得られにくいことが知られている．その問題を克服する工夫を導入した罰則付き尤度を構築し，交互方向乗数法や修正自然勾配法などを使って効率的な最適化法を提案している．数値実験により，高次元スパース構造の状態では，提案手法が既存手法よりも優れていることを確認している.

第 4 章では，逆確率重み付け推定や二重頑健推定において，反応変数が外れ値で汚染さ

れている場合の平均処置効果のロバスト推定が議論されている．外れ値に関しては，外れ値の比率は必ずしも小さくなく，共変量に依存する不均一汚染の場合も議論している．これまでのロバスト推定は中央値を使ったアドホックな手法だったが，べき密度を利用することで効果的なロバスト推定が可能になることを提示している．提案手法は，適当な仮定の下で，均一汚染の下では外れ値によるバイアスをほぼもたず，不均一汚染の下でも影響関数に基づいた分析により外れ値にロバストなことが示されている．さらに，モンテカルロ・シミュレーションと実データ解析により，提案手法の性能が評価されている．中央値を利用した手法は外れ値への耐性が十分ではないが，提案手法は安定的な性能を与えていることを確認している．

[論文の評価]

LiNGAM にスパース構造を想定することで効果的な推定手法を提案している．先行研究では上手く行っていなかった一致性とスパース性の両立という問題をうまく克服して効果的な手法を構築したことは評価できる．因果推論において，平均処置効果推定という基本的な問題でも外れ値への対処は簡単ではなく，それまでのアドホックな中央値に基づく方法を優越し，べき密度を利用した手法を提案して性質を十分に調べた上で効果的な手法を提示したことは評価できる．

[その他]

第 3 章の内容は査読付国際学術雑誌 Neurocomputing に採択されている．第 4 章の内容は投稿後に査読報告書が来て改稿中の状態にある．

以上をもち，審査委員会は，本論文が博士（統計科学）の学位の授与に値すると判断した．