Statistical estimation for causal relationships under sparsity and contamination

Kazuharu Harada

Thesis for the degree of doctor of philosophy
Submitted to the Department of Statistical Science,
School of Multidisciplinary Sciences,
The Graduate University for Advanced Studies, SOKENDAI

March 2022

Abstract

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.

Acknowledgement

This thesis consists of the research achievements obtained while the author belongs to the Ph.D. course in the Department of Statistical Science, School of Multidisciplinary Sciences, the Graduate University for Advanced Studies (SOKENDAI). I would like to express my sincere gratitude to Prof. Hironori Fujisawa of the Institute of Statistical Mathematics and the Department of Statistical Science for giving me the opportunity to conduct this research as an advisor and co-author. I often consulted with him about various matters not directly related to my research, such as how to write applications and financial matters, and I have been able to complete my Ph.D. course thanks to his great help. I would like to express my sincere gratitude to Prof. Yoshiyuki Ninomiya and Associate Prof. Masayuki Henmi, the Institute of Statistical Mathematics, and Prof. Shohei Shimizu, Shiga University, and Prof. Masataka Taguri, Yokohama City University, for their advice as members of the doctoral degree review committee, and for their help in various occasions, including discussions about individual research projects and group readings of relevant materials. I would also like to express my gratitude to the members and alumni of Fujisawa's laboratory and the members of the Institute of Statistical Mathematics for supporting my research life in various ways, from group reading sessions and daily discussions to private exchanges. I have been able to complete my doctoral thesis thanks to the support of many people. This work is partly supported by Grant-in-Aid for JSPS Fellows [Grant No. 21J10457]. Chapter 3 of this thesis is a revised version of a paper published in Nerocomputing. In accordance with the copyright policy of the publisher, Elsevier, the acknowledgment of the published version is quoted below:

> The authors are grateful to the associate editor and anonymous reviewers for their helpful comments. Kazuharu Harada was supported by Grant-in-Aid for JSPS Fellows [Grant No. 21J10457]. Hironori Fujisawa was supported in part by JSPS KAKENHI [Grant No. 17K00065].

Contents

1 Introduction **6**

1.1 Big Picture and Motivation 6

1.1.1 Fallacy of Causation 6

1.1.2 Frameworks for Causal Inference 8

1.1.3 Statistical Estimation of Causal Quantities 9

1.2 Our Contributions 10

1.3 Outline 11

2 Preparation **13**

2.1 Causal Discovery 13

2.1.1 Structural Causal Model and Causal Discovery 13

2.1.2 Linear Non-Gaussian Acyclic Model (LiNGAM) 15

2.2 Causal Inference 19

2.2.1 Potential Outcome Model 19

2.2.2 Propensity Score-Based Estimators for ATE 21

3 Sparse Estimation of LiNGAM **24**

3.1 Background 24

3.2 Proposed Method 26

3.2.1 ICA and Consistent Estimation 26

3.2.2 ICA with Sparse Penalty for Causal Discovery 28

3.3 Algorithm 29

3.3.1 How to Obtain Parameter Estimates 30

3.3.2 Tuning Parameter Selection 32

3.3.3 Post-Processing 33

3.4 Experiments 33

3.4.1 Comparison of the Methods on Synthetic Data 35

3.4.2 Scalability of the proposed method 39

3.4.3 Real Data 39

3.5 Conclusion 42

3.6 Additional Sources for sICA-LiNGAM 43

3.6.1 Pattern for Tuning Parameters of GOLEM 43

3.6.2 Definitions of Evaluation Measures 43

3.6.3 Full Results of the Experiment 5.1 44

3.6.4 Sensitivity Study on Cutoff Threshold 46

3.6.5 AR(1) Recovery at the Other Sites 47

4 Outlier-resistant Estimation of ATE **49**
4.1 Background . . . 49
4.2 Outlier-resistant Estimation . . . 50
4.2.1 Non-causal Setting . . . 50
4.2.2 Causal Setting . . . 50
4.3 Proposed Methods . . . 51
4.3.1 Assumptions on Outliers . . . 51
4.3.2 DP-IPW Estimator . . . 52
4.3.3 DP-DR Estimator . . . 53
4.3.4 Summary . . . 56
4.4 Influence Function-based Analysis . . . 56
4.5 Asymptotic Properties . . . 59
4.6 Algorithm . . . 60
4.6.1 General Form . . . 60
4.6.2 Gaussian Weight . . . 61
4.7 Monte-Carlo Simulations . . . 62
4.7.1 Comparative Methods . . . 62
4.7.2 Simulation Model . . . 63
4.7.3 Results . . . 64
4.8 Real Data Analysis . . . 71
4.9 Additional Sources for Outlier-resistant Estimator for ATE . . . 74
4.9.1 Proof of Theorem 4.1 . . . 74
4.9.2 Proof of Theorem 4.2 . . . 74
4.9.3 Proof of Theorem 4.3 . . . 75
4.9.4 Proof of Theorem 4.4 . . . 76
4.9.5 Derivation of Influence functions in Section 4.4 . . . 77
4.9.6 Influence Functions Under Homogeneous Contamination . . . 78
4.9.7 Regularity Conditions for Theorem 4.5 . . . 80
4.9.8 Proof of Theorem 4.6 . . . 81
4.9.9 Further Discussion on Asymptotic Variance . . . 81
4.9.10 Remaining Results of Monte-Carlo Simulation . . . 85

5 Conclusion **97**

1 Introduction

Quest for causality is a fundamental concern of the natural and social sciences. In addition to academic fields, causal understanding is a key factor to the success of businesses; marketers sometimes use randomized controlled trials for effective advertisement, and manufacturers use statistical analysis to identify the causes of product anomalies. However, it is not easy to infer causality, and naive inference based on observable facts may lead to an erroneous conclusion about causality. In this chapter, we introduce a framework for inferring causality through some examples and discuss its relationship to statistical inference. Then, we introduce the main concern of this thesis, the issue of causal models when dealing with various data difficulties (e.g., high dimensionality, outliers, missing, etc.).

1.1 Big Picture and Motivation

1.1.1 Fallacy of Causation

Statistical methods usually involve either simply aggregating the data or learning a model to fit the data. The model is evaluated based on its goodness of fit to the data or its predictive performance under the uniformity of nature. While these approaches are very useful, it can be seriously biased if one wants to know causality. Simpson's paradox [66, 10] is a well-known example that illustrates naive group comparison can lead to an erroneous conclusion about causality in the presence of confounding.

Example 1: Simpson's paradox Table 1.1 shows an artificial example of Simpson's paradox. Looking at the upper table, it appears that the new drug has no effect compared to the existing drug. However, looking at the lower table, it suggests an opposite conclusion about the effect of the new drug. This reversal of conclusion can be explained by the confounding of baseline disease severity. In other words, the new drug tends to be used more often in the severe group, but at the same time, the severe group tends to have a higher mortality rate than the mild group, regardless of which drug is used; as a result, without stratification, it appears as if the new drug does not improve the mortality rate. Usually, we are interested in whether the new drug works or not, but not in the differences between the treatment groups caused by confounding, so the conclusion based on the upper table is considered a fallacy.

Treatment	Death	Survival	% Death
New drug	16	34	32%
Existing drug	15	35	30%

Baseline Severity	Treatment	Death	Survival	% Death
Severe	New drug	15	25	37.5%
	Existing drug	10	5	66.7%
Mild	New drug	1	9	10%
	Existing drug	5	30	14.3%

Table 1.1: An artificial example of Simpson's paradox.

In Example 1, it is mentioned that our concern is "whether the new drug works or not". We would like to think about this question a little more. What makes us judge whether the new drug works or not? To answer this question, it would require a philosophical discussion of causality; however, for the purposes of this thesis, we consider the case of "taking a new drug" and the case of "not taking a new drug," and compare the former with the latter to determine whether the new drug works or not. If the new drug works, we say the new drug has a causal effect compared with the existing drug. Unfortunately, it is impossible to answer this question for a single patient. This is because the patient can either "take the new drug" or "not take the new drug," and it is impossible to observe both outcomes simultaneously. This fact has been called the fundamental problem of causal inference [61, 33]. Fortunately, however, as we will see later, it is possible to infer causal effects on a population. This is why statistics plays an important role in causal inference.

Then, is it possible to deal with causality within a purely statistical framework? Actually, this is difficult. Let us consider causality and statistics through the following example.

Example 2: Rainfall and umbrella sales When it rains, sales of umbrellas increase because people who forget their umbrellas buy them. We now try to express this causal relationship in a statistical model. Let X be a binary random variable indicating whether it rains or not, and Y be the sales of umbrellas. Here, if there is a causal effect from X to Y, we define $P_{Y|X} \neq P_Y$, where P_Y is the marginal distribution of Y, and $P_{Y|X}$ is the conditional distribution of Y given X (the same applies in the opposite causal direction). Since our intuition is that there is a causal effect from X to Y, $P_{Y|X} \neq P_Y$ holds in this setting. Then, what about causality in the opposite direction? $P_{Y|X} \neq P_Y$ implies that X and Y are not independent, so $P_{X|Y} \neq P_X$ is also true. This means "increased sales of umbrellas bring about rainfall", but this conclusion is apparently unacceptable.

There are many other possible ways to define causal dependence in the framework of statistics. However, one that seems to work intuitively turns out not to be a good representation of causality. This encourages us to consider another framework to represent causal relationships. In the next section, we see that a framework involving a "what if...?" world successfully represents causal relationships.

1.1.2 Frameworks for Causal Inference

One way to express causality without contradiction is to consider a framework directly incorporating the world of "what if...?." Let us look at the example above again.

Example 2: Rainfall and umbrella sales (continue) Let $P_Y^{(1)}$ denote the marginal distribution of Y in the world where God has sent rain regardless of the laws of nature, and $P_Y^{(0)}$ denote the marginal distribution of Y in the world where God has not sent rain. If there is a causal relationship from X to Y, we define $P_Y^{(1)} \neq P_Y^{(0)}$. Since rainfall has a causal effect on umbrella sales, one can conclude $P_Y^{(1)} \neq P_Y^{(0)}$. Conversely, the marginal distribution of X in the world where God has increased the sales of umbrellas is $P_X^{(1)}$, and the marginal distribution of X in the world where God has not increased the sales is $P_X^{(0)}$. Similarly, we define the existence of causal effect from Y to X by $P_X^{(1)} \neq P_X^{(0)}$. Since the previous conclusion $P_Y^{(1)} \neq P_Y^{(0)}$ does not imply anything about the relationship between $P_X^{(1)}$ and $P_X^{(0)}$, we can follow our intuition and conclude $P_X^{(1)} = P_X^{(0)}$.

Unlike the statistical framework, this example indicates that the "what if...?" framework enables us to express the causal relationship without any contradiction. For further understanding on causality, we have to discuss more details of this causal framework.

In Section 2, we introduce the structural causal model (SCM) [44, 56] and the potential outcome model [71, 61, 41] as examples of such frameworks. SCM uses the tools of "do-operation" and "counterfactual" to describe the "what if...?" world in a relatively direct way. On the other hand, the potential outcome model is rarely interpreted as a representation of the "what if...?" world; however, the "potential outcome" in its name refers to the outcome that would have been obtained if a certain treatment had been taken, which is nothing but a variable in the "what if...?" world. In fact, it has been pointed out that both models represent essentially the same concept [28, 29, 56].

1.1.3 Statistical Estimation of Causal Quantities

In the previous section, we see that a framework involving the world of "what if...?" enables us to describe a causal relationship without contradiction. In this section, we review the relationship between this framework and statistical inference. In statistical inference, especially in frequentist statistics, the data are assumed to be a sample from a fixed population with the probability distribution. Statistical inference is a framework for estimating the probability distribution of this population based on the sample. Causal inference, on the other hand, is concerned with the "what if...?" world. If we assume that this imaginary population (certainly the sampling population is also "imaginary" in the sense that it is just a model, but we refer to something of the "what if...?" world as "imaginary") entails a certain probability distribution, how can we infer the properties of that distribution? One way, as scientists have long done, is randomized experiments [22, 59]. Randomized experiments can be viewed as a way of pseudo-sampling from the imaginary population of interest. If we conduct a randomized experiment, we can assume that we have a sample from the population of interest, and we can make inferences on the causal effects of treatment using the usual statistical methods (estimation of expected value by sample mean, t-test, analisys of variance, etc.).

However, randomized experiments are often impossible for various reasons: including ethics, cost, effort, and the nature of the research object. In such cases, inferences are based only on incomplete experiments, natural experiments, or non-experimentally obtained data [59, 19]. One of the main challenges of statistical causal inference is the identification of the causal quantity of interest. This is the question of how to represent a quantity of interest defined in the imaginary population using only the quantities of the sampling population, and what assumptions are necessary. If a causal quantity is identifiable, then an estimator can be constructed by approximating with the quantity which can be estimated by the observed data. For example, the stratified estimator in Example 1 is a consistent estimator of the mortality rate in each stratum because the identification assumption is satisfied by the stratification by the baseline severity.

Even if a causal quantity is identifiable, it is often difficult to estimate. For example, the average treatment effect (ATE) is the most fundamental causal quantity, and it is often estimated using the conditional expectation of the outcome or a model of the treatment assignment probability. If the model is misspecified, the estimator loses consistency. In some cases, some of the assumptions necessary for identification do not seem to hold for the sampling distribution.

Above, we discuss the inference on the "what if...?" world, but even if we are

interested in the sampling distribution, we sometimes want to estimate a model with causal implications. In other words, we assume there is a causal structure (SCM in most cases) behind the sampling population and infer the causal structure. This kind of problem is called statistical causal discovery. As we see in the previous section, models with causal implications have higher expressive power than ordinary statistical models. This high expressive power is related to the difficulty of identification and estimation. For example, among SCMs, a nonparametric model and a linear Gaussian model are not completely identifiable [70, 16, 56]. For complete identification, it is necessary to assume nonlinearity with certain conditions or linear non-Gaussianity, for example. We discuss this identification matter in causal discovery in Section 2.1.1. Such a model requires some devices, as it is necessary to construct the estimation method with the identification condition in mind. For example, based on the assumption of non-Gaussianity, we have to use a loss function that can leverage the information on higher-order moments.

In practical data analysis, we have to deal with not only the difficulties of causality, but also various difficulties of data (high dimensionality, missing data, outliers, etc.). This is difficult because it is not enough to simply deal with causal difficulties and data difficulties; we must also pay attention to the difficulties that arise from the combination of the two. Frangakis and Rubin have introduced the framework of "principal stratification" to estimate causal effects in the presence of dropouts [23]. This framework enables us to consider systematically what should be estimated and how it should be estimated, but it requires additional models to deal with dropouts because the principle strata of each subject is not known a priori. Another example is estimating causal effects with high-dimensional covariates [9, 8, 3]. Regularized regression [74, 24] is widely used in ordinary statistics when the dimensionality of the covariates is large. However, the parameter estimates by regularized regression are generally biased; then the estimates of causal effects using such regularized estimators are also biased. The basic concern of this paper is similar to these two examples, and we are interested in how to deal with the difficulties associated with estimating causal relationships while dealing with the difficulties in the data. This thesis focuses on the causal discovery for high-dimensional and sparse data and estimation of the average treatment effect under outlier contamination.

1.2 Our Contributions

We have made some contributions to statistical estimation methods for causal relationships.

The first work is the development of an algorithm to estimate causal graphs of the

linear non-Gaussian acyclic model (LiNGAM), which is one of the fully identifiable causal models. This is a joint work with Prof. Fujisawa, the tutor of my Ph.D. course. The estimation problem of LiNGAM can be formulated as the estimation of parameter matrix in independent component analysis (ICA). We have focused on the sparsity of causal graphs and have proposed an efficient estimation method based on the log-likelihood with sparse penalty. Since the estimation of ICA has a unique instability, we incorporated various devices for stabilization in addition to the sparse penalty. This work appeared in *Neurocomputing*, which is an academic journal on machine learning.

> Kazuharu Harada and Hironori Fujisawa. Sparse Estimation of Linear Non-Gaussian Acyclic Model for Causal Discovery. *Neurocomputing*, 459: 223-233, 2021.

The second work is the development of an estimator for the ATE under outlier contamination. This is also a joint work with Prof. Fujisawa. The ATE is a fundamental quantity of causal inference, and the inverse probability weighting (IPW) estimator and the doubly robust (DR) estimator are widely used to estimate the ATE. Since these estimators are vulnerable to outliers, we should use alternative methods under contamination. For this purpose, some IPW/DR estimators for median are available; however, the outlier resistance of the median is limited, especially when the ratio of outliers is not ignorably small. Our estimators, the density-powered IPW/DR estimator, and its variant, the εDP-DR estimator, effectively reduce the bias due to outliers by incorporating a density power weighting. This work is now under review, and the manuscript is available on arXiv.org.

> Kazuharu Harada and Hironori Fujisawa. Outlier-Resistant Estimators for Average Treatment Effect in Causal Inference. *arXiv preprint arXiv:2106.13946*, 2021.

1.3 Outline

This thesis is organized as follows. In Section 2, we introduce some fundamental notions of statistics for causality. Statistical causal discovery and statistical causal inference are both statistical methods to deal with causal relationships, but their purposes are different and the methodological differences are not small. In this section, we review the basics of each of the two fields. In particular, we discuss LiNGAM and the IPW/DR estimators, which are highly related to our contributions. In Section 3, our first contribution, the sparse estimation algorithm for LiNGAM, is

presented. In Section 4, our second contribution, the outlier-resistant estimator for the ATE is presented. Section 5 is a conclusion.

2 Preparation

2.1 Causal Discovery

2.1.1 Structural Causal Model and Causal Discovery

In statistical causal discovery, a system of variables is represented by a structural causal model (SCM). The SCM is defined as follows.

Definition 2.1 (Structural causal model; SCM)**.** *Let X be the d-dimensional random variable. A SCM consists of a set of assignments:*

$$X_j := f_j(\mathbf{PA}(j), E_j), \tag{2.1}$$

where $\mathbf{PA}(j) \subseteq \{X_1, ..., X_d\} \backslash \{X_j\}$ is called parents of X_j and $E_j \in \{E_1, ..., E_d\}$ are independent noises. The corresponding graph $\mathcal{G}$ is constructed by setting the variable as nodes and drawing directed edges from each parent to its child.

Unlike algebraic equations, an assignment implies the substitution of the right-hand side to the left-hand side. Given an SCM, the joint distribution of the variables in the system is uniquely determined. The proof is based on so-called ancestral sampling (e.g., see Appendix C.2. of [56]).

In this thesis, only the case where the graph $\mathcal{G}$ does not have a cyclic structure is considered. Besides, the independence of the error terms implies that there are no unobserved confounders. Causal discovery in the presence of cyclic structure has been discussed in [48, 53, 36, 39, 11], for example. Causal discovery in the presence of unobserved confounding is also discussed in [14, 63, 73, 62], for example.

The joint distribution entailed with an SCM has a probabilistic structure that corresponds to the graph structure. In other words, there is an (conditional) independence between the variables in the system that corresponds to SCM, which can be deduced from the graph structure. A simple example is given below. Suppose that an SCM containing variables A, B, and C is defined by the following assignments:

$$\begin{aligned} A &:= f_A(E_A), \\ B &:= f_B(A, E_B), \\ C &:= f_C(B, E_C). \end{aligned}$$

The corresponding graph, excluding the error terms, is $A \to B \to C$. From the independence of the error terms, we can say the following conditional independence:

$$A \perp\!\!\!\perp C | B.$$

Conversely, it is a natural idea to recover the original graph from the set of conditional independence statements inferred from the data. Assuming that the data were *i.i.d.* samples from the entailed joint distribution, the conditional independence between the variables can be inferred from the data. In fact, this idea directly forms the basis of a classical approach: constraint-based methods for causal discovery. For example, the PC algorithm [70] estimates the graph structure based on the conditional independence between variables. Assuming a linear Gaussian structure on the assignments, conditional independence can be tested based on partial correlation [49]. Alternatively, nonparametric tests based on the kernel method are available [27, 86]. However, there is an important problem with the conditional independence-based approach. It is the problem of identifiability that different SCMs can generate the same joint distribution [70, 44, 56]. In the above example, $A \perp\!\!\!\perp C|B$ also holds for SCMs satisfying $A \leftarrow B \leftarrow C$ or $A \leftarrow B \rightarrow C$, instead of $A \rightarrow B \rightarrow C$. This unidentifiable set of graphs is called the Markov equivalence class.

In addition to the constraint-based approach, there is another classical approach to causal discovery, which is based on some scores computed from the data and the model. For example, the GES algorithm [16] assumes a linear Gaussian model for the SCM and searches for the best model based on the BIC. Since it is computationally infeasible to cover all DAG structures when there are many variables in the system, the GES algorithm uses the greedy method for the search. As well as the constraint-based approach, the score-based method does not solve the problem of unidentifiability of Markov equivalences as long as it assumes a linear Gaussian model. Let us give an example. Suppose that the data is generated according to the following SCM:

$$\begin{aligned} A &:= E_A, \\ B &:= \beta_{BA} A + E_B, \end{aligned}$$

where $E_A \sim \mathcal{N}(0, \sigma_A^2), E_B \sim \mathcal{N}(0, \sigma_B^2)$. The entailed joint distribution of A and B is $\mathcal{MN}(\mathbf{0}, \Sigma_{A \rightarrow B})$, where

$$\Sigma_{A \rightarrow B} = \begin{pmatrix} \sigma_A^2 & \beta_{BA}\sigma_A^2 \\ 0 & \beta_{BA}^2 \sigma_A^2 + \sigma_B^2 \end{pmatrix}.$$

However, the following SCM with opposite causal direction is also entailed by the same joint distribution:

$$\begin{aligned} A &:= \beta_{AB} B + E'_A, \\ B &:= E'_B, \end{aligned}$$

where $E_A \sim \mathcal{N}(0, \sigma_A^2\sigma_B^2/(\beta_{BA}^2+\sigma_B^2))$, $E_B \sim \mathcal{N}(0, \beta_{BA}^2\sigma_A^2+\sigma_B^2)$, and $\beta_{AB} = \beta_{BA}\sigma_A^2/(\beta_{BA}^2+\sigma_B^2)$. This means that the causal direction is not identifiable even in the two-variable case under the linear Gaussian assumption.

In recent studies, it has been shown that the causal structure can be fully indentified by certain assumptions on its assignments. For example, the linear non-Gaussian acyclic model (LiNGAM), which assumes a linear function for the assignments and non-Gaussianity for all but one of the error terms, is fully identifiable. We discuss the details of LiNGAM in the next subsection. For linear models, Gaussian model with equal error variances is also being fully identifiable [55]. Other identifiable models are based on some nonlinear functions. The nonlinear additive noise model (ANM) [57],

$$X_j := f_j(\mathbf{PA}(j)) + E_j, \tag{2.2}$$

is also identifiable. The postnonlinear model [85] is a more general class of identifiable SCM models:

$$X_j := g_j(f_j(\mathbf{PA}(j)) + E_j). \tag{2.3}$$

Identifiable models are comprehensively discussed in Chapters 4 and 7 of [56], for example. In this thesis, we focus on LiNGAM because of two reasons: (1) it is easy to interpret by virtue of its linearity, and (2) the assumption of non-Gaussianity is milder than the assumption of equal variance, which is the key assumption of another identifiable linear model [55]. Because of this preferable aspects, LiNGAM has several known applications [e.g. 50, 80, 83].

2.1.2 Linear Non-Gaussian Acyclic Model (LiNGAM)

LiNGAM is an acyclic SCM with independent non-Gaussian errors and linear assignments. Let $X \in \mathbb{R}^d$ be observed variables of the system, and then the following equation holds for the entailed joint distribution of LiNGAM

$$X = \mathbf{B}X + E. \tag{2.4}$$

$\mathbf{B} \in \mathbb{R}^{d\times d}$ is the coefficient matrix in the assignments; if the elements of X are aligned from upper to lower in the causal ordering, then $\mathbf{B}$ is a strictly lower triangular matrix. In other words, suppose that there exists a permutation matrix $\mathbf{Q}$ (which is an orthogonal matrix) and multiply $\mathbf{Q}$ by (2.4) from the left as

$$\mathbf{Q}X = (\mathbf{Q}\mathbf{B}\mathbf{Q}^T)\mathbf{Q}X + \mathbf{Q}E. \tag{2.5}$$

Then the matrix $\mathbf{Q}\mathbf{B}\mathbf{Q}^T$ becomes strictly lower triangular. $E \in \mathbb{R}^d$ is a vector of independent non-Gaussian errors. Estimation of $\mathbf{B}$ can be seen as a problem of independent component analysis (ICA) by solving (2.4) with respect to E and letting $\mathbf{W} = \mathbf{I} - \mathbf{B}$:

$$E = (\mathbf{I} - \mathbf{B})X = \mathbf{W}X. \tag{2.6}$$

This formula can be seen that the observed variable X has been generated by mixing the independent signals E by the matrix $(\mathbf{W})^{-1}$. The identifiability of LiNGAM relies on the identifiability of ICA. ICA is identifiable except for ordering and scale of the independent components. Thus, the solution of ICA $\mathbf{W}_{ICA}$ results from disarranging the order and scale of the rows of $\mathbf{I} - \mathbf{B}$. The epochal point of LiNGAM is that the row ordering and scale, which are not identified by ICA, can be identified by utilizing the acyclicity of the SCM. If the true SCM is acyclic, then $\mathbf{B}$ can uniquely be transformed into an exact lower triangular matrix. This implies that the diagonal elements of $\mathbf{W} = \mathbf{I} - \mathbf{B}$ are all 1. There is a unique transformation $\mathcal{T}(\cdot)$ involving reordering and rescaling the rows of $\mathbf{W}_{ICA}$ that makes all diagonal components of $\mathcal{T}(\mathbf{W}_{ICA})$ equal to 1 (see Appendix A of [64]). Therefore, the coefficient matrix $\mathbf{B}$ of LiNGAM is completely identifiable by virtue of the combination of the identifiability of ICA and the acyclicity of SCM.

Various estimation methods have been proposed for LiNGAM. In this section, we introduce two major, but different types of algorithms: ICA-LiNGAM[64] and DirectLiNGAM[65]. The first method, ICA-LiNGAM, is the one proposed in the original paper of LiNGAM. ICA-LiNGAM estimates B using ICA as well as its identifiability proof. The algorithm is shown below.

1. Obtain an estimate $\widehat{\mathbf{W}}_{ICA}$ for $\mathbf{W}_{ICA}$ from the data matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ by a standard ICA algorithm (e.g. FastICA[37]).

2. Apply the transformation $\mathcal{T}$ via the following two steps.

   (a) Obtain a permutation $\mathbf{P}$ such that the diagonal elements of $\mathbf{P}\widehat{\mathbf{W}}_{ICA}$ are non-zero by minimizing $\sum |(\mathbf{P}\widehat{\mathbf{W}}_{ICA})_{ii}|^{-1}$. This can be done by a classical linear assignment algorithm like Hungarian method (See [13]).

   (b) Next, rescale the rows of $\mathbf{P}\widehat{\mathbf{W}}_{ICA}$ by its diagonal elements. Let $\widehat{\mathbf{W}}_{pre}$ be the permuted and rescaled parameter matrix and let $\widehat{\mathbf{B}}_{pre} = \mathbf{I} - \widehat{\mathbf{W}}_{pre}$.

3. Obtain causal ordering of $\{X_1, ..., X_d\}$ by searching for a permutation matrix $\mathbf{Q}$ which makes $\mathbf{Q}\widehat{\mathbf{B}}_{pre}\mathbf{Q}^T$ lower triangular. This step is formulated as the minimization of the absolute sum of upper triangular elements of $\mathbf{Q}\widehat{\mathbf{B}}_{pre}\mathbf{Q}^T$.

4. Based on the obtained causal ordering, estimate the coefficient matrix by sparse linear regression [74, 92]. Namely, let $k(j)$ be the causal ordering of X_j, and regress X_j on $\{X_l; k(j) > k(l), l \in \{1, ..., d\}\}$ for all js.

The 3rd step of ICA-LiNGAM can be quite heavy in computation because the number of possible permutations grows so rapidly as the dimension d increases. This is intractable even in non-high dimensional cases such as $d > 10$. In that case, the permutation matrix $\mathbf{Q}$ can be obtained by a fast but somewhat ad hoc way:

3'-1 Replace the value closest to 0 in $\widehat{\mathbf{B}}_{pre}$ with 0.

3'-2 Test whether the matrix $\mathbf{Q}$ exists such that $\mathbf{Q}\widehat{\mathbf{B}}_{pre}\mathbf{Q}^T$ is strictly lower triangular.

3'-3 If the step 4'-2 fails, replace the next smallest (in absolute value) value with 0 and return to Step 4'-2.

Step 3'-2 can be done fast in a simple algorithm (See Algorithm B in [64]).

The another approach, DirectLiNGAM, is based on Darmois-Skitovich theorem [67, 68, 69].

Theorem 2.1 (Darmois-Skitovich)**.** *Let Z and W as*

$$Z = \sum_{j=1}^{d} \alpha_j S_j$$
$$W = \sum_{j=1}^{d} \beta_j S_j,$$

where S_j $(j = 1, ..., d)$ are the independent random variables and α_j, β_j $(j = 1, ..., d)$ are the coefficient constants. If Z and W are independent, all S_js that satisfy $\alpha_j\beta_j \neq 0$ are normally distributed.

This theorem is also used to show the identifiability of ICA [17]. In contraposition, the theorem implies that Z and W are not independent if there exists a non-Gaussian S_j with $\alpha_j\beta_j \neq 0$. Using this theorem, it can be shown the necessary and sufficient conditions for X_k to be exogenous, i.e., $X_k := E_k$ in the SCM.

Theorem 2.2 (Lemma 1 in [65]). *Assume that the observed variable X follows the LiNGAM (2.4). Let R_{jk} be the residual for which X_j is regressed on X_k:*

$$R_{jk} = X_j - \frac{\mathrm{Cov}[X_j, X_k]}{\mathrm{Var}[X_k]} X_k \tag{2.7}$$

Then, the variable X_k is independent of other variables if and only if X_k is independent of its residuals R_{jk} for all $j \neq k$.

Based on this result, the following algorithm called DirectLiNGAM has been proposed in [65].

1. Initialize the vector of active variables as $X_{\mathcal{A}} = X$, and let $U = \{1, ..., d\}$ and $K = \emptyset$ be an ordered list of the indices of X_j.

2. Repeat the following steps until $d - 1$ variables are appended to K:

   (a) For all k, regress X_j on X_k for all $j \in U \backslash K (j \neq k)$ and compute the residual vector $R_{\cdot k} \in \mathbb{R}^{|U \backslash K|}$ (note that the kth element R_{kk} is 0).

   (b) Compute the following independence measure:

   $$T(X_k; U \backslash K) = \sum_{j \in U \backslash K, j \neq k} MI(X_k, R_{jk}), \tag{2.8}$$

   where the mutual information MI is estimated by the kernel-based method [4], for example.

   (c) Find l that X_l minimizes $T(X_l; U \backslash K)$, and append l to K.

   (d) Update $X_{\mathcal{A}}$ with $R_{\cdot l}$.

3. Append the last variable to the end of K.

4. Estimate the coefficient matrix in the same manner as ICA-LiNGAM.

In summary, DirectLiNGAM finds the most upstream variable of the causal ordering among the set of active variables $X_{\mathcal{A}}$ by running through Step (a) to Step (d). The most upstream variable is found by minimizing the sum of the independence measure between the regressor and the regression residuals. Unlike ICA-LiNGAM, DirectLiNGAM does not involve an iterative optimization step in its algorithm, so it has the advantage of being completed in a fixed number of computations for a given data set. Another advantage is that it is easy to incorporate prior knowledge.

These methods, ICA-LiNGAM and DirectLiNGAM, are very convenient to estimate $\mathbf{B}$ of LiNGAM, however, they are based on two different statistical criteria;

the step defining the causal ordering is based on non-Gaussianity (independence), and the parameter estimation step is based on a squared loss with a sparse penalty. As discussed in detail in Section 3, such features of these methods can lead to undesirable information loss. In our work, we seek to improve this aspect and propose an efficient estimation method.

2.2 Causal Inference

2.2.1 Potential Outcome Model

Whereas statistical causal discovery aims at estimating the causal structure of the system itself, statistical causal inference assumes the causal structure is known and is mainly concerned with the definition of appropriate causal effects and how to identify and estimate such quantities. It is also common to discuss methods of statistical causal inference within the framework of SCM as well as statistical causal discovery; however, Rubin's causal model [41] is also widely used when we are interested in the effect of a particular treatment on the outcome. In this thesis, the estimation of treatment effects is discussed based on Rubin's causal model. Note that the notations of the variables are different from those in the previous subsection.

Now we introduce Rubin's causal model under the common setting that causes Simpson's paradox. Let $Y_i \in \mathbb{R}$ be the observed outcome, let $T \in \{0, 1\}$ be the dichotomous treatment indicator, and let $X_i \in \mathcal{X}$ be the confounders. Rubin's causal model introduces the potential outcome $Y^{(t)} \in \mathbb{R}$, which denotes the hypothetical outcome of what if an individual received the treatment t. $Y^{(t)}$ can be observed only when the individual actually receives the treatment t. For example, suppose that the dataset in Table 2.1 is available. Due to missingness, the individual-level treatment

ID	Actual Treatment	Outcome Y	$Y^{(1)}$	$Y^{(0)}$
1	New Drug	Died	Died	(missing)
2	New Drug	Survived	Survived	(missing)
3	Existing Drug	Died	(missing)	Died
4	Existing Drug	Died	(missing)	Died

Table 2.1: The potential outcome can only be observed for the actually received treatment for each individual. $Y^{(1)}$ is the potential outcome with the new drug, and $Y^{(0)}$ is that with the existing drug.

effect $Y_i^{(1)} - Y_i^{(0)}$ cannot be computed in reality. This is the fundamental problem of causal inference [61, 33]. Instead, distributional parameters like the average treatment effect (ATE), which is defined as $\mathbb{E}[Y^{(1)} - Y^{(0)}]$, are used to express the group-level causal effect. Although there are some other types of causal effects, which

are sometimes more informative than the ATE, such as the quantile treatment effect [1, 15, 21, 18] and the density treatment effect [47], the classical ATE is often the center of interest.

If the treatment is randomly assigned, the ATE is equivalent to $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$, which can be estimated by *i.i.d.* observations. Random assignment is common in experimental settings like randomized controlled trials, but it usually does not hold in observational settings. Example 1 in Section 1.1.1 is a typical case for which the random assignment does not hold, and it falls into Simpson's paradox. To identify the ATE from observational data, it is necessary to adjust the confounding. In Rubin's causal model, ATE is identified under the following three assumptions:

1. Exchangeability: $Y^{(t)} \perp\!\!\!\perp T|X \quad \text{for all} \quad t \in \{0,1\}$.
2. Consistency: $Y = Y^{(t)} \quad \text{if} \quad T = t$.
3. Positivity: $P(T=1|X) > c$ for some c>0.

Exchangeability, also known as ignorability or unconfounded, means that the treatment is randomly assigned conditional on confounders. Consistency ensures that the observed outcome reflects the value of the potential outcome corresponding to the actually received treatment. When the treatment is dichotomous, it can also be written as $Y = TY^{(1)} + (1-T)Y^{(0)}$. Positivity represents any individuals have a chance to be assigned to either treatment, no matter what their values of confounders are. Under these assumptions, ATE can be transformed as follows:

$$\begin{aligned}
\mathbb{E}[Y^{(1)} - Y^{(0)}] &= \mathbb{E}[\mathbb{E}[Y^{(1)} - Y^{(0)}|T, X]] \\
&= \mathbb{E}[\mathbb{E}[Y^{(1)}|T, X] - \mathbb{E}[Y^{(0)}|T, X]] \\
&= \mathbb{E}[\mathbb{E}[Y^{(1)}|T=1, X] - \mathbb{E}[Y^{(0)}|T=0, X]] \\
&= \mathbb{E}[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X]].
\end{aligned}$$

The third equality holds because of exchangeability, and the fourth equality holds because of consistency. Stratification enables us to estimate this estimand in a relatively straight manner. In the stratification approach, the sample is divided so that the confounders are equally distributed across the treatment groups within each stratum, and the sample mean by group is taken in each stratum to estimate $\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]$. Then, depending on the sample size of each stratum, the strata are combined to estimate ATE. Matching involves pairing samples of the treatment group and the control group that have the same values of confounders or are close to each other based on certain criteria. And then $\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]$ is estimated based on the difference in outcomes between the pairs. Stratification and

matching are preferable in the sense that they do not depend on a particular statistical model. However, they have the disadvantage that there is some arbitrariness in how to form strata or pairs. Moreover, stratification and matching are practically impossible when the confounding variables are high-dimensional.

Outcome regression is an estimation method that can deal with the problem of high dimensionality. It models $\mathbb{E}[Y|T=1,X]$ and $\mathbb{E}[Y|T=0,X]$ by some statistical models $m_1(X;\beta_1)$ and $m_0(X;\beta_0)$, respectively. Then, ATE is estimated by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{m_1(X_i;\hat{\beta}_1)-m_0(X_i;\hat{\beta}_0)\right\}, \tag{2.9}$$

where $\hat{\beta}_1$ and $\hat{\beta}_0$ are obtained by a standard estimation method like least squares or maximum likelihood estimation. In this way, ATE can be estimated by outcome regression even with high-dimensional counfounders. However, it is necessary to model the conditional expectation correctly for a consistent estimation of ATE. In general, a deep scientific consideration on the variables is essential to model the relationship between the outcome and confounding variables correctly. This modeling is often very difficult. On the other hand, the propensity score, reviewed in the next subsection, models the mechanism between treatment assignment and confounding variables. The mechanism of treatment assignment may be more tractable than modeling the relationship between the outcome and confounding variables. For example, guidelines for clinical practice exist in a medical setting. An additional advantage is that even when multiple outcomes are of interest, it is not necessary to model the relationship between each outcome and the confounders one by one if the correct propensity score is obtained.

2.2.2 Propensity Score-Based Estimators for ATE

As discussed above, stratification and matching are practically impossible when the confounders are high dimensional. Outcome regression can deal with this problem, but in turn, it faces the problem of model misspecification. Then, we introduce the propensity score as another approach to estimating ATE. The propensity score models the conditional probability of assignment $P(T=1|X)$. Typically, logistic regression with maximum likelihood estimation is assumed. The propensity score has a balancing property, which states that the distributions of the covariates among the treatment groups are equal when conditioned on the propensity score.

While the propensity score can be used for stratification and matching, this section mainly discusses the inverse probability weighting (IPW) method [58, 60] and its extensions. ATE can also be identified in different ways than in the previous

section:

$$\begin{aligned}
\mathbb{E}[Y^{(1)}] &= \iint y dP_{Y^{(1)}|X}(y|x) dP_X(x) \\
&= \iint y dP_{Y|T=1,X}(y|T=1,x) dP_X(x) \\
&= \iint \frac{1 \cdot y}{P(T=1|x)} \cdot P(T=1|x) \\
&\qquad + \frac{0 \cdot y}{P(T=0|x)} \cdot P(T=0|x) dP_{Y|T=1,X}(y|T=1,x) dP_X(x) \\
&= \iiint \frac{ty}{P(T=1|x)} dP_{Y|T,X}(y|t,x) dP_{T|X}(t|x) dP_X(x) \\
&= \mathbb{E}\left[\frac{TY}{P(T=1|X)}\right].
\end{aligned}$$

Since the expectation of $Y^{(0)}$ is also identified as

$$\mathbb{E}[Y^{(0)}] = \mathbb{E}\left[\frac{(1-T)Y}{1-P(T=1|X)}\right],$$

ATE is identified by using the propensity score. Let $\pi(X;\hat{\alpha})$ be the propensity score with estimated parameter $\hat{\alpha}$, and then the IPW estimator is

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_i Y_i}{\pi(X_i;\hat{\alpha})} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-T_i)Y_i}{1-\pi(X_i;\hat{\alpha})}. \tag{2.10}$$

In the following, we only discuss estimation of $\mathbb{E}[Y^{(1)}]$ for simplicity. The IPW estimator has some different forms [51]. In particular, the following form has a smaller asymptotic variance than (2.10). Moreover, it can be regarded as a solution to the following estimating equation:

$$\begin{aligned}
\hat{\mu}_{IPW} &= \left(\sum_{i=1}^{N}\frac{T_i Y_i}{\pi(X_i;\hat{\alpha})}\right)\left(\sum_{i=1}^{N}\frac{T_i}{\pi(X_i;\hat{\alpha})}\right)^{-1}, \\
&\sum_{i=1}^{N}\frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i-\mu) = 0.
\end{aligned} \tag{2.11}$$

The doubly robust (DR) estimator [5] is an extension of the IPW estimator that is resistant to model misspecification. The typical DR estimator for $\mathbb{E}[Y^{(1)}]$ is usually

defined as follows:

$$\hat{\mu}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{T_i Y_i}{\pi(X_i;\hat{\alpha})} - \frac{T_i - \pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})} m_1(X_i;\hat{\beta})\right\}, \tag{2.12}$$

where $m_1(X;\beta)$ is a model for $\mathbb{E}[Y|T=1,X]$ and $\hat{\beta}$ is an estimator of β. The DR estimator combines the IPW estimator and the outcome regression. This type of estimator is also called the augmented IPW (AIPW) estimator. The DR estimator is consistent with $\mu^{(1)}$ if either the propensity score model or the outcome regression model is correctly specified. The name "doubly robust" comes from this property. In addition, semiparametric theory ensures that the DR estimator has the lowest asymptotic variance as an estimator for $\mu^{(1)}$ if both models are correctly specified[77, 5, 76]. The proof of the double robustness is given in Chapter 13 of [32] or Chapter 6 of [76], for example. The DR estimator is also represented as the solution of the following estimating equation:

$$\sum_{i=1}^{N}\left\{\frac{T_i}{\pi(X;\hat{\alpha})}(Y_i - \mu) - \frac{T_i - \pi(X;\hat{\alpha})}{\pi(X;\hat{\alpha})}(m_1(X_i;\hat{\beta}) - \mu)\right\} = 0.$$

Since the IPW and DR estimators can be regarded as solutions to the estimating equations, they can be considered as one of the M-estimators, which is discussed in Section 4. This allows us to apply the theory of M-estimation to the proof of key asymptotic properties such as consistency and asymptotic normality.

The IPW and DR estimators are very useful, however, they are vulnerable to outliers because they involve the sample average. As we see in Section 4, it is insufficient to estimate $m_1(X_i;\hat{\beta})$ in an outlier-resistant way. Quantile-based approach is relatively resistant to outliers, but the resistance is limited in case the contamination ratio is not small. We propose novel IPW and DR type estimators that are more resistant to outliers than IPW/DR quantile-based estimators.

3 Sparse Estimation of LiNGAM

3.1 Background

In this work, the goal is to estimate the structure and parameter matrix of LiNGAM with sparse connectivity. Notations are taken over from Section 2.1.

As introduced in Chapter 2, LiNGAM is one of the causal discovery models that is completely identifiable and is characterized by its interpretability. We introduced two popular methods for the estimation of LiNGAM: ICA-LiNGAM and DirectLiNGAM. There are other methods similar to DirectLiGAM: Pairwise LiNGAM [39] and High-dimensional LiNGAM [81], but these are two-step methods that first use a discrete algorithm to infer the causal ordering, and then use penalized least squares to estimate the presence or absence of directed edges and path coefficients. Here, we focus on a sparse structure of causal relationships in high-dimensional data. The two-stage methods based on different criteria are not efficient in this case for two reasons: (i) the sparse structure is not always incorporated in causal order estimation, and (ii) the information of higher-order moments is not used in parameter estimation although the model is assumed to be non-Gaussian. To address these issues, we develop a likelihood-based and one-criterion algorithm incorporating the sparse structure. A sparse estimation method based on ICA for LiNGAM was already discussed [87, 88]. However, their methods do not satisfy the prerequisites for consistency, so that their estimation seems unstable. This work addresses the issue of consistency by developing a sparse estimation algorithm using whitened data with two penalty terms: a generalized lasso [75] type penalty and another penalty which is related to the consistency condition.

Another approach, the new characterization of "DAGness" has been introduced by [91], which uses the continuous function $h : \mathbb{R}^{d\times d} \to \mathbb{R}$ such that $h(\mathbf{B}) = 0$ if and only if the corresponding graph is acyclic. Previously, combinatorial optimization techniques have been essential for finding the structure of a DAG, but by using the function h, a DAG can be obtained by solving a fully continuous optimization. Their novel structure learning method is formulated as the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{B}\in\mathbb{R}^{d\times d}} \quad & \frac{1}{2N}\|\mathbf{X} - \mathbf{X}\mathbf{B}^T\|_F^2 + \lambda\|\mathbf{B}\|_1 \\ \text{subject to} \quad & h(\mathbf{B}) = 0, \end{aligned} \tag{3.1}$$

where $\mathbf{X} \in \mathbb{R}^{N\times d}$ is the data matrix which is supposed to be drawn from the entailed distribution $P_{\mathfrak{C}}(X)$ of an unknown linear SCM $\mathfrak{C}$. This is called Non-

combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS). In addition, a new algorithm called Gradient-based Optimization of dag-regularized Likelihood for learning linEar dag Model (GOLEM) has been proposed [54] recently. GOLEM uses h as a penalty term in combination with L1 regularization. The main term of the loss function is the log-likelihood of Gaussian distribution. There are two types of GOLEMs: GOLEM-EV (Equal Variance) and GOLEM-NV (Not equal Variance). They investigated some significant roles of h and the L1 penalty in structure learning.

As mentioned in Section 2.1, an identifiable linear SCM requires some additional assumptions on noise, namely, Gaussian with equal variance [55] or non-Gaussianity [64]. Since GOLEM and NOTEARS do not assume non-Gaussianity, they may not be appropriate as comparative methods in this work. Nevertheless, we included these methods in the numerical experiments because these methods are said to perform better than LiNGAM in some linear non-Gaussian settings.

In this work, we propose a method for estimating LiNGAM based on the penalized log-likelihood of ICA. The proposed method consists of a single statistical criterion, which leverages the sparse structure and the information of higher-order moments. By virtue of this single criterion, we can (i) incorporate the sparse structure in the estimation of causal structure and (ii) use the information of higher-order moments in parameter estimation. Besides, a generalized lasso type penalty is employed as the sparse penalty instead of an ordinary L1 penalty, and an "orthogonal penalty" is incorporated to suppress the correlation of the estimated independent components. These devices aim to bring the estimator closer to the consistent one. Our method requires an iterative optimization algorithm. We construct an efficient algorithm based on gradient descent and Alternating Direction Method of Multipliers [ADMM; 12]. ADMM is a widely used algorithm for sparse estimation. A modified natural gradient is introduced to search in the matrix space with some restrictions. We also propose an objective procedure for selecting a tuning parameter via likelihood cross-validation (CV). In order to verify the effectiveness and scalability of the proposed method, we demonstrate exhaustive numerical experiments. The proposed method is compared with some estimation methods of LiNGAM and other approaches to structure learning including NOTEARS [91] and GOLEM [54]. The proposed method outperforms the comparative methods in almost all cases and shows stable performance even in high-dimensional cases. The proposed method is also applied to real data. Finally, we give a conclusion.

3.2 Proposed Method

In this section, we propose a new method, which we call sparsely mixing ICA-LiNGAM (sICA-LiNGAM), to estimate the linear DAG model with non-Gaussian noises. The parameter is estimated based on the log-likelihood of ICA with two penalties related to sparsity and orthogonality.

3.2.1 ICA and Consistent Estimation

Here we start with a brief review of ICA and its consistent estimation. Let $S = (S_1, \ldots, S_d)^T$ be the vector of independent components (ICs) with zero means. Note that S corresponds to the error term E of linear SCM (2.1). Suppose that we observe $X = \mathbf{A}S$. The purpose of ICA is to recover the ICs S by $\mathbf{M}X$, where $\mathbf{M}$ is the parameter matrix.

Let p_j be the probability density function of S_j, and let $\mathbf{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_d)^T$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$. The maximum likelihood estimator (MLE) $\hat{\mathbf{M}}$ is given by

$$\hat{\mathbf{M}} = \arg\max_{\mathbf{M}} \ell(\mathbf{M}; \mathbf{X}), \tag{3.2}$$

$$\ell(\mathbf{M}; \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d} \log p_j(\boldsymbol{m}_j^T \boldsymbol{x}_i) + \log |\det \mathbf{M}|. \tag{3.3}$$

In ICA, we assume that p_js are unknown, so that it seems impossible to estimate $\mathbf{M}$ consistently. However, even when p_js are unknown, the following theorem tells that MLE has consistency except for the indeterminacy of the scale and the order of the estimated ICs if we use appropriate probability density functions $\tilde{p}_j$s instead of p_js.

Theorem 3.1 (Theorem 9.1 in Chapter 9 of [38])**.** *Let $Y_j = \boldsymbol{m}_j^T X$ be the estimated IC for all $j \in \{1, ..., d\}$. Suppose that Y_js are uncorrelated with unit variance. The MLE $\hat{\mathbf{M}}$ has consistency except for the indeterminacy of the scale and the ordering of ICs, if*

$$\mathbb{E}[S_j(\tilde{g}_j(S_j)) - \tilde{g}_j'(S_j)] > 0 \quad \textit{for } j = 1, \ldots, d, \tag{3.4}$$

where

$$\tilde{g}_j(s) = \frac{\partial}{\partial s} \log \tilde{p}_j(s). \tag{3.5}$$

In the likelihood-based ICA algorithm, the density functions $\tilde{p}_j$s are adaptively selected from two candidates such that (3.4) is satisfied for all ICs. For example, the

next functions are suitable for this purpose:

$$\log p_j^+(s) = a_1 - 2\log\cosh(s) \tag{3.6}$$

$$\log p_j^-(s) = a_2 - [s^2/2 - \log\cosh(s)], \tag{3.7}$$

where a_1 and a_2 are related to the normalization constant, which vanishes in (3.4). Density p^+ is used for super-Gaussian components and p^- for sub-Gaussian ones. Detailed discussion is provided in Chapter 9 of [38].

Pre-whitening is often employed in ICA algorithms to convert the assumptions on Y_js in Theorem 3.1 into a convenient parameter constraint. Consider the spectral decomposition of $(1/N)\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$, where $\mathbf{V}$ is the orthogonal matrix and $\mathbf{D}^2$ is the diagonal matrix whose diagonal entries are the eigenvalues of $(1/N)\mathbf{X}^T\mathbf{X}$. Let

$$\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)^T = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}, \tag{3.8}$$

where the diagonal elements of $\mathbf{D}$ are positive. Then we have $(1/N)\mathbf{Z}^T\mathbf{Z} = I_d$, so that the transformed variable $\mathbf{Z}$ is uncorrelated with unit variance. Let $\mathbf{W} = \mathbf{M}\mathbf{V}\mathbf{D}$. Then the MLE of $\mathbf{W}$ is reformulated as

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \tilde{\ell}(\mathbf{W}; \mathbf{Z}), \tag{3.9}$$

$$\tilde{\ell}(\mathbf{W}; \mathbf{Z}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} \log p_j(\boldsymbol{w}_j^T \boldsymbol{z}_i) + \log|\det \mathbf{W}|, \tag{3.10}$$

where $\boldsymbol{w}_j$ is the jth row of $\mathbf{W}$. We also have $\hat{\mathbf{M}} = \hat{\mathbf{W}}\mathbf{D}^{-1}\mathbf{V}^T$.

Let $Y_{ij} = \boldsymbol{m}_j^T \boldsymbol{x}_i$ and $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{N\times d}$. We have $\mathbf{Y} = \mathbf{X}\mathbf{M}^T = \mathbf{Z}\mathbf{W}^T$. Here, we suppose $\mathbf{W}$ is an orthogonal matrix. Considering the pre-whitening, the orthogonality of $\mathbf{W}$ keeps $\mathbf{Y}$ whitened:

$$\frac{1}{N}\mathbf{Y}^T\mathbf{Y} = \mathbf{W}\left(\frac{1}{N}\mathbf{Z}^T\mathbf{Z}\right)\mathbf{W}^T = \mathbf{W}\mathbf{W}^T = \mathbf{I}. \tag{3.11}$$

This implies that if we assume $\mathbf{W}$ is an orthogonal matrix, the estimated ICs are whitened even in a finite sample. Many ICA algorithms efficiently obtain $\hat{\mathbf{W}}$ by utilizing this orthogonality [38]. The previous works [87, 88] are based on the likelihood-based ICA with sparse penalty as well; however, since these works do not whiten the observed variables and pay no attention to the correlation of ICs, it should not generally have consistency. Note that (3.11) should hold at the population level, as seen in Theorem 3.1, and hence the orthogonality constraint is not necessary to be satisfied strictly in a finite sample.

In our approach, it is difficult to make $\mathbf{W}$ to be strictly orthogonal because of sparsity, and therefore we relax the orthogonal constraint to a penalty term, which is detailed in the next section. By making $\mathbf{W}$ closer to an orthogonal matrix, we expect our estimator for $\mathbf{W}$ to get closer to a consistent one.

3.2.2 ICA with Sparse Penalty for Causal Discovery

Returning to the linear SCM estimation, we consider the sparse structure of $\mathbf{M}$, because about the half of the entries of $\mathbf{M}$ $(= \mathbf{I} - \mathbf{B})$ should be zero due to the acyclicity. In addition, the causal structure may be simple, which leads to more sparsity into $\mathbf{M}$. Unfortunately, we cannot impose sparsity on $\mathbf{B}$ directly because the MLE cannot identify the order of ICs, and $\mathbf{B}$ cannot be derived correctly in the likelihood-maximization process. Therefore, we impose sparsity on $\mathbf{M}$ instead of $\mathbf{B}$, and, the order of the ICs is decided in a similar way to ICA-LiNGAM after the estimation of $\mathbf{B}$ is obtained.

For sparse estimation, we add to the log-likelihood (3.10) a sparse penalty of the adaptive lasso [92]:

$$\mathcal{P}_\gamma(\mathbf{M}) = \sum_{j,k=1}^{d} c_{jk}^{\gamma} |m_{jk}|, \quad c_{jk} > 0. \tag{3.12}$$

A typical example of the weight is $c_{jk} = 1/|m_{jk}^0|$, where m_{jk}^0 is an initial estimator for m_{jk}. A candidate for m_{jk}^0 is the MLE or the estimate with unweighted L1 regularization. The tuning parameter γ can be selected by cross-validation, however, we employ $\gamma = 1$ in this work for simplicity, which has been used in some relevant papers [88, 40]. The penalty term of the adaptive lasso is known to reduce the bias associated with L1 regularization and has oracle properties, especially in the case of (generalized) linear regression. Recalling the data is pre-whitened in ICA, the sparse penalty must be imposed on $\mathbf{M} = \mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T$, which is the linear transform of the parameter matrix $\mathbf{W}$. This type of sparse estimation is called generalized lasso [75].

Here we review the principal component analysis (PCA) with sparsity in order to understand the sparse estimation of an orthogonal parameter matrix. The PCs are usually obtained under the orthogonality constraint as well as ICs. Jolliffe et al. (2003) [42] has proposed a method for obtaining sparse PCs by maximizing the explained variances with an L_1 penalty under the orthogonality constraint. It has been reported in Zou et al. (2006) [93] that it is difficult to obtain sufficiently sparse PCs by the method of [42], and then they have proposed a smart idea to obtain sparse PCs by relaxing the orthogonality. Unfortunately, such an idea cannot be applied to our situation directly. The most important message is that the sparse

penalty may not work well under the strict orthogonality constraint. To address this issue, a novel method is proposed in this work, which contains an additional penalty related to the orthogonality constraint.

To obtain a sparse structure on $\mathbf{M}$, we relax the orthogonality constraint on $\mathbf{W}$. More precisely, we impose a unit-norm constraint on each row of $\mathbf{W}$ and relax the off-diagonal orthogonality constraint by adding a penalty term $\|\mathbf{P}^T\mathbf{W} - \mathbf{I}\|_F^2$ with $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ to the loss function. The unit-norm constraint fixes the variance of the estimated ICs to 1. The penalty $\|\mathbf{P}^T\mathbf{W} - \mathbf{I}\|_F^2$ reduces the correlation of estimated ICs. These make the estimate partially satisfy the condition of Theorem 3.1. Note that the additional penalty term brings the parameter matrix closer to the orthogonal matrix because the minimizer of $\|\mathbf{P}^T\mathbf{W} - \mathbf{I}\|_F^2$ under $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ satisfies $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

Finally, we summarize the proposed method. The observed data is whitened first. The loss function consists of the negative log-likelihood (3.10), the adaptive lasso penalty (3.12) on the matrix $\mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T$, and the orthogonality penalty $\|\mathbf{P}^T\mathbf{W} - \mathbf{I}\|_F^2$ under $\mathbf{P}^T\mathbf{P} = I$. Denote the loss function by

$$F(\mathbf{W}) = -\tilde{\ell}(\mathbf{W}; \mathbf{Z}) + \lambda \left\{ \alpha \mathcal{P}_\gamma(\mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T) + \tfrac{(1-\alpha)}{2} \|\mathbf{P}^T\mathbf{W} - \mathbf{I}\|_F^2 \right\}.$$

Let $\mathcal{N} \subset \mathbb{R}^{d \times d}$ be the set of non-singular matrices whose row vectors are normalized. The parameter estimation problem is defined by

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W} \in \mathcal{N}} F(\mathbf{W}) \qquad \text{subject to} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{3.13}$$

$\tilde{p}_j$s are adaptively selected from the candidates: (3.6) and (3.7). The tuning parameter λ controls the total extent of the penalty, and α balances the sparsity and the orthogonality. These parameters can be selected via cross validation (CV). After estimating $\mathbf{M}$ by $\hat{\mathbf{M}} = \hat{\mathbf{W}}\mathbf{D}^{-1}\mathbf{V}^T$, we recover the full causal causal structure with a non-statistical procedure. The CV and the recovering procedure are presented in the next section. Note that each column of $\mathbf{X}$ is centered and normalized before whitening to impose the sparse penalty evenly, as is usually employed in lasso.

3.3 Algorithm

In this section, we show the whole algorithm of the proposed method. First, we show how to obtain the minimizer $\hat{\mathbf{W}}$ with given tuning parameters λ and α. Next, we illustrate how to select the tuning parameters. Finally, we explain how to recover the total causal structure.

3.3.1 How to Obtain Parameter Estimates

We derive an optimization algorithm based on the Alternating Direction Method of Multipliers [ADMM; 12]). ADMM is applied to the optimization of the objective function formulated as $f_1(x) + f_2(z)$ under a linear constraint on (x, z). In particular, ADMM is a powerful tool when one of the two functions is simple, hence ADMM is one of the standard optimization algorithms for sparse estimation problems [12, 31] including generalized lasso.

First, the optimization problem (3.13) is equivalently transformed to the ADMM form:

$$\begin{aligned} &\min_{\mathbf{W}\in\mathcal{N},\mathbf{M},\mathbf{P}\in\mathbb{R}^{d\times d}} F_1(\mathbf{W},\mathbf{P}) + F_2(\mathbf{M}) \\ &\text{subject to} \quad \mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{M},\ \mathbf{P}^T\mathbf{P} = \mathbf{I}, \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} F_1(\mathbf{W},\mathbf{P}) &= -\tilde{\ell}(\mathbf{W};\mathbf{Z}) + \lambda\frac{(1-\alpha)}{2}\|\mathbf{P}^T\mathbf{W}-\mathbf{I}\|_F^2, \\ F_2(\mathbf{M}) &= \lambda\alpha\mathcal{P}_\gamma(\mathbf{M}). \end{aligned}$$

Then, we update $(\mathbf{W},\mathbf{M})$ and $\mathbf{P}$ alternately. For given $\mathbf{W}$ (and $\mathbf{M}$), the updated matrix of $\mathbf{P}$ is obtained by minimizing $\|\mathbf{P}^T\mathbf{W}-\mathbf{I}\|_F^2$ under $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. Let $\mathbf{U}_W\mathbf{D}_W\mathbf{V}_W^T$ denote the singular value decomposition of $\mathbf{W}$. We see that the updated matrix of $\mathbf{P}$ is given by $\mathbf{P} = \mathbf{U}_W\mathbf{V}_W^T$. For given $\mathbf{P}$, we define the augmented Lagrangian as

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{W},\mathbf{M},\mathbf{U}) &= F_1(\mathbf{W},\mathbf{P}) + F_2(\mathbf{M}) \\ &+ \operatorname{tr}\left[\mathbf{U}^T(\mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T-\mathbf{M})\right] + \frac{\rho}{2}\|\mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T-\mathbf{M}\|_F^2, \end{aligned} \tag{3.15}$$

where $\mathbf{U}\in\mathbb{R}^{d\times d}$ is a Lagrange multiplier matrix and ρ is a fixed tuning parameter. From the optimality condition of ADMM, the updates of $\mathbf{W},\mathbf{M}$, and $\mathbf{U}$ are given by

$$\begin{cases} \mathbf{W}_{t+1} = \arg\min_{\mathbf{W}\in\mathcal{N}} \mathcal{L}_\rho(\mathbf{W},\mathbf{M}_t,\mathbf{U}_t) \\ \mathbf{M}_{t+1} = \arg\min_{\mathbf{M}\in\mathbb{R}^{d\times d}} \mathcal{L}_\rho(\mathbf{W}_{t+1},\mathbf{M},\mathbf{U}_t) \\ \mathbf{U}_{t+1} = \mathbf{U}_t + \rho(\mathbf{W}_{t+1}\mathbf{D}^{-1}\mathbf{V}^T - \mathbf{M}_{t+1}) \end{cases}. \tag{3.16}$$

In the following, the first and second updates are discussed in detail.

The first update in (3.16) is based on the gradient descent in Algorithm 1. Since the parameter space of $\mathbf{W}$ is restricted to the space of non-singular matrices, the gradient descent can be improved by incorporating the structure of the parameter space.

Algorithm 1 Gradient Descent for $\mathbf{W}$

1: **Input:**
$\mathbf{W}_t$, $\mathcal{L}_\rho(\mathbf{W}, \mathbf{M}_t, \mathbf{U}_t)$, $u_{\max}$, and learning rate $\eta > 0$
2: **Output:** $\mathbf{W}_{t+1}$

3: Calculate the natural gradient $\Delta\tilde{\mathbf{W}}$ and its modification $\Delta_{mod}\mathbf{W}$.
4: **for** $u = 1$ to $u_{\max}$ **do**
5: $\quad \mathbf{W}_t^{(u+1)} \leftarrow \mathbf{W}_t^{(u)} - \eta\Delta_{mod}\mathbf{W}$
6: $\quad$ **break** if the convergence criteria for $\mathbf{W}$ is satisfied.
end for
7: Update $\mathbf{W}$ as $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t^{(u_{\max})}$
8: $\boldsymbol{w}_{j,t+1} \leftarrow \boldsymbol{w}_{j,t+1}/\|\boldsymbol{w}_{j,t+1}\|$ for all j

We use the natural gradient [2], given by

$$\begin{aligned}\Delta\tilde{\mathbf{W}} &= \frac{\partial\mathcal{L}_\rho(\mathbf{W}, \mathbf{M}_t, \mathbf{U}_t)}{\partial\mathbf{W}}\mathbf{W}^T\mathbf{W} \\ &= -\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\tilde{g}(\boldsymbol{w}_j^T\boldsymbol{z}_i)\boldsymbol{w}_j^T\boldsymbol{z}_i + \mathbf{I}\right)\mathbf{W} \\ &\quad + \lambda\frac{(1-\alpha)}{2}(\mathbf{W} - \mathbf{P})\mathbf{W}^T\mathbf{W} \\ &\quad + \left\{\mathbf{U}_t + \rho(\mathbf{W}\mathbf{D}^{-1}\mathbf{V}^T - \mathbf{M}_t)\right\}\mathbf{V}\mathbf{D}^{-1}\mathbf{W}^T\mathbf{W}.\end{aligned} \tag{3.17}$$

The updated matrix using the natural gradient on the non-singular matrix space is usually not in $\mathcal{N}$. Thus it will be more efficient to make a gradient such that the updated matrix is in $\mathcal{N}$. Consider a small change $\mathbf{W} + \varepsilon\Delta\mathbf{W}$ ($\varepsilon > 0$). When it is in $\mathcal{N}$, the diagonal entries of $(\mathbf{W} + \varepsilon\Delta\mathbf{W})(\mathbf{W} + \varepsilon\Delta\mathbf{W})^T$ must be one. Here we ignore a very small term related to ε^2, which yields $\mathrm{diag}(\mathbf{W}\Delta\mathbf{W}^T) = \mathbf{0}$. Let $\Delta\mathbf{W} = (\delta\boldsymbol{w}_1, \ldots, \delta\boldsymbol{w}_d)^T$ with

$$\delta\boldsymbol{w}_j = \delta\tilde{\boldsymbol{w}}_j - \frac{\langle\boldsymbol{w}_j, \delta\tilde{\boldsymbol{w}}_j\rangle}{\|\boldsymbol{w}_j\|^2}\boldsymbol{w}_i \quad \text{for } j = 1, 2, \ldots, d. \tag{3.18}$$

Then we can easily see $\mathrm{diag}(\mathbf{W}\Delta\mathbf{W}^T) = \mathbf{0}$. Let $\Delta_{mod}\mathbf{W}$ be the gradient satisfying (3.18), and we use it in Algorithm 1 instead of the ordinary natural gradient $\Delta\tilde{\mathbf{W}}$. Rigorously, the updated matrix is slightly out of $\mathcal{N}$. Hence, at the 8th step in Algorithm 1, it is pulled back to $\mathcal{N}$.

Although the first update in (3.16) requires the minimization of $\mathcal{L}_\rho(\mathbf{W}, \mathbf{M}_t, \mathbf{U}_t)$ at every step, it is computationally heavy. Instead, we set a fixed upper limit $u_{\max}$ on the number of iterations to make the algorithm fast. From our experience, even if $u_{\max}$ is not large, the algorithm converges well.

The second update in (3.16) can be expressed in a closed form. The subgradient equation of $\mathcal{L}_\rho(\mathbf{W}_{t+1}, \mathbf{M}, \mathbf{U})$ with respect to $\mathbf{M}$ is

$$\lambda\alpha\partial_{\mathbf{M}}\mathcal{P}_\gamma(\mathbf{M}) - \mathbf{U}_t - \rho(\mathbf{W}_{t+1}\mathbf{D}^{-1}\mathbf{V}^T - \mathbf{M}) = O. \tag{3.19}$$

From this equation, the update for $\mathbf{M}$ takes the form

$$\mathbf{M}_{t+1} = \mathcal{S}\left(\mathbf{W}_{t+1}\mathbf{D}^{-1}\mathbf{V}^T + \frac{1}{\rho}\mathbf{U}_t \ ; \ \frac{\lambda\alpha}{\rho}\mathbf{C}_\gamma\right), \tag{3.20}$$

where $\mathbf{C}_\gamma = (c_{jk}^\gamma)$, and $\mathcal{S}(\cdot)$ is the soft-thresholding operator given by

$$\{\mathcal{S}(X;\mathbf{C})\}_{jk} = \begin{cases} X_{jk} - c_{jk} & (X_{jk} > c_{jk}) \\ 0 & (-c_{jk} \leq X_{jk} \leq c_{jk}) \\ X_{jk} + c_{jk} & (X_{jk} < -c_{jk}) \end{cases} .$$

ADMM usually has one stopping criterion related to the linear constraint, but the proposed algorithm additionally requires another one because we update $\mathbf{W}$ by gradient descent. Detailed criteria are described in the numerical experiments.

3.3.2 Tuning Parameter Selection

The tuning parameter λ and α can be selected by K-fold CV. Since $\lambda(1-\alpha)$ imposes orthogonality, and α balances sparsity and orthogonality, it is necessary to pay attention to the search range. The parameter λ must be large to some extent, and $0 \leq \alpha < 1$. In this strategy, the tuning parameter for the sparsity, $\lambda\alpha$, can be taken from zero to a large value.

Many sparse estimation methods search for the tuning parameter in descending order [31], but our algorithm does in ascending order. There are two reasons to this point. One reason is that it is impossible to obtain a maximum of $\lambda\alpha$. For example, in lasso [74], the maximum tuning parameter is obtained so that all parameter estimates are shrunk to zero. In contrast, our methods expect $\hat{\mathbf{M}}$ to be transformed to $\hat{\mathbf{B}}$, which means $\hat{\mathbf{M}}$ must have at least d non-zero elements. Another reason is that we have often encountered that the algorithm falls into an inappropriate local minimum with a large $\lambda\alpha$.

There is another remark on K-fold CV in the proposed method. We usually use the average of the log-likelihood paths over the K folds, and we choose the α which maximizes the averaged path. However, we observed that a log-likelihood path rarely changed more steeply than other paths. In that case, the averaged path is influenced by the irregular one. For a robust selection of α, we obtain K points of α that give

the maximum value of each path and then take the median of them.

3.3.3 Post-Processing

We obtain the estimate $\hat{\mathbf{W}}$ from the optimization problem (3.13). However, we cannot estimate $\mathbf{B}$ directly via the relation $\hat{\mathbf{B}} = \mathbf{I} - \hat{\mathbf{M}}$ because of two problems: (i) rows of $\hat{\mathbf{M}}$ have to be rearranged and rescaled so that all diagonal entries are one, and (ii) the estimate $\hat{\mathbf{B}}$ may not be acyclic, even if the true $\mathbf{B}$ is acyclic. These problems can be solved by the method proposed in [64]. Here, we only show the outline of this method.

The problem (i) is solved as follows. First, in order to obtain the matrix with non-zero diagonal entries, we search for a row permutation π minimizing $\sum_{j=1}^{d} 1/|\{\pi(\hat{\mathbf{M}})\}_{jj}|$. Next, we rescale the diagonal entries of $\pi(\hat{\mathbf{M}})$ to be one and divide each row by the same scale.

The problem (ii) is solved as follows. To ensure the acyclicity, we repeatedly apply a test-and-cutoff procedure:

1. Test whether $\hat{\mathbf{B}}$ is acyclic or not.
2. If $\hat{\mathbf{B}}$ is not acyclic, replace the non-zero smallest absolute value of $\hat{\mathbf{B}}$ to 0, and return to step 1.

For the test of acyclicity, an efficient algorithm was proposed by [64].

There is an additional device in the proposed method. After $\hat{\mathbf{B}}$ is made acyclic, the cutoff threshold is obtained. When this value is larger than pre-specified criteria $\omega_1 (> 0)$, such as $\omega_1 = 0.05$, we can improve the estimate by increasing α and again estimating $\hat{\mathbf{M}}$ until $\hat{\mathbf{B}}$ is made acyclic with a smaller cutoff than ω_1. This truncation seems ad hoc, but relevant methods like GOLEM set larger criteria like $\omega_1 = 0.3$. The proposed method worked well with much smaller cutoffs in numerical experiments as shown in Figure 3.4. Furthermore, we can reduce the false discovery of directed edges by additional cutoff $\omega_2 > 0$ for which the entries of $\hat{\mathbf{B}}$ is truncated to 0 if their absolute values are smaller than ω_2.

Figure 3.1 shows the flowchart of sICA-LiNGAM. If we select the tuning parameters via CV, we have to add the CV-step before calculating $\mathbf{M}$. The increment step of α can be skipped.

3.4 Experiments

In this section, we show the performance of the proposed method by numerical experiments. On synthetic data, the proposed method (sICA-LiNGAM) is compared with ICA-LiNGAM, DirectLiNGAM, the method of Zhang et al.(2009) [88]

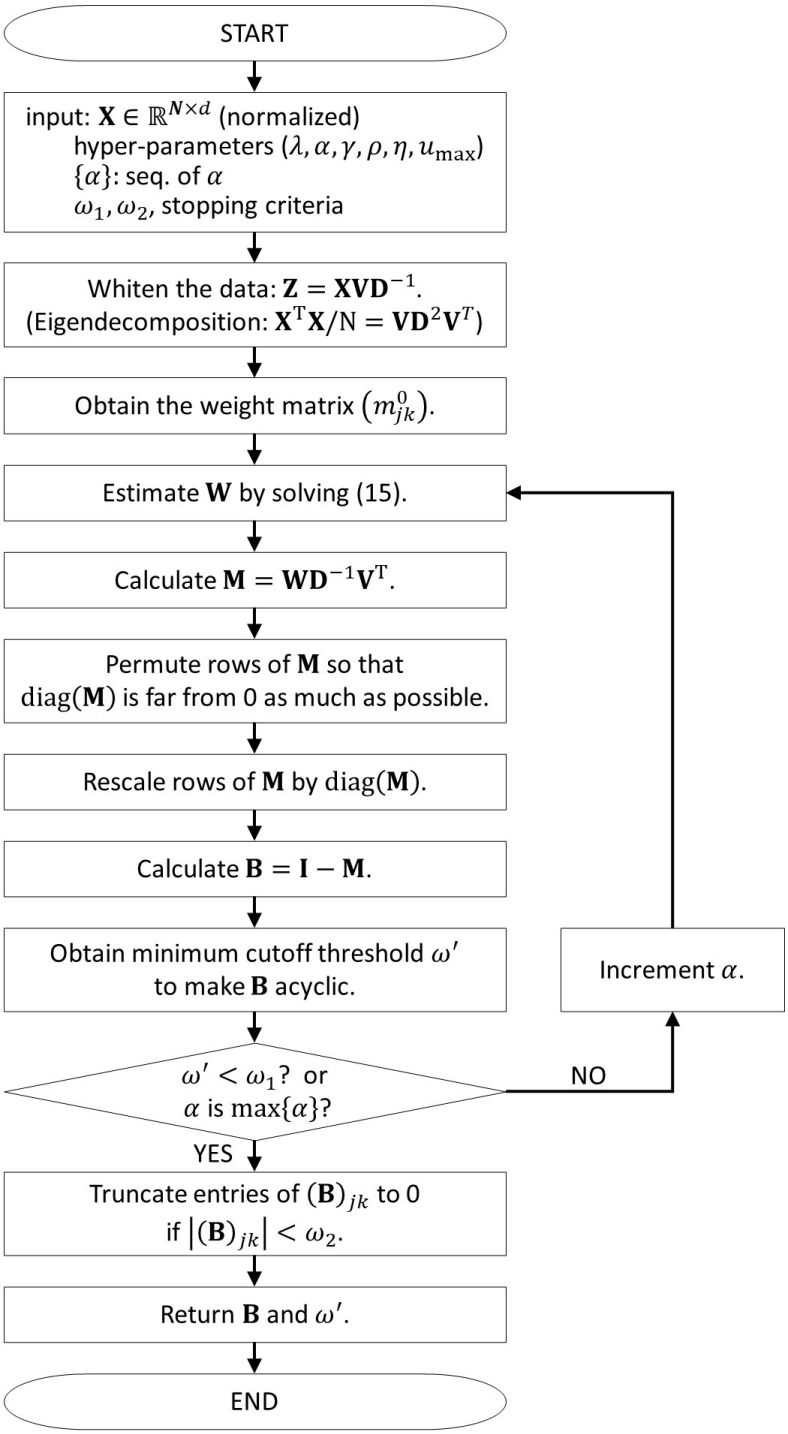

Figure 3.1: Flowchart of sICA-LiNGAM

(Zhang+(2009)), NOTEARS, and two GOLEMs. We do not include Pairwise LiNGAM and High-dimensional LiNGAM because the first two methods would show similar performance to DirectLiNGAM in our setting, as seen in [39, 81]. The scalability of the proposed method is examined in high dimensional data up to $d = 500$ with a fixed sample size. The application to real data is also conducted.

3.4.1 Comparison of the Methods on Synthetic Data

Numerical experiments are conducted on synthetic data. We basically mimic the simulation settings of [65, 91]. Observed variables are generated from a linear DAG model with non-Gaussian noises. True graphs are generated from *Erdös-Rényi* [ER; 20] or *Scale-Free* [SF; 6] models of different sizes ($d = 10, 20, 50, 100$), whose expected number of edges are d or $2d$. The graph type and the number of edges are denoted by ER1, ER2, SF1, and SF2. The weight parameters of the generated graph are uniformly drawn from the interval $[-1.5, -0.5] \cup [0.5, 1.5]$. (The true non-zero weight parameters are avoided to be around zero.) The distribution of each noise is randomly selected from three non-Gaussian distributions (*Laplace, uniform*, and *exponential*). The noise variances are uniformly drawn from the interval $[1, 3]$. We generate 10 datasets with a sample size of $N = 1,000$. The noises are randomly drawn from the selected distributions.

The models are estimated by sICA-LiNGAM, ICA-LiNGAM, DirectLiNGAM, Zhang+(2009), NOTEARS, GOLEM-EV, and GOLEM-NV. We describe the settings for these methods. For the proposed method, the tuning parameter λ and α are selected by 5-fold CV. The search range for λ is $[0.1, 0.25, 0.5, 0.75, 1.0]$, and that of $\log_{10} \alpha$ is set to
$[0, -4, -3.5, -3, -2.5, -2, -1.5, -1, -0.5]$. The initial estimate ($m_{jk}^0$) is obtained by the proposed method with $\gamma = 0$, $\alpha = 0$ for $d \leq 50$, and $\alpha = 0.01$ for $d \geq 100$. The step size is set at $\eta = 0.001$. If the resulting estimate $\hat{\mathbf{B}}$ is not acyclic, we compare two policies: (i) the initial tuning parameters (λ, α) are selected via CV, and increment α until $\hat{\mathbf{B}}$ became acyclic (sICA-LiNGAM(CV+)), (ii) the selected (λ, α) is used even if the cutoff threshold is larger than ω_1 (sICA-LiNGAM(CV)). The two cutoff parameters ω_1 and ω_2 are both set at 0.05. Two stopping criteria for parameter updates are set to $\max |\mathbf{W}_t \mathbf{D}^{-1} \mathbf{V}^T - \mathbf{M}_t| < 10^{-4}$ and $\max |\mathbf{W}_{t+1} - \mathbf{W}_t| < 10^{-6}$. Other tuning parameters are fixed at $u_{\max} = 10$ and $\rho = 1$. We use FastICA [37] results for the initial value except for $d = 100$. For $d = 100$, since we found FastICA was difficult to be converged, we used the result of DirectLiNGAM for the initial value for $\mathbf{W}$. The proposed method is implemented in Python 3.6.8. We also implemented Zhang+(2009). For other methods, the authors' implementations are

used [1]. ICA-LiNGAM and DirectLiNGAM do not require tuning. For Zhang+(2009) and NOTEARS, the tuning parameter defining sparsity is searched in the same range as that of α by 5-fold CV. Tuning of GOLEM is more complicated. GOLEM-EV and GOLEM-NV have three tuning parameters, $\lambda_{EV}, \lambda_{NV}$, and λ_{acyc}. We search for the optimal tuning parameters from 25 patterns, which are shown in the supplementary material. The parameter λ_{acyc}, which imposes acyclicity on the coefficient matrix, is fixed at 5. This value is used in the original literature [54]. Note that we did not conduct CV for NOTEARS and GOLEM on the data with $d \geq 50$ due to computation time. We use $\lambda_{NOTEARS} = 0.1$ for NOTEARS, $(\lambda_{EV}, \lambda_{acyc}) = (10^{-1.5}, 5)$ for GOLEM-EV, and $(\lambda_{EV}, \lambda_{NV}, \lambda_{acyc}) = (10^{-1.5}, 10^{-1}, 5)$ for GOLEM-NV. These values were once selected by 5-fold CV in our pilot experiment. The estimates of NOTEARS and GOLEM are truncated to reduce false positives if their absolute values are less than 0.3.

The estimation methods are evaluated in terms of estimation error based on the Frobenius norm between the estimated and true weight matrices (Distance), Structural Hamming Distance (SHD), False Discovery Rate (FDR) and True Positive Rate (TPR). The metrics are detailed in the supplementary material.

Figure 3.2 shows the main results on ER1 and SF1 graphs. The proposed method achieved the best performance among all methods. This result was probably because the proposed method efficiently used the information of the data as we intended. The proposed method succeeded in improving ICA-LiNGAM by virtue of sparsity and, in particular, largely improved in the high dimensional setting ($d = 100$). Zhang+(2009), NOTEARS, and GOLEM were behind the others. ICA-LiNGAM and Zhang+(2009) were omitted from the result of $d = 100$ due to seriously bad performance (full results are available in the supplementary material). The low performance of NOTEARS and GOLEM would be because these methods cannot leverage the non-Gaussianity assumption.

Figure 3.3 shows the results on ER2 and SF2 graphs. In this setting, the proposed method also achieved the best performance among all methods in almost all settings.

Figure 3.4 shows the actual cutoff threshold necessary to make $\hat{\mathbf{B}}$ acyclic at CV-selected parameter. Although we used $\omega_1 = 0.05$, the estimate $\hat{\mathbf{B}}$ could be acyclic with a much smaller cutoff threshold in most cases. Besides, the proposed method requires a much smaller cutoff threshold than other methods. This indicates that we can estimate the acyclic graph in a much less ad hoc manner. ICA-LiNGAM, DirectLiNGAM, and NOTEARS are excluded because the authors' implementations produce acyclic estimates directly.

[1] ICA- and DirectLiNGAM: https://github.com/cdt15/lingam, NOTEARS: https://github.com/xunzheng/notears, GOLEM: https://github.com/ignavier/golem

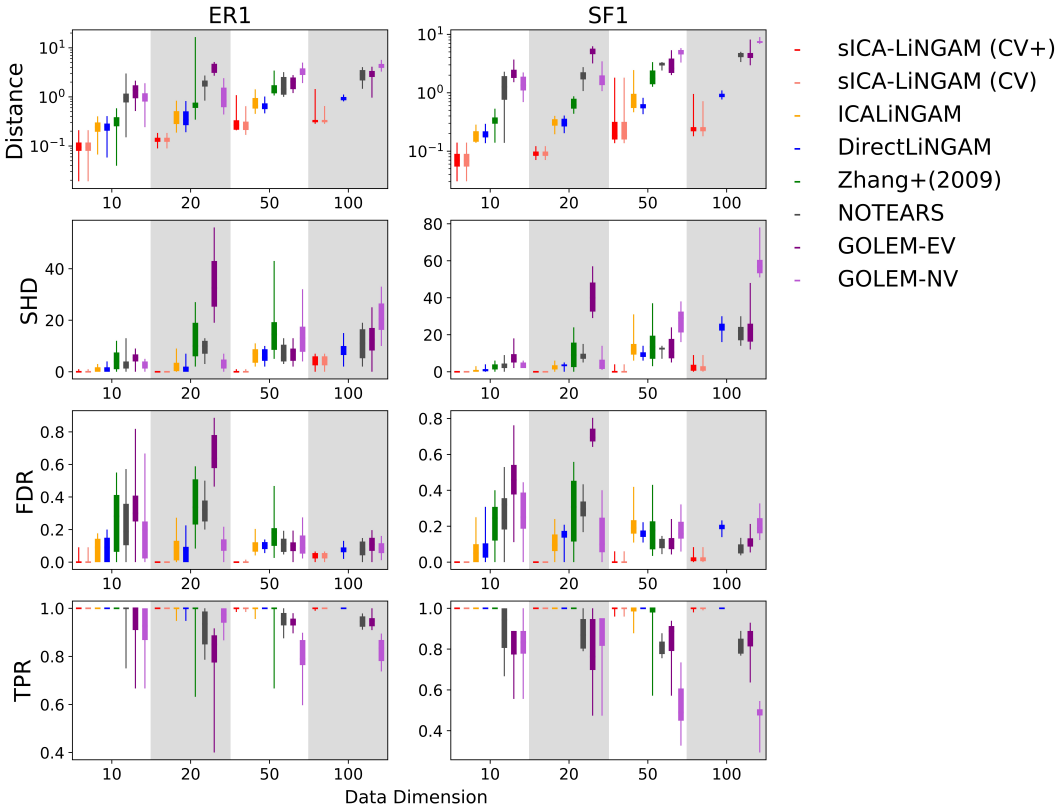

Figure 3.2: Four evaluation measures (Distance, SHD, FDR, TPR) over two graph types (ER1, SF1) and four graph sizes (10, 20, 50, 100). The x-axis is the graph size, and the y-axis is the value of each measure. The thick bar and the thin bar are the interquartile range and min-max range, respectively.

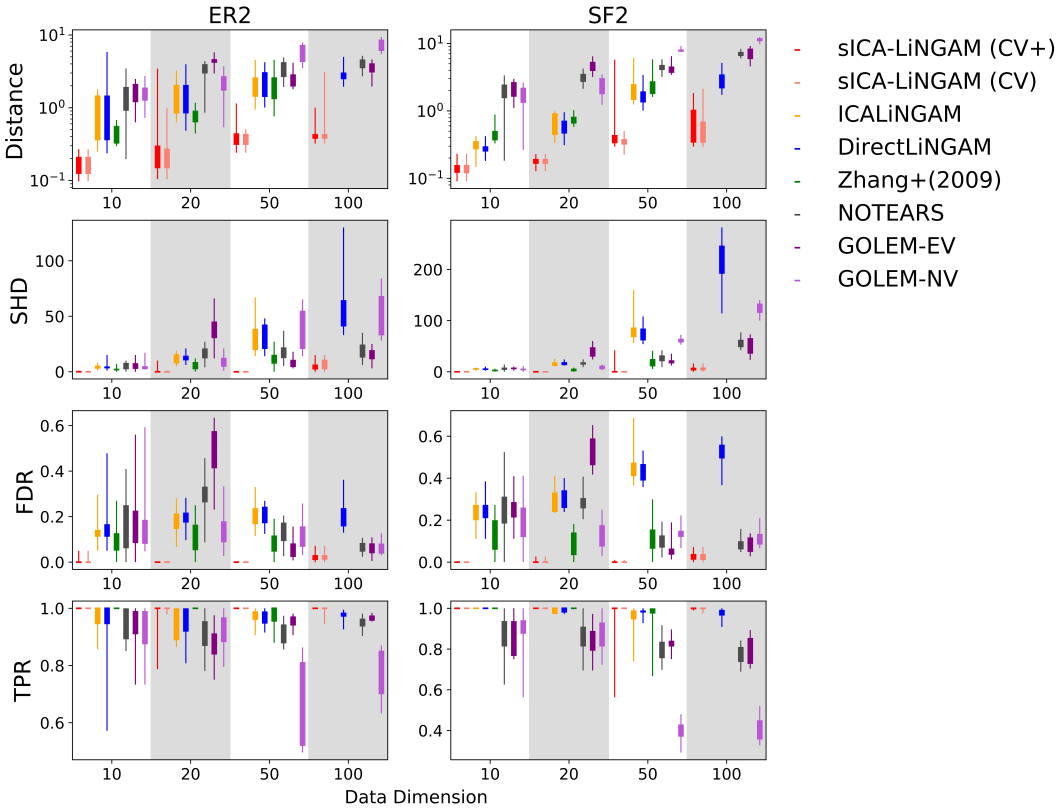

Figure 3.3: Four evaluation measures (Distance, SHD, FDR, TPR) over two graph types (ER2, SF2) and four graph sizes (10, 20, 50, 100). The x-axis is the graph size, and the y-axis is the value of each measure. The thick bar and the thin bar are the interquartile range and min-max range, respectively.

The threshold can be affected by the scale of the true $\mathbf{B}$. After investigating the threshold and scale sensitivities, we found that the thresholds were small enough even when each value of $\mathbf{B}$ was multiplied by 0.1. The results of the sensitivity study are found in the supplementary material.

3.4.2 Scalability of the proposed method

The scalability of the proposed method is examined in simpler and higher dimensional settings. The graphs are generated from the ER model with $d = 100, 200, 500$, and noises are drawn only from the *Laplace* distribution. The expected number of directed edges is 50 or 100. The sample size and other settings are the same as the previous experiment. We evaluate the estimation error and computational time of the proposed method and DirectLiNGAM because these methods performed well for $d = 100$ in the previous experiment. In order to evaluate the computational time conveniently, the tuning parameters of the proposed method were fixed at $(\lambda, \alpha) = (0.1, 0.1)$. The experiments are conducted on a single 3.6 GHz Intel Core i7 CPU and a 32GB memory.

The results of 10 simulations are shown in Figure 3.5. Our method performed well even in high-dimensional settings. On the other hand, the estimates of DirectLiNGAM were unstable when $d = 500$. Both methods finished in realistic computational time. In high dimensions, it seems difficult to reduce false discovery by truncation with a small threshold like $\omega_2 = 0.05$, but the proposed method did not overlook any causal relationships even in the $d = 500$ case.

3.4.3 Real Data

As described in [64], a time series can be approximated by a linear DAG model, especially if the time series is a stationary autoregressive model of order 1 (AR(1)). For example, when a time series is sliced into time windows with three time points, the AR(1) structure can be approximated by the linear DAG model of

$$\begin{cases} X_1 &= \varepsilon_1 \\ X_2 &= b_{21} X_1 + \varepsilon_2 \\ X_3 &= b_{32} X_2 + \varepsilon_3 \end{cases} . \tag{3.21}$$

Therefore, if the noises are independent and non-Gaussian, LiNGAM can express the AR(1) structure and then recover the correct order from the sliced time series.

We applied the four methods to the Beijing Multi-Site Air-Quality Data [89]. This data includes hourly concentration measures of major air pollutants, such as nitrogen dioxide (NO_2) and sulfur dioxide (SO_2), recorded at national monitoring

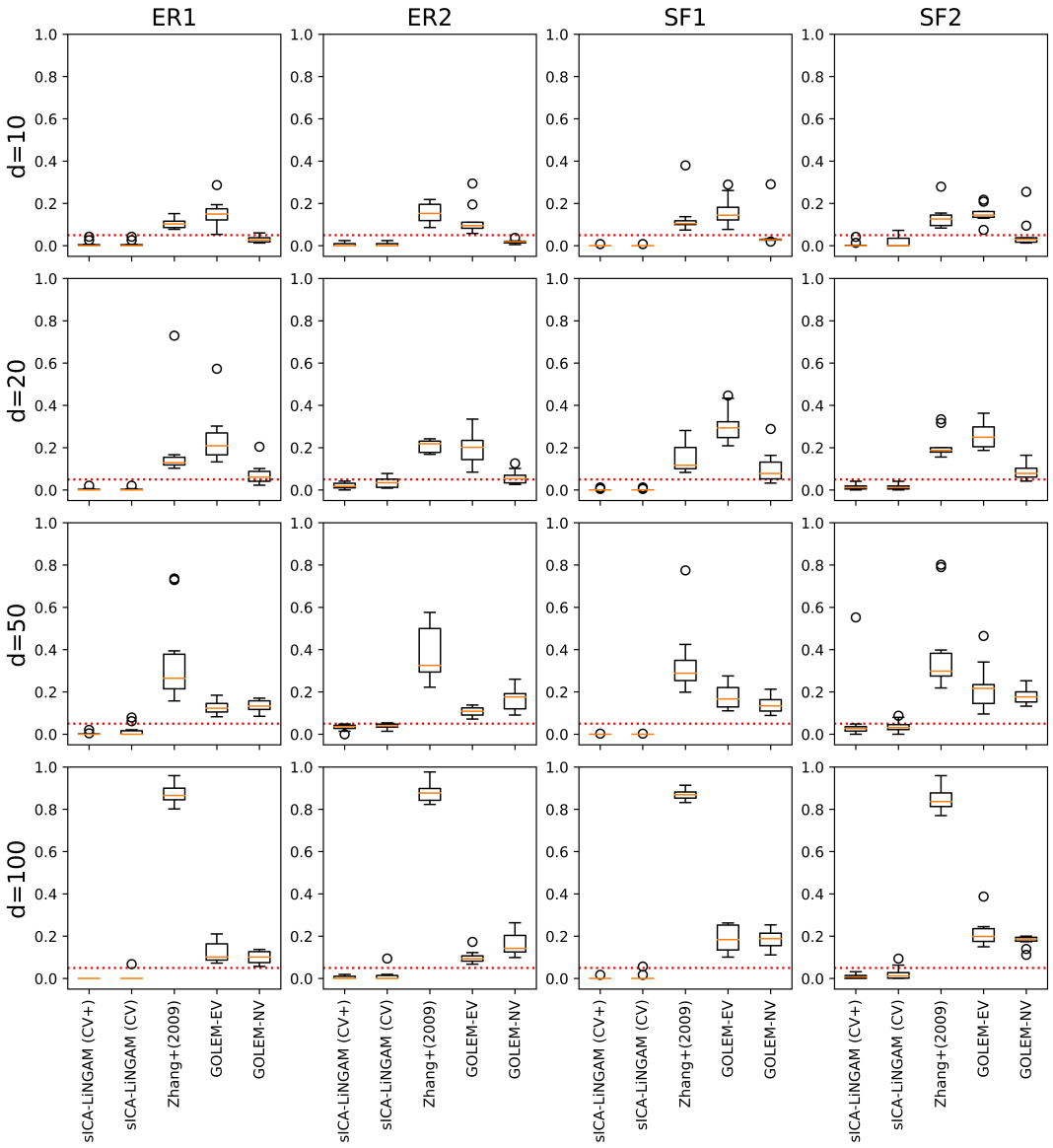

Figure 3.4: Box plots of cutoff thresholds necessary to make $\hat{\mathbf{B}}$ acyclic at CV-selected α. Each unit contains 10 results. The lower and the upper edges of each box are quartile 1 (Q1) and quartile 3 (Q3), and the orange lines are the medians. Circles are outliers, which are out of the range $(Q1 - 1.5 \cdot \mathrm{IQR}, Q3 + 1.5 \cdot \mathrm{IQR})$. Please see the reference of Matplotlib 3.1.1 for details.

If the cutoff exceeded the criteria (red dotted line), we increased α.

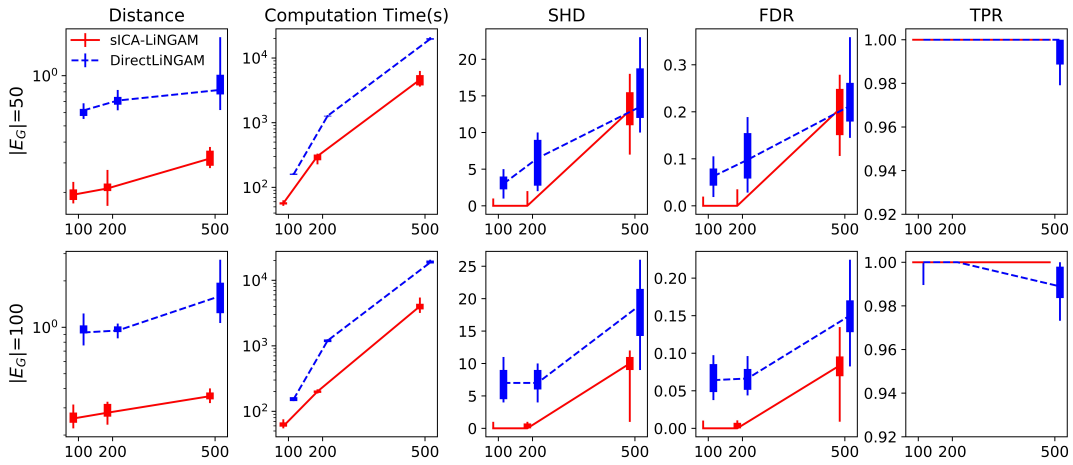

Figure 3.5: Estimation error and computation time based on 10 simulations with $\omega_2 = 0.05$. The x-axis is the dimension of the data.

sites in Beijing. Each site provides every pollutant's time series from March 1st, 2013 to February 28th, 2017. Each time series was sliced into 1,461 time windows with 24 time points so that each window consists of hourly measures of one day. The values were transformed by the function $\log(1 + x)$. When a time window contained missing values, the window was removed. Every method was evaluated by a heatmap visualizing $\hat{\mathbf{B}}$. Suppose the AR(1) structure is recovered by the linear DAG model, only the $(j+1, j)$th cell is colored for $j = 1, ..., 23$, and the other cells are not colored. Note that the proposed method used the CV-selected α because it was difficult to obtain the acyclic estimates by increasing α.

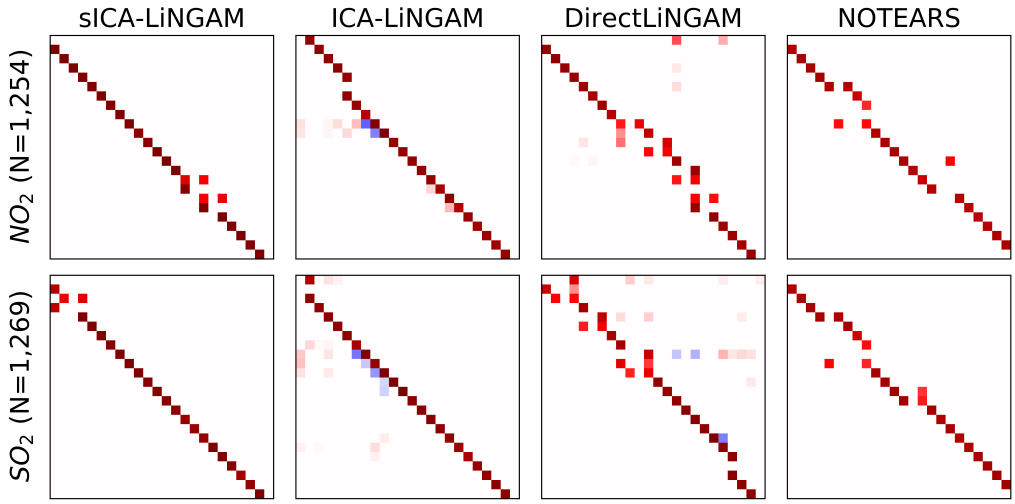

Figure 3.6: Heatmaps of $\hat{\mathbf{B}}$s at Tiantan station. Red/Blue indicates a positive/negative value. If the AR(1) structure is recovered correctly, the $(j+1, j)$th cell is colored for $j = 1, ..., 23$ and the other cells are not colored.

Figure 3.6 shows the results of the data at Tiantan station. The proposed method succeeded in recovering the AR(1) structure very well, and non-AR(1) cells were almost shrunk to zero. ICA-LiNGAM also recovered the structure, but some non-AR(1) cells had non-zero values. By virtue of sparsity, the proposed method shrunk the small non-zero estimates of non-AR(1) cells to zero and recovered the AR(1) structure better than ICA-LiNGAM. DirectLiNGAM and NOTEARS failed to recover about or more than half of the AR(1) structure. As seen in the supplementary material, the proposed method also showed the best performance clearly at the other monitoring sites.

3.5 Conclusion

In this work, we have proposed a new estimating algorithm for a linear DAG model with non-Gaussian noise. The proposed method is based on the penalized log-likelihood of ICA and estimates the causal structure and the parameter values based on a single statistical criterion. Several devices for stable and efficient learning are introduced, such as a penalty on the orthogonality of the parameter matrix and the modified natural gradient. The proposed method achieved the best performance among the existing methods in the numerical experiments. For future work, it is significant to extend the method to non-DAG structures, such as data with cyclic structures and/or latent confounders.

3.6 Additional Sources for sICA-LiNGAM

3.6.1 Pattern for Tuning Parameters of GOLEM

Table 3.6.1 shows all triples of tuning parameters of GOLEM ($\log_{10}\lambda_{EV}, \log_{10}\lambda_{NV}, \lambda_{acyc}$) in Section 5.1.

Table 3.1: The search pattern for tuning parameters of GOLEM

$\log_{10}\lambda_{EV}$	$\log_{10}\lambda_{NV}$	λ_{acyc}
$-\infty$	$-\infty$	5
-4.0	-4.5	5
-4.0	-4.0	5
-4.0	-3.5	5
-3.5	-4.0	5
-3.5	-3.5	5
-3.5	-3.0	5
-3.0	-3.5	5
-3.0	-3.0	5
-3.0	-2.5	5
-2.5	-3.0	5
-2.5	-2.5	5
-2.5	-2.0	5
-2.0	-2.5	5
-2.0	-2.0	5
-2.0	-1.5	5
-1.5	-2.0	5
-1.5	-1.5	5
-1.5	-1.0	5
-1.0	-1.5	5
-1.0	-1.0	5
-1.0	-0.5	5
-0.5	-1.0	5
-0.5	-0.5	5
-0.5	0.0	5

3.6.2 Definitions of Evaluation Measures

We evaluated the estimates $\hat{\mathbf{B}}$ and the corresponding graph by four metrics: 1) *Distance*, 2) *Structural Hamming Distance* (SHD), 3) *False Discovery Rate* (FDR), 4) *True Positive Rate* (TPR). The Distance is a measure for the estimation error. The other three measures are employed to evaluate the performance of causal discovery.

- *Distance (Frobenius Norm of the Difference Between Two Matrices)* : Distance is defined as the Frobenius norm of the difference between two matrices. The

Distance between the estimate and the truth is evaluated:

$$Distance = \|\hat{\mathbf{B}} - \mathbf{B}_{\mathbf{true}}\|_F. \tag{3.22}$$

- *Structured Hamming Distance* (SHD): SHD indicates the number of steps to transform the estimated graph into the true graph. The steps include edge addition, deletion, and reversals.
- *False Discovery Rate* (FDR): FDR is the proportion of false positives and reversed edges over the estimated edges.
- *True Positive Rate* (TPR): TPR is the proportion of true positive edges over the true edges.

3.6.3 Full Results of the Experiment 5.1

Figure 3.7 and 3.8 show the full results of experiment 1 in Section 5.1.

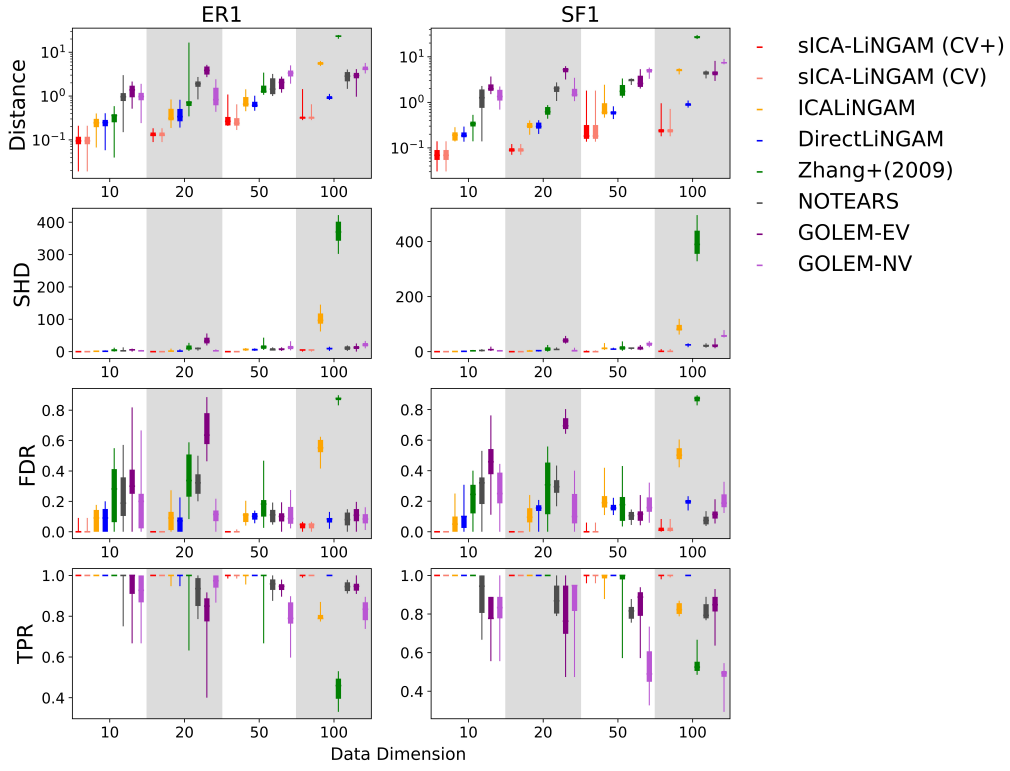

Figure 3.7: Four evaluation measures (Distance, SHD, FDR, TPR) over two graph types (ER1, SF1) and four graph sizes (10, 20, 50, 100). The x-axis is the graph size, and the y-axis is the value of each measure. The thick bar and the thin bar are the interquartile range and min-max range, respectively.

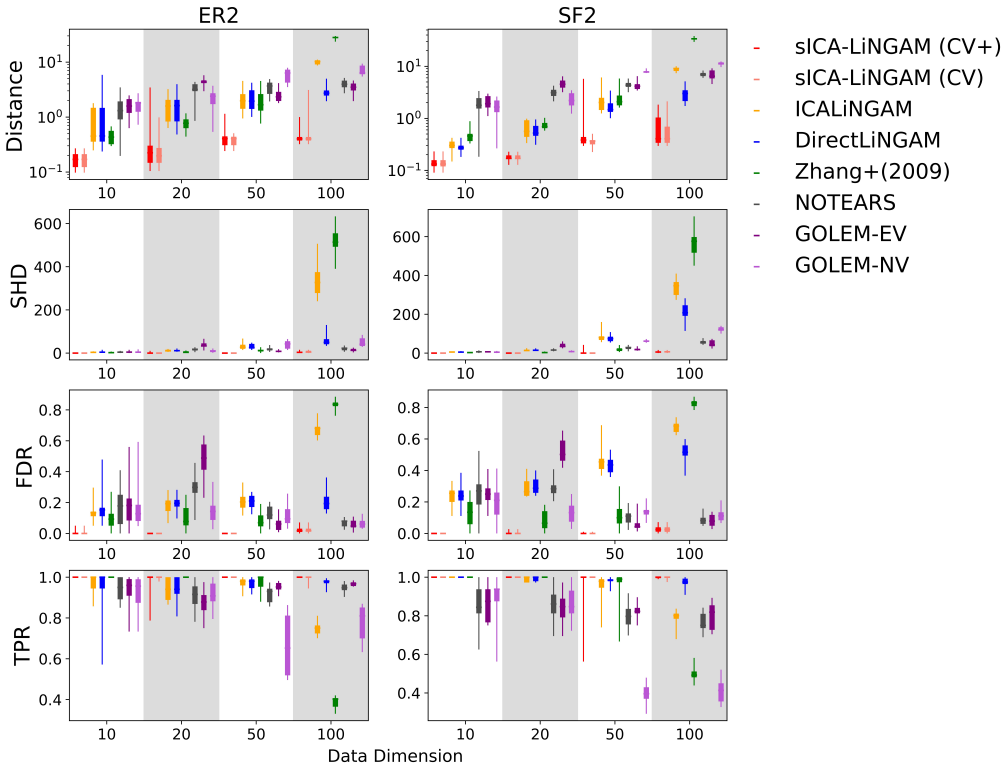

Figure 3.8: Four evaluation measures (Distance, SHD, FDR, TPR) over two graph types (ER2, SF2) and four graph sizes (10, 20, 50, 100). The x-axis is the graph size, and the y-axis is the value of each measure. The thick bar and the thin bar are the interquartile range and min-max range, respectively.

3.6.4 Sensitivity Study on Cutoff Threshold

We investigated the sensitivity of the cutoff threshold to obtain acyclic $\mathbf{B}$. Let $\kappa \in \{0.1, 0.2, 0.5, 0.75, 1.0\}$ be the scale factor, and each value of $\mathbf{B}$ is multiplied by κ. Then we generated 10 datasets for each κ. The graph types are ER1 and ER2, and the dimension is $d = 20$. Figure 3.9 shows the cutoff threshold necessary to obtain acyclic $\mathbf{B}$. The threshold tends to large when the scale is small. It seems difficult to estimate when the true scale of $\mathbf{B}$ is small.

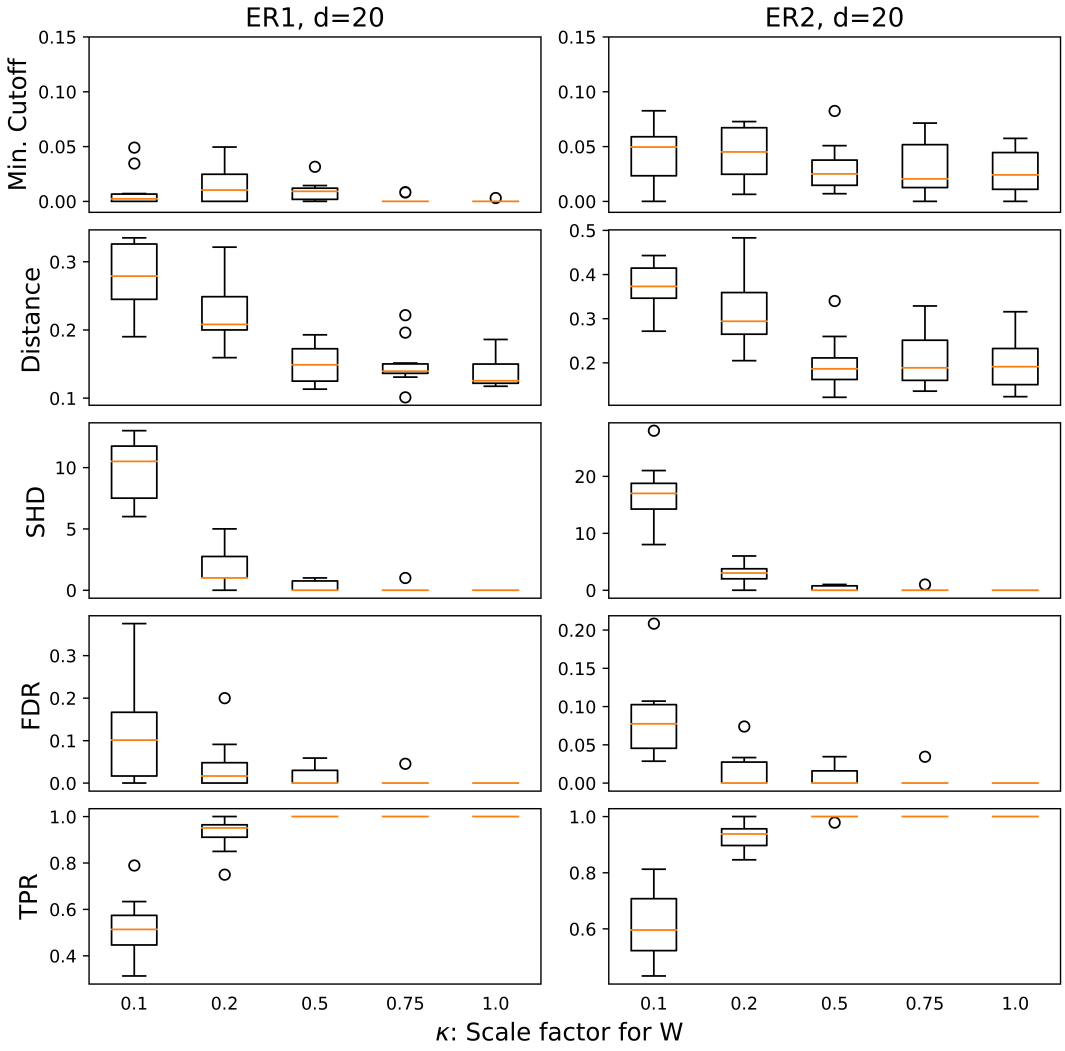

Figure 3.9: Cutoff threshold and four evaluation measures over two graph types (ER1, ER2). The x-axis is the scale factor κ, and the y-axis is the value of each measure. Each box shows 10 results.

3.6.5 AR(1) Recovery at the Other Sites

We estimated $\hat{\mathbf{B}}$ by four methods (sICA-LiNGAM (proposed), ICA-LiNGAM, DirectLiNGAM, NOTEARS) at three additional sites. If the AR(1) structure is recovered by the linear DAG model, only the $(j+1, j)$th cell is colored for $j = 1, ..., 23$, and the other cells are not colored.

- At Aotizhongxin station (Figure 3.10), the proposed method successfully recovered the time series, especially on SO_2. ICA-LiNGAM partly recovered the structure, but some non-AR(1) cells had non-zero values. DirectLiNGAM also recovered the structure on NO_2, but failed to recover about two third of the AR(1) structure on SO_2. NOTEARS failed to recover about or more than half of the AR(1) structure.

- At Nongzhanguan station (Figure 3.11), the proposed method completely recovered the AR(1) structure except for the points of (1,24) and (21,22) on SO_2. None of the other methods could recovered the AR(1) sequence successfully.

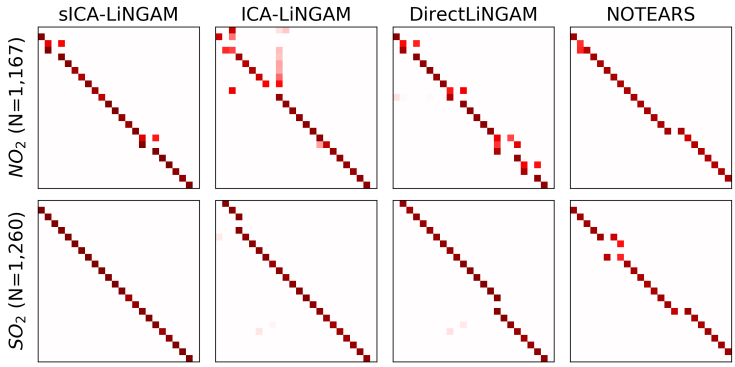

Figure 3.10: Heatmaps of $\hat{\mathbf{B}}$s at Aotizhongxin station. Red/Blue indicates a positive/negative value.

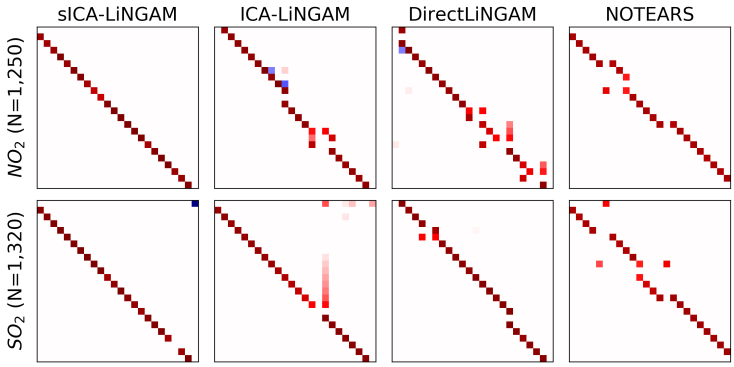

Figure 3.11: Heatmaps of $\hat{\mathbf{B}}$s at Nongzhanguan station. Red/Blue indicates a positive/negative value.

4 Outlier-resistant Estimation of ATE

4.1 Background

In this work, the goal is to estimate the ATE under contamination by a propensity score-based approach. Notations are taken from Section 2.2.

As introduced in Section 2, the IPW and DR methods are popular for the estimation of the ATE. Many estimators for causal quantities, including the IPW and DR estimators, are based on the sample mean and are therefore influenced by outliers. Outlier-resistant estimators have been studied for long; however, these are mostly applicable to non-causal settings [35, 30, 52]. They are not directly applicable to causal settings owing to confounding. The ATE can be estimated in an outlier-resistant manner only by using the causal version of the sample median [21, 90, 18, 72]. In other words, the existing ATE estimators have limited outlier resistance.

In this work, we propose outlier-resistant extensions of the IPW and DR estimators for the ATE whose outlier resistance is beyond the sample median. We discuss the outlier resistance of these estimators from the viewpoint of the unbiasedness of the estimating equation and influence function (IF). The theoretical assumptions we make, such as heterogeneous contamination and non-small contamination ratio, are generally challenging in outlier-resistant statistics. Nonetheless, our estimators can effectively reduce the bias caused by outliers even under these assumptions. In particular, our estimators are almost consistent with the true ATE under homogeneous contamination. No existing estimators for causal inference show outlier resistance beyond the sample median. In other words, ours are the first methods to overcome this problem. In order to tackle this problem, we incorporate a positively powered density function into the estimating equations of the IPW and DR estimators. Our results show that this density power approach is viable for outlier-resistant estimation of causal quantities. Furthermore, the theoretical advantages of our estimators are verified through Monte-Carlo simulations and real data analysis.

The remainder of Section 4 is organized as follows. In Section 4.2, we introduce the basic concept of outliers and expand it to a causal setting. In Section 4.3, we propose new estimators and discuss the outlier resistance from the viewpoint of the unbiasedness of the estimating equations. In Section 4.4, we evaluate the outlier resistance in terms of the IF. In Section 4.5, we discuss asymptotic properties. In Section 4.6, we present the numerical algorithms. Finally, in Sections 4.7 and 4.8, we present the experimental results.

4.2 Outlier-resistant Estimation

4.2.1 Non-causal Setting

This subsection provides a brief review on outlier-resistant estimation in a one-variable and non-causal setting. Let $\tilde{g}$ be the density function of a random variable $Z \in \mathbb{R}$. Assume that the density is contaminated as $\tilde{g}(z) = (1-\varepsilon)f_{\theta^*}(z) + \varepsilon\delta(z)$, where f_{θ^*} is the density of interest indexed by the parameter θ^*, ε is the contamination ratio, and δ is the density of outliers. Our goal is to estimate the parameter θ^* from *i.i.d.* observations $\{Z_1, ..., Z_n\}$. Let $\hat{\theta}_\psi$ be a root of $\sum_{i=1}^n \psi(Z_i, \theta) = 0$. This type of estimator is called an *M-estimator*. We assume the unbiasedness of the estimating equation:

$$\mathbb{E}_{f_\theta}[\psi(Z, \theta)] = 0. \tag{4.1}$$

The IPW, DR, and the proposed estimators are all M-estimators, and they satisfy the unbiasedness of the estimating equation under no contamination. If the estimating equation is unbiased and some regularity conditions hold, the M-estimator has consistency and asymptotic normality [79]. However, under contamination, we generally have $\mathbb{E}_{\tilde{g}}[\psi(Z, \theta^*)] \neq 0$. Let θ^*_ψ denote a root of $\mathbb{E}_{\tilde{g}}[\psi(Z, \theta)] = 0$; then, the latent bias is defined as $\theta^*_\psi - \theta^*$. We hope that the latent bias is small even under contamination. If δ is Dirac's delta and ε is sufficiently small, the latent bias is approximated by the IF. The IF-based discussion in Section 4.4 provides some insight into the outlier resistance of the estimators when the contamination ratio is small. The latent bias and M-estimators are discussed in detailed elsewhere [35, 30, 26, 25].

4.2.2 Causal Setting

Next, we consider a causal setting. In this work, we assume that only the outcome Y may be contaminated. Let $\delta_{Y|TX}$ be the conditional density of outliers given (T, X), then the contaminated conditional density given (T, X) is defined as

$$\tilde{g}_{Y|TX}(y|T, X) = (1 - \varepsilon(T, X))g_{Y|TX}(y|T, X) + \varepsilon(T, X)\delta_{Y|TX}(y|T, X), \tag{4.2}$$

where g denotes the density without contamination, and $\varepsilon(T, X)$ is the ratio of outliers. The tilde indicates that the distribution is contaminated. To simplify the notation, we often drop the subscripts of density functions as long as there would be no confusion. The ratio of the outliers ε and their density δ depend on the treatment T and the confounder X. Since we estimate $\mu^{(t)}$ for each treatment separately, the dependence on T is tractable. In contrast, the dependence on X is critical

in our analysis. The X-dependent contamination is referred to as heterogeneous contamination. We write $\varepsilon_t(x) = \varepsilon(t, x)$ and $\delta_t(y|x) = \delta_{Y|TX}(y|T = t, x)$. We also discuss the special case in which ε and δ are not dependent on X, called homogeneous contamination. Note that we do not assume $\varepsilon_t(x)$ to be small enough to be negligible, except in Section 4.4.

We are interested in the marginal mean of $Y^{(1)}$, and let $f_{Y^{(1)}}(y; \mu^{(1)})$ be the true marginal density of $Y^{(1)}$. It is obtained by integrating X out from $g_{Y|TX}(y|T, X)$ under $T = 1$:

$$f_{Y^{(1)}}(y; \mu^{(1)}) = \int g_{Y^{(1)}|X}(y|x) g_X(x) dx = \int g_{Y|TX}(y|T = 1, x) g_X(x) dx. \tag{4.3}$$

The second equality holds from the causal consistency and the exchangeability assumption. We often write $f_{Y^{(1)}}(y; \mu^{(1)})$ as $f_1(y)$ for simplifying the notation.

Under contamination, the IPW estimating equation is severely biased because the conditional expectation $\mathbb{E}_{-g+\delta}[\cdot|X]$ is accompanied by the density of outliers:

$$\mathbb{E}_{\tilde{g}}\left[\frac{T}{\pi(X|\alpha^*)}(Y - \mu^{(1)})\right] = \mathbb{E}_g\left[\varepsilon_1(X)\mathbb{E}_{-g+\delta}\left[(Y - \mu^{(1)})|X\right]\right] \neq 0. \tag{4.4}$$

The DR estimating equation is similarly biased. To estimate $\mu^{(1)}$ accurately, we have to remove the influence of contamination.

4.3 Proposed Methods

4.3.1 Assumptions on Outliers

Below, we assume that the true marginal density $f_1(y)$ is symmetric about $\mu^{(1)}$. This is a common assumption in outlier-resistant estimation, and it is also a prerequisite to use the sample median as an estimator for the population mean.

Let $h(y; \mu)^\gamma$ ($\gamma > 0$) be a density power weight for $Y^{(1)}$, where $h(y; \mu)$ is a symmetric density function with the location parameter μ. The density $h(y; \mu^{(1)})$ is not necessarily equal to the true marginal density $f_1(y)$. Any symmetric density is suitable for $h(y; \mu)$ if it satisfies Assumption 1 below. Typically, we assume a Gaussian density. The density power weight is used to enhance the outlier resistance in noncausal settings [82, 7, 43, 25]. The tuning parameter γ controls the variability of the weight; this leads to the trade-off between outlier resistance and asymptotic efficiency. Before we propose novel estimators, we introduce an assumption to remove the influence of outliers. Suppose that the outliers are sufficiently far from the weighting distribution of $Y^{(1)}$. Then, we use the following assumption.

Assumption 1. *Let $h(y;\mu)$ be a weighting density symmetric about μ. Then, there exists $\gamma > 0$ such that*

$$\xi_1(X) = \int \delta_1(y|X) h(y;\mu^{(1)})^\gamma (y - \mu^{(1)}) dy \approx 0 \quad a.e. \tag{4.5}$$

This assumption implies

$$\nu_1(\phi) := \mathbb{E}[\phi(X)\xi_1(X)] = \int \phi(x)\xi_1(x)g(x)dx \approx 0, \tag{4.6}$$

for any bounded function $\phi(x)$. In particular, let $\phi(x) = 1$; then, the outliers are marginally negligible:

$$\nu_1(1) = \mathbb{E}[\xi_1(X)] = \int \delta_1(y) h(y;\mu^{(1)})^\gamma (y - \mu^{(1)}) dy \approx 0. \tag{4.7}$$

Throughout this work, we assume that γ is sufficiently large so that Assumption 1 holds.

Furthermore, Assumption 1 is reduced to a simpler form when $\delta_1(y|X)$ is Dirac's delta at y_0; this is one of the core assumptions in Section 4.4.

Assumption 1'. *Let $h(y;\mu)$ be a weighting density that is symmetric about μ, and assume that the density of outliers is Dirac's delta at y_0 that is sufficiently far from $\mu^{(1)}$. Then, there exists $\gamma > 0$ such that*

$$\int \delta_{y_0}(y) h(y;\mu^{(1)})^\gamma (y - \mu^{(1)}) dy = h(y_0;\mu^{(1)})^\gamma (y_0 - \mu^{(1)}) \approx 0 \tag{4.8}$$

For example, if $h(y;\mu^{(1)})$ is a Gaussian density with mean $\mu^{(1)}$, the condition (4.8) holds since y_0 is sufficiently far from $\mu^{(1)}$. All proofs for the theorems in this section are provided in Section 4.9.

4.3.2 DP-IPW Estimator

First, we introduce an extension of the IPW estimator, called the density-powered inverse probability weighting (DP-IPW) estimator. The DP-IPW estimator is defined as a root of the following estimating equation:

$$\sum_{i=1}^{n} \frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})} (Y_i - \mu) = 0. \tag{4.9}$$

Under no contamination, the DP-IPW estimating equation is unbiased by the following theorem.

Theorem 4.1. *Assume that $h(y;\mu^{(1)})$ and $f_1(y)$ are both symmetric about $\mu^{(1)}$ and that the true propensity score $\pi(X;\alpha^*)$ is given. Then, under no contamination, we have*

$$\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\right]=0. \tag{4.10}$$

Although only an estimate $\pi(X;\hat{\alpha})$ is available in practice, the asymptotic consistency of (DP-)IPW still holds if the model $\pi(X;\alpha)$ is correctly specified.

Now we consider the contaminated case. Suppose the conditional density of Y is (4.2). Then, the bias of the estimating equation takes a different form from (4.4).

Theorem 4.2. *Suppose Y is contaminated as* (4.2)*. Under the same assumptions as those in Theorem 4.1, the expectation of the DP-IPW estimating equation is expressed as*

$$\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\right]=-\int\varepsilon_1(x)\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1(\varepsilon_1). \tag{4.11}$$

In particular, if the contamination ratio is independent of X, the right-hand side of (4.11) *reduces to $\nu_1(\varepsilon_1)$.*

Thus, the DP-IPW estimating equation is almost unbiased under homogeneous contamination since we assume that $\nu_1(\varepsilon_1)$ is negligible. Under heterogeneous contamination, it is biased because the non-negligible first term of (4.11) remains. However, compared to (4.4), the dominant bias of DP-IPW does not contain δ_1. This implies that the bias of DP-IPW is not strongly affected by the absolute value of outliers. Furthermore, if the contamination ratio is small, the first term can also be small.

4.3.3 DP-DR Estimator

Next, we introduce the density-powered doubly robust (DP-DR) estimator. This is a special case of the doubly robust M-estimator [77, 76, 34]. The DP-DR estimator is defined as a root of the following estimating equation:

$$\sum_{i=1}^n\left\{\frac{T_ih(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu)-\frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}\left\{m_{1,\mu}(X_i;\hat{\beta})-\mu m_{0,\mu}(X_i;\hat{\beta})\right\}\right\}=0, \tag{4.12}$$

where $m_{0,\mu}(X;\hat{\beta})$ and $m_{1,\mu}(X;\hat{\beta})$ are the estimators for $\mathbb{E}_g[h(Y^{(1)};\mu)^\gamma|X]$ and $\mathbb{E}_g[h(Y^{(1)};\mu)^\gamma Y^{(1)}|X]$, respectively. The estimators of $m_{0,\mu}$ and $m_{1,\mu}$ are obtained by direct calculation or Monte Carlo approximation [34] based on the conditional density $q(y|T=1,X;\beta)$ of the outcome regression. Section 4.6 presents explicit forms of $m_{0,\mu}$ and $m_{1,\mu}$ when the conditional distribution is supposed to be Gaussian. The parameter β is usually estimated in an outlier-resistant manner, for example, by using Huber regression [35], MM estimator [84], density-power regression [7, 45], and γ-regression [25, 46]. Unlike the existing density power approaches, DP-DR does not multiply the whole estimating equation by the weight h^γ. Instead, we use h^γ as a multiplicative factor on the first term of (4.12), which is usual, but incorporate h^γ inside the conditional expectation on the second term of (4.12), which is unusual.

Theorem 4.3. *Suppose $h(y;\mu^{(1)})$ and $f_1(y)$ are both symmetric about $\mu^{(1)}$, and either the true PS or the true OR model is given. Then, if there is no contamination, the DP-DR estimating equation is unbiased.*

Now, we evaluate the bias of the DP-DR estimating equation under contamination.

Theorem 4.4. *Suppose that Y is contaminated as given by (4.2). Under the same assumptions as in Theorem 4.3, and if the true PS model is given, the expectation of the DP-DR estimating equation is expressed as*

$$\begin{aligned}&\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})-\frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta)-\mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\right]\\ =&-\int\varepsilon_1(x)\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1(\varepsilon_1).\end{aligned}\tag{4.13}$$

In particular, if the contamination ratio is independent of X, the right-hand side of (4.13) reduces to $\nu_1(\varepsilon_1)$.

If the true OR model is given, the expectation of the DP-DR estimating equation is expressed as

$$\begin{aligned}&\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha)}(Y-\mu^{(1)})-\frac{T-\pi(X;\alpha)}{\pi(X;\alpha)}\left\{m_{1,\mu^{(1)}}(X;\beta^*)-\mu^{(1)}m_{0,\mu^{(1)}}(X;\beta^*)\right\}\right]\\ =&\,\mathbb{E}_g\left[-\varepsilon_1(X)\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right]+\nu_1(\varepsilon_1P(T=1|\cdot)/\pi(\cdot;\alpha)).\end{aligned}\tag{4.14}$$

If the contamination ratio is independent of X, the right-hand side of (4.14) *becomes*

$$-\varepsilon_1 \mathbb{E}_g \left[\frac{P(T=1|X)}{\pi(X;\alpha)} \mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma (Y^{(1)}-\mu^{(1)})|X] \right] + \nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)). \tag{4.15}$$

Suppose that $\pi(\cdot;\alpha)$ is bounded away from 0 and 1; then, we find that $P(T=1|\cdot)/\pi(\cdot;\alpha)$ is bounded. Therefore, from Assumption 1, $\nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha))$ is negligible. As in the case of DP-IPW, the dominant term of the bias is independent of δ, indicating that the influence of outliers is reduced. Unfortunately, DP-DR is still biased if only the OR model is correct, even when the contamination ratio is constant.

Considering this result, we propose a variant of DP-DR called the εDP-DR estimator. This estimator is designed to cancel the dominant bias under homogeneous contamination. The εDP-DR estimator is a root of the following estimating equation:

$$\sum_{i=1}^{n} \left\{ \frac{T_i h(Y_i;\mu)^\gamma}{\pi(X_i;\hat{\alpha})}(Y_i-\mu) - \frac{T_i-\pi(X_i;\hat{\alpha})}{\pi(X_i;\hat{\alpha})}(1-\hat{\varepsilon}_1)\left\{ m_{1,\mu}(X_i;\hat{\beta}) - \mu m_{0,\mu}(X_i;\hat{\beta}) \right\} \right\} = 0, \tag{4.16}$$

where $\hat{\varepsilon}_1$ is a consistent estimator of the expected ratio of outliers: $\overline{\varepsilon}_1 = \int \varepsilon_1(x) g(x) dx$. The expected ratio of outliers can be estimated using the an outlier-resistant regression proposed in [45], for example. Under no contamination, the εDP-DR estimating equation is identical to the DP-DR estimating equation. The εDP-DR estimating equation is also biased under heterogeneous contamination; however, the bias takes a different form.

Corollary 4.1. *If the true PS model is given, the expectation of the εDP-DR estimating equation is equal to* (4.13)*. If the true OR model is given, the expectation of the εDP-DR estimating equation is expressed as*

$$\mathbb{E}_g \left[(\overline{\varepsilon}_1 - \varepsilon_1(X)) \frac{P(T=1|X)}{\pi(X;\alpha)} \mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma (Y^{(1)}-\mu^{(1)})|X] \right] + \nu_1(\varepsilon_1 P(T=1|\cdot)/\pi(\cdot;\alpha)). \tag{4.17}$$

The first term disappears if $\varepsilon_1(X)$ is constant.

Proof. Derivation is the same as that of Theorem 4.4. If $\varepsilon_1(X)$ is constant, the first term disappears because $\overline{\varepsilon}_1 = \varepsilon_1 \int g(x)dx = \varepsilon_1$. □

Similar to (4.15), the second term of (4.17) is approximately zero if we assume that $\pi(\cdot;\alpha)$ is bounded away from 0 and 1.

Remark One may believe that "$\varepsilon(X)$"DP-DR would work better than εDP-DR under X-dependent contamination. In fact, the bias (4.17) will disappear if we replace $\bar{\varepsilon}$ with $\varepsilon(X)$. However, it is necessary to model $\varepsilon(X)$ correctly for consistent estimation of "$\varepsilon(X)$"DP-DR. To the best of our knowledge, no easy method is available for this purpose.

4.3.4 Summary

We have proposed three types of outlier resistant semiparametric estimators: DP-IPW, DP-DR, and εDP-DR. Table 4.1 shows the bias of the estimating equations under the conditions discussed above. εDP-DR improves DP-DR in the OR-correct case under homogeneous contamination. However, we discuss DP-DR further below for two reasons: the contamination ratio is sometimes hard to estimate, and the simulation results presented in Section 4.7 indicate that DP-DR remains better than the existing methods even in the OR-correct case. Unfortunately, it is difficult to remove the influence of outliers under heterogeneous contamination. However, the bias of the estimating equations is hardly influenced by the absolute value of outliers. Furthermore, as discussed in Section 4.4, outliers have negligible influence if the contamination ratio is sufficiently small.

Contamination	model	DP-IPW	DP-DR	εDP-DR
No contam.	PS-correct	≈ 0	≈ 0	≈ 0
	OR-correct	-	≈ 0	≈ 0
homo.	PS-correct	≈ 0	≈ 0	≈ 0
	OR-correct	-	$\approx \varepsilon\mathbb{E}[\phi(X)]$	≈ 0
hetero.	PS-correct	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$
	OR-correct	-	$\approx \mathbb{E}[\varepsilon(X)\phi(X)]$	$\approx \mathbb{E}[(\bar{\varepsilon} - \varepsilon(X))\phi(X)]$

Table 4.1: Summary of bias of proposed estimating equations. The function $\phi(X)$ differs cell-by-cell. PS-correct means that the PS model is correctly specified and the OR model may not be; OR-correct means the opposite.

4.4 Influence Function-based Analysis

As discussed in the previous section, the three estimators are less suffered from outliers compared with ordinary estimators from the viewpoint of the unbiasedness of the estimating equation. In this section, we demonstrate that they are outlier-resistant from the viewpoint of IF.

Here, we briefly review the IF for the univariate M-estimator [35]. Further, we expand it to evaluate our estimators. Let G be the distribution of $Z \in \mathbb{R}$, and let

$T(G)$ be a functional of G, which is the parameter of interest. If of $T(G)$ is defined as

$$IF(z_0; G) := \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)G + \varepsilon\Delta_{z_0}) - T(G)}{\varepsilon} = \frac{\partial}{\partial \varepsilon}\{T((1-\varepsilon)G + \varepsilon\Delta_{z_0}) - T(G)\}\Big|_{\varepsilon=0}, \tag{4.18}$$

where Δ_{z_0} is a degenerate distribution at z_0. We also see that the latent bias $T((1-\varepsilon)G + \varepsilon\Delta_{z_0}) - T(G)$ can be approximated by $\varepsilon IF(z_0; G)$. Therefore, the behavior of the IF can approximately imply that of the latent bias. In a population, the M-estimator $T_M(G)$ satisfies $\int \psi(z, T_M(G))dG(z) = 0$. Then, the IF for $T_M(G)$ is obtained by differentiating $\int \psi(z, T_M((1-\varepsilon)G + \varepsilon\Delta_{z_0})d\{(1-\varepsilon)G + \varepsilon\Delta_{z_0}\}(z) = 0$ with respect to ε. This yields

$$IF(z_0; G) = -\mathbb{E}\left[\frac{\partial}{\partial \eta}\psi(Z, \eta)\Big|_{\eta=T_M(G)}\right]^{-1} \psi(z_0, T_M(G)). \tag{4.19}$$

The function ψ is said to have a redescending property if $\psi(z_0, T_M(G))$ approaches zero as the outlier $|z_0|$ increases. Therefore, when ψ has a redescending property and z_0 is an outlier, the latent bias is sufficiently small. This is favorable for outlier resistance.

Since ε_1 is dependent on X in our setting, we cannot apply the IF directly to our estimators. To overcome this issue, we consider the influence under fixed covariates $\{X_i\}_{i=1}^n$; this approach is similar to the fixed carrier model discussed in [30]. Consider the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{g}}\left[\psi(Y, T, X_i; \mu)\middle| X_i\right] = 0. \tag{4.20}$$

If the fixed sample $\{X_i\}_{i=1}^n$ consists of *i.i.d.* observations, then the left-hand side of (4.20) converges to $\mathbb{E}_{\tilde{g}}[\psi(Y, T, X; \mu)]$ as $n \to \infty$. Let $\tilde{\mu}_n^{(1)}$ denote a root of (4.20), and let $\tilde{\mu}^{(1)}$ be a root of $\mathbb{E}_{\tilde{g}}[\psi(Y, T, X; \mu)]$. Then, $\tilde{\mu}_n^{(1)}$ also converges to $\tilde{\mu}^{(1)}$. Therefore, $\tilde{\mu}_n^{(1)}$ shows roughly the same behavior as that of the target estimator $\tilde{\mu}^{(1)}$. The contaminated density $\tilde{g}$ is defined as (4.2), and $\delta_1(y|X_i)$ is assumed to be Dirac's delta at y_0. The IF of $T_n(\tilde{G})$ at X_i is obtained by differentiating (4.20) with respect to $\varepsilon_1(X_i)$ at $\varepsilon_1(X_i) = 0$.

Accordingly, the IF of the DP-IPW estimator at X_i is

$$\mathbb{E}_g\left[\frac{\partial \psi}{\partial \mu}\Big|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1} h(y_0; \mu_n^{(1)})^{\gamma}(y_0 - \mu_n^{(1)}). \tag{4.21}$$

If $\mu_n^{(1)}$ is close to $\mu^{(1)}$, then from Assumption 1', $h(y_0;\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)})$ tends to zero as $|y_0|\to\infty$. Thus, the DP-IPW estimator has a redescending property.

The IF of the DP-DR estimator is

$$-\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left\{\frac{P(T=1|X_i)}{\pi(X_i;\alpha)}h(y_0;\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)})\right.$$
$$\left.-\frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\{m_{1,\mu_n^{(1)}}(X_i;\beta)-m_{0,\mu_n^{(1)}}(X_i;\beta)\mu_n^{(1)}\}\right\}. \tag{4.22}$$

In the PS-correct case, the second term in the large curly brackets is equal to zero, and the IF tends to zero as $|y_0|\to\infty$. However, in the OR-correct case, the second term does not disappear. Considering the limit of $|y_0|\to\infty$, the IF converges to

$$-\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left\{-\frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\mathbb{E}[h(Y;\mu_n^{(1)})^\gamma(Y-\mu_n^{(1)})|X_i]\right\}. \tag{4.23}$$

Thus, the DP-DR estimator has a redescending property only in the PS-correct case. In the OR-correct case, the influence cannot be eliminated; however, the limit of the IF tends to a constant as $|y_0|$ tends to infinity, implying that the influence of the outlier is not serious.

The IF of the εDP-DR estimator is similar to that of the DP-DR estimator. Assume that $\overline{\varepsilon}_1=\frac{1}{n}\sum_{i=1}^n\varepsilon_1(X_i)$, then the IF is

$$-\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left\{\frac{P(T=1|X_i)}{\pi(X_i;\alpha)}h(y_0-\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)})\right.$$
$$\left.-\frac{n-1}{n}\frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\{m_{1,\mu_n^{(1)}}(X_i;\beta)-m_{0,\mu_n^{(1)}}(X_i;\beta)\mu_n^{(1)}\}\right\}; \tag{4.24}$$

this has a redescending property only in the PS-correct case. In the OR-correct case, the influence of outliers is not large, like in the case of the DP-DR estimator.

In conclusion, even when the contamination ratio depends on the confounder X, the proposed estimators are outlier-resistant when the contamination ratio is sufficiently small. The derivations of all IFs are presented in Section 4.9.

Under homogeneous contamination, the ordinary IF is applicable. As discussed above, we see that the proposed estimators have a redescending property under homogeneous contamination. Furthermore, εDP-DR has a redescending property even in the OR-correct case; this result is consistent with Corollary 4.1. The IF-based

analysis under homogeneous contamination is presented in Section 4.9.

4.5 Asymptotic Properties

We discuss the asymptotic properties of the εDP-DR estimators. For the other proposed estimators, we obtain similar results with small changes. The asymptotic properties can be obtained in a manner similar to that described in [34]. Assume that the PS and OR models are regular and are estimated consistently if the models are correctly specified. Furthermore, the contamination ratio ε_1 is known. Note that when the contamination ratio is consistently estimated simultaneously with the OR model by [45], we can replace β with $(\varepsilon_1, \beta^T)^T$ in the following discussion.

Denote (4.16) by $\frac{1}{n}\sum_{i=1}^n \psi_i(\mu; \hat{\alpha}, \hat{\beta})$, and let $\frac{1}{n}\sum_{i=1}^n s_i^{PS}(\alpha) = 0$ and $\frac{1}{n}\sum_{i=1}^n s_i^{OR}(\beta) = 0$ be the estimating equations for the PS and OR models, respectively. Let $\lambda = (\mu, \alpha^T, \beta^T)^T$ be the parameter vector, and define the full estimating equation as

$$\sum_{i=1}^n S_i(\lambda) = \sum_{i=1}^n \begin{pmatrix} \psi_i(\mu; \alpha, \beta) \\ s_i^{PS}(\alpha) \\ s_i^{OR}(\beta) \end{pmatrix} = \mathbf{0}. \tag{4.25}$$

Let $\lambda^* = (\mu^*, \alpha^{*T}, \beta^{*T})^T$ be a root of (4.25) in population. Note that, in this section, $*$ does not mean that the model is correctly specified. With the results presented in [79], the following theorem holds under some regularity conditions.

Theorem 4.5. *Under the regularity conditions presented in Appendix 4.9.7, the following asymptotic properties hold:*

$$\hat{\lambda} \xrightarrow{p} \lambda^*, \tag{4.26}$$

$$\sqrt{n}(\hat{\lambda} - \lambda^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}^{\tilde{g}}(\lambda^*)\right), \tag{4.27}$$

$$\mathbf{V}^{\tilde{g}}(\lambda^*) = \mathbf{J}^{\tilde{g}}(\lambda^*)^{-1}\mathbf{K}^{\tilde{g}}(\lambda^*)\{\mathbf{J}^{\tilde{g}}(\lambda^*)^T\}^{-1}, \tag{4.28}$$

$$\mathbf{J}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[\partial S_i(\lambda^*)/\partial \lambda^T\right], \tag{4.29}$$

$$\mathbf{K}^{\tilde{g}}(\lambda^*) = \mathbb{E}_{\tilde{g}}\left[S_i(\lambda^*)S_i(\lambda^*)^T\right]. \tag{4.30}$$

Under homogeneous contamination, by applying the results presented in Section 4.3.3, we find that the limit μ^* is in the neighborhood of $\mu^{(1)}$.

Theorem 4.6. *Let $\lambda^{**} = (\mu^{(1)}, \alpha^{*T}, \beta^{*T})^T$ and assume that $\mathbf{J}_{11}^{\tilde{g}}(\lambda)$ is non-zero within the interval $[\lambda^*, \lambda^{**}]$. Under Assumption 1 and homogeneous contamination, if either the PS or the OR model is correct, it then holds that*

$$\mu^* = \mu^{(1)} + \mathcal{O}(\nu_1(\phi)), \tag{4.31}$$

where $\phi(\cdot) = \varepsilon_1$ (constant) in the PS-correct case and $\phi(\cdot) = \varepsilon_1 P(T = 1|\cdot)/\pi(\cdot;\alpha)$ in the OR-correct case.

The proof of Theorem 4.6 and further discussions on the asymptotic variance are available in Section 4.9.

4.6 Algorithm

4.6.1 General Form

Because the proposed estimating equations cannot be solved explicitly, we use an iterative algorithm. Various algorithms are available; however, we propose a standard algorithm for M-estimators [35, 30]. The algorithm for the DP-IPW estimator is given by the following updates:

$$\hat{\mu}^{[a+1]} = \left\{ \sum_{i=1}^{n} w_i^{[a]} Y_i \right\} \left\{ \sum_{i=1}^{n} w_i^{[a]} \right\}^{-1}, \tag{4.32}$$

$$w_i^{[a+1]} = \frac{T_i h(Y_i; \hat{\mu}^{[a+1]})^{\gamma}}{\pi(X_i; \hat{\alpha})} \quad \text{for all} \;\; i. \tag{4.33}$$

We recommend to obtain the initial values $(\mu^{[0]}, w_i^{[0]})$ in an outlier-resistant manner. For example, $\mu^{[0]}$ can be obtained by the IPW median [21, 90], and $w_i^{[0]}$ is obtained using (4.33). If the weighting density is indexed by other parameters, it must be estimated in advance or be updated simultaneously to μ and w. In the next section, we present an algorithm in which we assume that h is Gaussian.

The (ε)DP-DR estimator is obtained in a similar manner. Let $h(\cdot; \mu^{(1)})$ be fixed and solve (4.16) with respect to μ:

$$\begin{aligned} \mu = \frac{1}{n} \sum_{i=1}^{n} & \left\{ \frac{T_i h(Y_i; \mu^{(1)})^{\gamma} Y_i}{\pi(X_i; \hat{\alpha})} - \frac{T_i - \pi(X_i; \hat{\alpha})}{\pi(X_i; \hat{\alpha})} (1 - \hat{\varepsilon}) m_{1,\mu^{(1)}}(X_i; \hat{\beta}) \right\} \\ & \times \left\{ \frac{T_i h(Y_i; \mu^{(1)})^{\gamma}}{\pi(X_i; \hat{\alpha})} - \frac{T_i - \pi(X_i; \hat{\alpha})}{\pi(X_i; \hat{\alpha})} (1 - \hat{\varepsilon}) m_{0,\mu^{(1)}}(X_i; \hat{\beta}) \right\}^{-1}. \end{aligned} \tag{4.34}$$

Then, the following algorithm is obtained:

$$\hat{\mu}^{[a+1]} = \left\{ \sum_{i=1}^{n} w_{1,i}^{[a]} Y_i - w_{2,i} \hat{m}_{1,\mu^{[a]}}(X_i; \hat{\beta}) \right\} \left\{ \sum_{i=1}^{n} w_{1,i}^{[a]} - w_{2,i} \hat{m}_{0,\mu^{[a]}}(X_i; \hat{\beta}) \right\}^{-1}, \tag{4.35}$$

$$w_{1,i}^{[a+1]} = \frac{T_i h(Y_i; \hat{\mu}^{[a+1]})^{\gamma}}{\pi(X_i; \hat{\alpha})} \quad \text{for all} \;\; i, \tag{4.36}$$

$$w_{2,i} = \frac{T_i - \pi(X_i; \hat{\alpha})}{\pi(X_i; \hat{\alpha})} (1 - \hat{\varepsilon}_1) \quad \text{for all} \;\; i. \tag{4.37}$$

Note that it is not necessary to update $w_{2,i}$ once it is computed. The initial values should be obtained in an outlier-resistant manner, as in DP-IPW. Recall that $\hat{m}_{1,\mu}$ and $\hat{m}_{0,\mu}$ are the estimates for the conditional expectation $\mathbb{E}_g[h(Y^{(1)}; \mu)^{\gamma} Y^{(1)} | X]$ and $\mathbb{E}_g[h(Y^{(1)}; \mu)^{\gamma} | X]$ given μ. These updates can be obtained from the estimated conditional density $q(y|X; \hat{\beta})$ through Monte-Carlo approximation [34] or direct calculations.

4.6.2 Gaussian Weight

When the weighting density is assumed to be Gaussian, some value must be assigned to the standard deviation σ. Under contamination, we suggest that σ is estimated in an outlier-resistant manner, such as by using the normalized median absolute deviation (MADN) [30]. MADN is an unbiased estimator for the standard deviation of a Gaussian random variable. When the weighting density is Gaussian, the parameters are updated as follows:

$$\hat{\mu}^{[a+1]} = \left\{ \sum_{i=1}^{n} w_i^{[a]} Y_i \right\} \left\{ \sum_{i=1}^{n} w_i^{[a]} \right\}^{-1}. \tag{4.38}$$

$$\hat{\sigma}^{[a+1]} = \text{IPW-MADN}(\{Y_i\}_{i=1}^{n}, \hat{\mu}^{[a+1]}). \tag{4.39}$$

$$w_i^{[a+1]} = \frac{T_i h(Y_i; \theta(\hat{\mu}^{[a+1]}, \hat{\sigma}^{[a+1]}))^{\gamma}}{\pi(X_i; \hat{\alpha})} \quad \text{for all} \;\; i. \tag{4.40}$$

The IPW-MADN is defined as

$$\text{IPW-MADN}(\{Y_i\}_{i=1}^{n}, \mu) = 1.483 \cdot \text{IPW-median}\left(\{|Y_i - \mu|\}_{i=1}^{n}\right), \tag{4.41}$$

where 1.483 is a normalization constant.

Similarly, the (ε)DP-DR estimator is updated as

$$\hat{\mu}^{[a+1]} = \left\{ \sum_{i=1}^{n} w_{1,i}^{[a]} Y_i - w_{2,i} \hat{m}_{1,\mu^{[a]}}(X_i; \hat{\beta}) \right\} \left\{ \sum_{i=1}^{n} w_{1,i}^{[a]} - w_{2,i} \hat{m}_{0,\mu^{[a]}}(X_i; \hat{\beta}) \right\}^{-1}, \tag{4.42}$$

$$\hat{\sigma}^{[a+1]} = \text{DR-MADN}(\{Y_i\}_{i=1}^n, \hat{\mu}^{[a+1]}), \tag{4.43}$$

$$w_{1,i}^{[a+1]} = \frac{T_i h(Y_i; \hat{\mu}^{[a+1]}, \hat{\sigma}^{[a+1]})^{\gamma}}{\pi(X_i; \hat{\alpha})} \quad \text{for all} \quad i, \tag{4.44}$$

$$w_{2,i} = \frac{T_i - \pi(X_i; \hat{\alpha})}{\pi(X_i; \hat{\alpha})} (1 - \hat{\varepsilon}_1) \quad \text{for all} \quad i. \tag{4.45}$$

The DR-MADN is obtained by using the DR-median [90, 18, 72]

$$\text{DR-MADN}(\{Y_i\}_{i=1}^n, \mu) = 1.483 \cdot \text{DR-median}\left(\{|Y_i - \mu|\}_{i=1}^n\right). \tag{4.46}$$

Further, the updates of $\hat{m}_{1,\mu}$ and $\hat{m}_{0,\mu}$ are expressed explicitly when $q(y|X; \hat{\beta})$ is assumed to be the conditional Gaussian distribution given X. Let $u(X) = \mathbb{E}_q[Y|X]$ and $v^2(X) = \text{Var}_q[Y|X]$. Then, we obtain

$$m_{0,\mu^{[a]}}(X) = (2\pi)^{-\frac{\gamma}{2}} \frac{(\sigma^{[a]^2})^{\frac{1-\gamma}{2}}}{\sqrt{\sigma^{[a]^2} + \gamma v^2(X)}} \cdot \exp\left\{ -\frac{\gamma(\mu^{[a]} - u(X))}{2(\sigma^{[a]^2} + \gamma v^2(X))} \right\}, \tag{4.47}$$

$$m_{1,\mu^{[a]}}(X) = (2\pi)^{-\frac{\gamma}{2}} \frac{(\sigma^{[a]^2})^{\frac{1-\gamma}{2}}}{\sqrt{\sigma^{[a]^2} + \gamma v^2(X)}} \cdot \frac{u(X)\sigma^{[a]^2} + \gamma \mu^{[a]} v^2(X)}{\sigma^{[a]^2} + \gamma v^2(X)} \cdot \exp\left\{ -\frac{\gamma(\mu^{[a]} - u(X))}{2(\sigma^{[a]^2} + \gamma v^2(X))} \right\}. \tag{4.48}$$

Notably, the conditional variance can be easily estimated because many general outlier-resistant methods can be applied for this purpose.

4.7 Monte-Carlo Simulations

4.7.1 Comparative Methods

We conduct Monte-Carlo simulations to evaluate the performance of the proposed estimators. We compare our methods with naive IPW and DR estimators and some existing outlier-resistant methods [21, 90, 18, 72]. These methods focus on the median of the potential outcome; therefore, they show limited outlier resistance. To the best of our knowledge, no method other than the proposed method has more

outlier resistance than the median. Firpo's IPW estimator [21] is defined as

$$\hat{\mu}_{\text{Firpo}} = \arg\min_{\mu} \sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}(Y_i - \mu)(0.5 - \mathbb{I}(Y_i \leq \mu)), \tag{4.49}$$

where the function $\mathbb{I}$ is an indicator. Zhang's IPW median [90] is based on the IPW-empirical distribution

$$\hat{F}_{\text{IPW}}(y) = \left(\sum_{i=1}^{n} \frac{T_i \mathbb{I}(Y_i \leq y)}{\pi(X_i;\hat{\alpha})}\right) \Bigg/ \left(\sum_{i=1}^{n} \frac{T_i}{\pi(X_i;\hat{\alpha})}\right), \tag{4.50}$$

and the median is estimated as y_0 such that $\hat{F}_{IPW}(y_0) = 0.5$. Firpo's IPW and Zhang's IPW are almost equivalent except for a slight difference in their computation. Furthermore, some methods have been proposed for the DR-median. Zhang's and Sued's DR methods [90, 72] estimate the empirical distribution in a doubly robust way. They incorporate an IPW-type estimator into the first term. The remaining term of Zhang's DR is based on the Gaussian cumulative distribution function of Y given X. By contrast, Sued's DR constructs the remaining term in a nonparametric manner. Diaz's DR median [18] is a largely different approach; it employs the targeted maximum likelihood estimator (TMLE) [78]. Nonetheless, all of these comparative methods focus on the median of the potential outcome; therefore, our method is the first one whose outlier resistance is more than the sample median. We implement our methods, Zhang's IPW/DR, and Sued's DR in R. For Firpo's IPW and TMLE, we use the *causalquantile* package [1].

4.7.2 Simulation Model

We generate random observations based on a simple causal setting. The confounders (X_1, X_2) are independently generated from a Gaussian or uniform distribution with mean zero and unit variance. The treatment T is assigned along with the conditional probability $P(T = 1|X_1, X_2)$ that is defined as a sigmoid function of $0.8X_1 + 0.2X_2$. The potential outcomes $(Y^{(1)}, Y^{(0)})$ are generated according to a linear function of (X_1, X_2) with Gaussian error: $Y^{(1)} = \mu^{(1)} + 1.2X_1 + 0.3X_2 + e$ and $Y^{(0)} = \mu^{(0)} + 1.2X_1 + 0.3X_2 + e$. The standard deviation (SD) of e is set to 0.72; therefore, $\text{SD}[Y^{(1)}] = \text{SD}[Y^{(0)}] = 1.5$. The potential means $\mu^{(1)}$ and $\mu^{(0)}$ are set to 3 and 0, respectively. When the confounders are not Gaussian, the target variable $Y^{(t)}$ is not Gaussian. The observed outcome Y is defined as $Y = TY^{(1)} + (1 - T)Y^{(1)}$ under no contamination. Outliers are generated from $\mathcal{N}(\mu^{(t)} + 10\sigma^{(t)}, 1)$, with $\sigma^{(t)} = \text{SD}[Y^{(t)}] =$

[1] https://github.com/idiazst/causalquantile (Updated on 31 Aug 2017)

1.5. For the homogeneous contamination settings, the contamination ratio is set to be a constant $\varepsilon_t \in \{0, 0.05, 0.1, 0.2\}$. For the heterogeneous contamination settings, the contamination ratio is set to be $1.5\varepsilon_t$ if $X_1 + X_2 \leq 0$ and $0.5\varepsilon_t$ if $X_1 + X_2 > 0$. The average contamination ratio is set at $\varepsilon_t \in \{0, 0.05, 0.1, 0.2\}$. The observations of Y are randomly replaced with outliers according to the contamination ratio. The sample size is fixed to $n = 100$ throughout the Monte Carlo simulations. Further, we generate several datasets in which outcome followed a heavy-tailed distribution. We draw the error term of $Y^{(t)}$ from the standard Cauchy distribution instead of inserting outliers.

4.7.3 Results

First, we perform a comparative study. The potential mean $\mu^{(1)}$ is estimated using the proposed methods and the comparative methods. In this experiment, we use all settings illustrated in the previous section. The propensity score is estimated by logistic regression. The outcome regression is conducted in two ways: Gaussian MLE with non-outliers or unnormalized Gaussian modeling (the tuning parameter was set to 0.5) [45]. For the DR estimators, we investigate three patterns of model misspecification: PS-correct/OR-correct (T/T), PS-correct/OR-incorrect (T/F), and PS-incorrect/OR-correct (F/T). In the model-correct case, we include an intercept and (X_1, X_2) as covariates. In the model-incorrect case, we include only an intercept and X_2. We perform 10,000 simulations for each setting and method. Tables 4.2 and 4.3 show the results of the comparative study when the covariates are Gaussian and the OR for the DR-type estimators was the Gaussian MLE with non-outliers. The estimation error is measured by the root mean square error (RMSE). The mean and SD of all estimates and the mean computation time, and the results for the other settings are provided in Section 4.9. In Table 4.2, the naive IPW estimator had a significantly larger RMSE under contamination. Both the median-based methods and DP-IPW dramatically reduced the RMSE. As the ratio of outliers increased, the RMSE increased. The RMSE tended to be larger for heterogeneous contamination than for homogeneous contamination. When the optimal γ was properly chosen, the proposed method outperformed the comparative methods and had the smallest RMSE for all settings. The results of the DR-type estimators were similar to those of the IPW estimators, as shown in Table 4.3. The proposed method with a proper γ outperformed the comparative methods and had the smallest RMSE in all settings. In particular, when $\gamma = 0.5$, the proposed method was always superior to the comparative methods. DP-DR and εDP-DR performed similarly, although εDP-DR was slightly superior in many settings. Among the median-based methods, TMLE

performed relatively well; however, it took much more time than the other methods, including the proposed methods, and occasionally (<1%) failed to converge. Table 4.4 shows the mean and SD of 10,000 simulated estimates of naive DR, DP-DR, and εDP-DR under homogeneous contamination. In that case, the average of the εDP-DR estimates was closer to 3 than that of DP-DR in F/T cases, which corresponds to Corollary 4.1.

Table 4.5 shows the RMSE of each method on the heavy-tailed data. As well as the above experiments, the proposed method performed better than the comparative methods. In this setting, we only used the unnormalized Gaussian modeling for OR for the DR-type estimators. Only in the PS-correct/OR-incorrect case with Gaussian X, the median (TMLE) performed slightly better than the proposed method.

Next, we show the result of a γ-sensitivity study. We estimate $\mu^{(1)}$ by the proposed method with different γs. X had a Gaussian distribution, and the contamination ratio varied in $\{0, 0.05, 0.1, 0.2\}$ under homogeneous contamination. For the DR estimators, the outcome regression was obtained by the Gaussian MLE with non-outliers. We perform 10,000 simulations for each setting and method. Table 4.6 shows the results of the γ-sensitivity study. As in the comparative study, when the ratio of outliers increased, the bias increased. Furthermore, a larger γ resulted in increased variance. When the contamination ratio was small ($\varepsilon = 0.05$), it was sufficient to use a small γ such as $\gamma = 0.1$ or 0.2 to remove the adverse effect of outliers. Even in highly contaminated cases, it seems unnecessary to use γ larger than 1.0. As in many other outlier-resistant statistical methods, parameter tuning is a challenging issue. Based on Figure 4.1, we suggest a possible policy on this issue. Figure 4.1 shows the solution paths of the first 100 simulations. The adverse effect of outliers decreased as γ increased, and each path became stable around the true value after reaching a specific γ value. Thus, we suggest using the smallest γ value among the γ values with stable estimates to avoid increasing the variance.

		Homogeneous			**Heterogeneous**		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
Naive	0.222	0.957	1.683	3.153	0.993	1.752	3.253
median (Firpo)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
median (Zhang-IPW)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
DP-IPW ($\gamma = 0.1$)	0.218	0.276	0.531	2.263	0.293	0.609	2.377
DP-IPW ($\gamma = 0.5$)	0.227	0.249	0.272	0.639	0.245	0.287	0.726
DP-IPW ($\gamma = 1.0$)	0.261	0.271	0.275	0.413	0.262	0.281	0.498

Table 4.2: RMSE of the IPW-type estimators. X was drawn from Gaussian distributions.

		Homogeneous			Heterogeneous		
ε	0.00	0.05	0.10	0.20	0.05	0.10	0.20
(PS-correct/OR-correct)							
Naive	0.184	0.957	1.684	3.154	0.997	1.758	3.265
median (Zhang-DR)	0.239	0.317	0.391	0.733	0.330	0.452	0.905
median (Sued)	0.238	0.316	0.388	0.693	0.329	0.450	0.869
median (TMLE)	0.237	0.280	0.359	0.603	0.295	0.402	0.701
DP-DR ($\gamma = 0.1$)	0.183	0.302	0.564	2.262	0.318	0.649	2.394
DP-DR ($\gamma = 0.5$)	0.202	0.285	0.326	0.697	0.274	0.349	0.834
DP-DR ($\gamma = 1.0$)	0.240	0.288	0.307	0.524	0.287	0.336	0.669
εDP-DR ($\gamma = 0.1$)	0.183	0.296	0.554	2.255	0.314	0.636	2.385
εDP-DR ($\gamma = 0.5$)	0.202	0.264	0.302	0.669	0.271	0.323	0.793
εDP-DR ($\gamma = 1.0$)	0.240	0.287	0.299	0.513	0.286	0.335	0.648
(correct/incorrect)							
Naive	0.237	0.963	1.686	3.156	1.001	1.758	3.262
median (Zhang-DR)	0.275	0.342	0.408	0.741	0.350	0.465	0.912
median (Sued)	0.275	0.342	0.407	0.699	0.350	0.464	0.872
median (TMLE)	0.242	0.284	0.363	0.622	0.297	0.404	0.719
DP-DR ($\gamma = 0.1$)	0.237	0.314	0.561	2.267	0.330	0.644	2.393
DP-DR ($\gamma = 0.5$)	0.247	0.319	0.349	0.714	0.319	0.361	0.839
DP-DR ($\gamma = 1.0$)	0.280	0.334	0.347	0.581	0.329	0.372	0.709
εDP-DR ($\gamma = 0.1$)	0.237	0.311	0.557	2.264	0.328	0.640	2.388
εDP-DR ($\gamma = 0.5$)	0.247	0.317	0.344	0.694	0.313	0.356	0.817
εDP-DR ($\gamma = 1.0$)	0.280	0.333	0.338	0.551	0.327	0.369	0.708
(incorrect/correct)							
Naive	0.181	0.879	1.591	3.026	0.826	1.490	2.813
median (Zhang-DR)	0.237	0.263	0.316	0.503	0.269	0.337	0.548
median (Sued)	0.236	0.272	0.346	0.599	0.277	0.364	0.627
median (TMLE)	0.234	0.260	0.309	0.478	0.265	0.328	0.522
DP-DR ($\gamma = 0.1$)	0.182	0.192	0.345	2.057	0.191	0.299	1.681
DP-DR ($\gamma = 0.5$)	0.199	0.206	0.218	0.366	0.203	0.209	0.283
DP-DR ($\gamma = 1.0$)	0.230	0.232	0.239	0.273	0.230	0.233	0.242
εDP-DR ($\gamma = 0.1$)	0.182	0.193	0.381	2.207	0.194	0.335	1.839
εDP-DR ($\gamma = 0.5$)	0.199	0.203	0.208	0.376	0.203	0.212	0.318
εDP-DR ($\gamma = 1.0$)	0.230	0.230	0.231	0.243	0.231	0.237	0.260

Table 4.3: RMSE of the DR-type estimators. X was drawn from Gaussian distributions.

	No contam.	**Homogeneous**		
ε	0.00	0.05	0.10	0.20
(PS-correct/OR-correct)				
Naive	2.999 (0.18)	3.745 (0.60)	4.489 (0.79)	5.979 (1.04)
DP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.029 (0.30)	3.140 (0.55)	4.465 (1.72)
DP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.997 (0.29)	3.000 (0.33)	3.060 (0.69)
DP-DR ($\gamma = 1.0$)	2.992 (0.24)	2.991 (0.29)	2.992 (0.31)	3.009 (0.52)
εDP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.028 (0.29)	3.138 (0.54)	4.464 (1.72)
εDP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.997 (0.26)	2.999 (0.30)	3.058 (0.67)
εDP-DR ($\gamma = 1.0$)	2.992 (0.24)	2.991 (0.29)	2.991 (0.30)	3.007 (0.51)
(correct/incorrect)				
Naive	3.004 (0.24)	3.750 (0.60)	4.494 (0.78)	5.984 (1.03)
DP-DR ($\gamma = 0.1$)	2.998 (0.24)	3.033 (0.31)	3.150 (0.54)	4.490 (1.71)
DP-DR ($\gamma = 0.5$)	2.986 (0.25)	2.989 (0.32)	2.992 (0.35)	3.059 (0.71)
DP-DR ($\gamma = 1.0$)	2.979 (0.28)	2.978 (0.33)	2.979 (0.35)	3.001 (0.58)
εDP-DR ($\gamma = 0.1$)	2.998 (0.24)	3.033 (0.31)	3.149 (0.54)	4.489 (1.71)
εDP-DR ($\gamma = 0.5$)	2.986 (0.25)	2.989 (0.32)	2.992 (0.34)	3.057 (0.69)
εDP-DR ($\gamma = 1.0$)	2.979 (0.28)	2.978 (0.33)	2.978 (0.34)	2.998 (0.55)
(correct/incorrect)				
Naive	2.999 (0.18)	3.725 (0.50)	4.451 (0.65)	5.902 (0.86)
DP-DR ($\gamma = 0.1$)	2.999 (0.18)	2.997 (0.19)	3.051 (0.34)	4.326 (1.57)
DP-DR ($\gamma = 0.5$)	3.001 (0.20)	2.975 (0.20)	2.950 (0.21)	2.907 (0.35)
DP-DR ($\gamma = 1.0$)	3.005 (0.23)	2.978 (0.23)	2.953 (0.23)	2.895 (0.25)
εDP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.020 (0.19)	3.108 (0.37)	4.541 (1.58)
εDP-DR ($\gamma = 0.5$)	3.001 (0.20)	2.998 (0.20)	2.998 (0.21)	3.020 (0.38)
εDP-DR ($\gamma = 1.0$)	3.005 (0.23)	3.001 (0.23)	3.001 (0.23)	3.003 (0.24)

Table 4.4: Mean and SD of 10,000 simulated estimates of $\mu^{(1)}$. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the Gaussian MLE using non-outliers. The complete table including the IPW-type estimators and the comparative methods is in Section 4.9

		Distribution of X	
		Gaussian	Uniform
IPW(T/-)	Naive	274.024	246.118
	median (Firpo)	0.414	0.438
	median (Zhang-IPW)	0.414	0.438
	DP-IPW ($\gamma = 0.1$)	0.443	0.425
	DP-IPW ($\gamma = 0.5$)	0.367	0.363
	DP-IPW ($\gamma = 1.0$)	0.380	0.383
DR(T/T)	Naive	275.447	247.011
	median (Zhang-DR)	0.415	0.431
	median (Sued)	0.408	0.430
	median (TMLE)	0.392	0.425
	DP-DR ($\gamma = 0.1$)	0.501	0.420
	DP-DR ($\gamma = 0.5$)	0.363	0.356
	DP-DR ($\gamma = 1.0$)	0.372	0.374
	εDP-DR ($\gamma = 0.1$)	0.487	0.420
	εDP-DR ($\gamma = 0.5$)	0.361	0.355
	εDP-DR ($\gamma = 1.0$)	0.370	0.374
DR(T/F)	Naive	275.446	247.011
	median (Zhang-DR)	0.456	0.443
	median (Sued)	0.436	0.441
	median (TMLE)	0.394	0.427
	DP-DR ($\gamma = 0.1$)	0.514	0.431
	DP-DR ($\gamma = 0.5$)	0.404	0.369
	DP-DR ($\gamma = 1.0$)	0.418	0.389
	εDP-DR ($\gamma = 0.1$)	0.503	0.430
	εDP-DR ($\gamma = 0.5$)	0.399	0.368
	εDP-DR ($\gamma = 1.0$)	0.412	0.388
DR(F/T)	Naive	263.629	177.037
	median (Zhang-DR)	0.390	0.429
	median (Sued)	0.373	0.400
	median (TMLE)	0.389	0.429
	DP-DR ($\gamma = 0.1$)	0.390	0.401
	DP-DR ($\gamma = 0.5$)	0.358	0.376
	DP-DR ($\gamma = 1.0$)	0.364	0.393
	εDP-DR ($\gamma = 0.1$)	0.377	0.385
	εDP-DR ($\gamma = 0.5$)	0.328	0.338
	εDP-DR ($\gamma = 1.0$)	0.334	0.351

Table 4.5: RMSE values of the comparative study using the heavy-tailed data. The covariates X were generated from Gaussian or uniform distributions. The OR model for the DR-type estimators were obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling, respectively.

	PS/OR	ε	$\gamma = 0.0$	0.1	0.2	0.5	1.0	1.5	2.0
DP-IPW	T/-	0.00	3.004 (0.22)	2.998 (0.22)	2.994 (0.22)	2.986 (0.23)	2.980 (0.26)	2.974 (0.30)	2.970 (0.34)
		0.05	3.749 (0.59)	3.030 (0.27)	2.999 (0.26)	2.987 (0.25)	2.978 (0.27)	2.970 (0.30)	2.963 (0.33)
		0.10	4.493 (0.78)	3.142 (0.51)	3.015 (0.32)	2.989 (0.27)	2.977 (0.27)	2.969 (0.30)	2.963 (0.33)
		0.20	5.983 (1.02)	4.492 (1.70)	3.536 (1.39)	3.052 (0.64)	2.990 (0.41)	2.978 (0.39)	2.971 (0.40)
DP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.029 (0.30)	3.002 (0.27)	2.997 (0.29)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.79)	3.140 (0.55)	3.017 (0.36)	3.000 (0.33)	2.992 (0.31)	2.986 (0.32)	2.981 (0.33)
		0.20	5.979 (1.04)	4.465 (1.72)	3.532 (1.41)	3.060 (0.69)	3.009 (0.52)	2.999 (0.51)	2.994 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.494 (0.78)	3.150 (0.54)	3.020 (0.37)	2.992 (0.35)	2.979 (0.35)	2.970 (0.37)	2.963 (0.39)
		0.20	5.984 (1.03)	4.490 (1.71)	3.546 (1.41)	3.059 (0.71)	3.001 (0.58)	2.985 (0.55)	2.975 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.725 (0.50)	2.997 (0.19)	2.976 (0.19)	2.975 (0.20)	2.978 (0.23)	2.982 (0.26)	2.986 (0.29)
		0.10	4.451 (0.65)	3.051 (0.34)	2.956 (0.21)	2.950 (0.21)	2.953 (0.23)	2.956 (0.26)	2.960 (0.28)
		0.20	5.902 (0.86)	4.326 (1.57)	3.301 (1.15)	2.907 (0.35)	2.895 (0.25)	2.897 (0.26)	2.900 (0.28)
εDP-DR	T/T	0.00	2.999 (0.18)	2.998 (0.18)	2.997 (0.19)	2.996 (0.20)	2.992 (0.24)	2.989 (0.28)	2.985 (0.31)
		0.05	3.745 (0.60)	3.028 (0.29)	3.002 (0.27)	2.997 (0.26)	2.991 (0.29)	2.985 (0.31)	2.980 (0.34)
		0.10	4.489 (0.78)	3.138 (0.54)	3.017 (0.35)	2.999 (0.30)	2.991 (0.30)	2.985 (0.32)	2.980 (0.33)
		0.20	5.978 (1.03)	4.464 (1.72)	3.531 (1.40)	3.058 (0.67)	3.007 (0.51)	2.998 (0.50)	2.993 (0.51)
	T/F	0.00	3.004 (0.24)	2.998 (0.24)	2.994 (0.24)	2.986 (0.25)	2.979 (0.28)	2.974 (0.32)	2.968 (0.36)
		0.05	3.750 (0.60)	3.033 (0.31)	3.001 (0.29)	2.989 (0.32)	2.978 (0.33)	2.970 (0.36)	2.963 (0.39)
		0.10	4.493 (0.78)	3.149 (0.54)	3.020 (0.36)	2.992 (0.34)	2.978 (0.34)	2.970 (0.37)	2.963 (0.39)
		0.20	5.983 (1.02)	4.489 (1.71)	3.543 (1.40)	3.057 (0.69)	2.998 (0.55)	2.984 (0.54)	2.976 (0.54)
	F/T	0.00	2.999 (0.18)	2.999 (0.18)	2.999 (0.18)	3.001 (0.20)	3.005 (0.23)	3.010 (0.26)	3.014 (0.29)
		0.05	3.746 (0.50)	3.020 (0.19)	2.998 (0.19)	2.998 (0.20)	3.001 (0.23)	3.005 (0.26)	3.009 (0.29)
		0.10	4.493 (0.66)	3.108 (0.37)	3.004 (0.20)	2.998 (0.21)	3.001 (0.23)	3.004 (0.26)	3.007 (0.28)
		0.20	5.986 (0.87)	4.541 (1.58)	3.486 (1.24)	3.020 (0.38)	3.003 (0.24)	3.005 (0.25)	3.008 (0.27)

Table 4.6: Results of γ-sensitivity study. Each figure displays the mean (SD) of 10,000 simulations for each setting. In the second column, "T" and "F" denote the correct and the incorrect modeling, respectively.

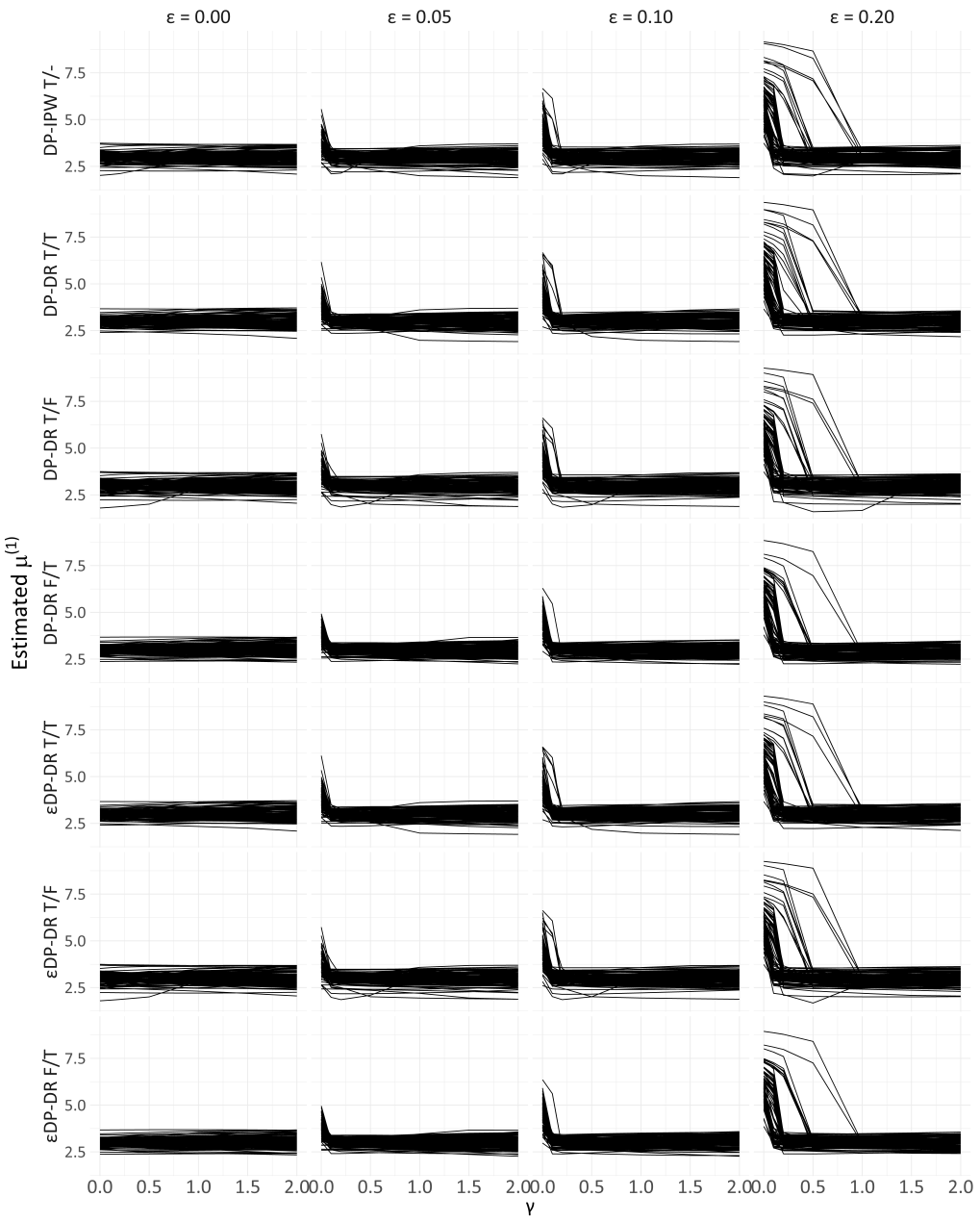

Figure 4.1: Solution paths of the first 100 simulations. The x-axis represents the tuning parameter γ and the y-axis, the estimates of $\mu^{(1)}$.

4.8 Real Data Analysis

In this section, we demonstrate an estimation of the ATE on a real dataset. We used the data of the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The NHEFS is a national longitudinal study that was performed by U.S. public agencies. We used the processed dataset available online [2] [32]. The NHEFS dataset contains 1,566 observations of smokers when the study started in 1971–75. By the follow-up visit in 1982, 403 (25.7%) participants had quit smoking. The goal was to evaluate the treatment effect of smoking cessation ($T = 1$) on weight gain (Y). Other than the treatment and the outcome, several baseline variables were collected, including sex, age, race, education level, intensity and duration of smoking, physical activity in daily life, recreational exercise, and baseline weight. We used all of them to control for confounding in a manner similar to [32]. We included linear and quadratic terms for all the continuous covariates (age, intensity and duration of smoking, and baseline weight) and dummy terms for the discrete covariates. The propensity score was estimated by logistic regression, and outcome regression was conducted by unnormalized Gaussian modeling [45] (tuning parameter was set to 0.2). The original dataset does not contain obvious outliers; therefore, we randomly replaced 10% observations with outliers extracted from $\mathcal{N}(100, 5^2)$. Then, we estimated $\mu^{(1)}$, $\mu^{(0)}$ and the ATE by the same methods in the Monte Carlo simulations. We repeated this process 10,000 times and summarized the results in Table 4.7. For reference, we estimated each quantity using the naive IPW/DR for the original data.

For the IPW-type estimators, the median-based methods tended to give larger estimates of $\mu^{(1)}$ and $\mu^{(0)}$ than those in the case of IPW (no outliers). In particular, $\mu^{(0)}$ was estimated to be much larger. As a result, when using the median-based methods, the ATE was estimated to be smaller than that in the case of IPW (no outliers). By contrast, DP-IPW tended to overestimate $\mu^{(1)}$ with $\gamma = 0.05$ and to underestimate $\mu^{(1)}$ with $\gamma \geq 0.10$; further, it overestimated $\mu^{(0)}$ compared to the case of IPW (no outliers). This tendency was strengthened by increasing γ. However, because the extent of overestimation of $\mu^{(0)}$ was smaller than that in the case of median-based methods, the estimate of the ATE by DP-IPW was closer to that obtained using IPW (no outliers) than to that obtained using median-based methods, even for $\gamma = 0.5$. The DR-type estimators showed similar results. The median-based methods overestimated $\mu^{(1)}$ and $\mu^{(0)}$. DP-DR and εDP-DR underestimated $\mu^{(1)}$ and overestimated $\mu^{(0)}$. The ATE was estimated better by DP-DR and εDP-DR than by the median-based methods. DP-DR and εDP-DR had the same tendency of

[2] https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

estimation bias and γ: a larger γ value increased the bias.

We briefly discuss the tendency of the estimation bias in real data analysis. The distribution of the outcome of NHEFS data was slightly skewed for each treatment. This skewed distribution violates the assumption that the potential outcomes are distributed marginally symmetrically about their mean. In the case of skewed data, the median is not equal to the mean; therefore, a median-based estimator will be biased for the "average" treatment effect. Our estimators also require symmetry for consistent estimation; however, the influence of asymmetry may be controlled by setting a small γ value. In fact, under asymmetry and no contamination, we can estimate the ATE consistently with $\gamma = 0$, whereas the median-based methods cannot. This flexibility is an advantage of our method. However, this flexibility does make it more difficult to choose the optimal γ under asymmetric settings; the estimates will shift as γ is increased even after removing the effect of outliers, as seen in Table 4.7.

	Target Quantities		
	$\mu^{(1)}$	$\mu^{(0)}$	ATE
IPW (no outliers)	5.221 (-)	1.780 (-)	3.441 (-)
IPW	14.718 (1.57)	11.607 (0.87)	3.111 (1.78)
median (Firpo)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
median (Zhang-IPW)	5.439 (0.21)	2.753 (0.10)	2.686 (0.24)
DP-IPW ($\gamma = 0.05$)	5.597 (0.30)	1.851 (0.07)	3.746 (0.31)
DP-IPW ($\gamma = 0.10$)	5.157 (0.15)	1.819 (0.07)	3.338 (0.17)
DP-IPW ($\gamma = 0.20$)	5.089 (0.15)	1.875 (0.06)	3.215 (0.16)
DP-IPW ($\gamma = 0.50$)	4.949 (0.15)	2.007 (0.06)	2.941 (0.16)
DR (no outliers)	5.136 (-)	1.772 (-)	3.364 (-)
DR	14.574 (1.57)	11.589 (0.90)	2.985 (1.81)
median (Zhang-DR)	5.352 (0.20)	2.743 (0.10)	2.609 (0.22)
median (Sued)	5.353 (0.20)	2.744 (0.10)	2.609 (0.23)
median (TMLE)	5.363 (0.21)	2.739 (0.10)	2.624 (0.23)
DP-DR ($\gamma = 0.05$)	5.478 (0.27)	1.842 (0.07)	3.636 (0.28)
DP-DR ($\gamma = 0.10$)	5.057 (0.16)	1.810 (0.07)	3.248 (0.17)
DP-DR ($\gamma = 0.20$)	4.983 (0.16)	1.865 (0.06)	3.119 (0.17)
DP-DR ($\gamma = 0.50$)	4.834 (0.16)	1.997 (0.06)	2.837 (0.17)
εDP-DR ($\gamma = 0.05$)	5.574 (0.29)	1.851 (0.07)	3.723 (0.30)
εDP-DR ($\gamma = 0.10$)	5.148 (0.15)	1.819 (0.07)	3.330 (0.17)
εDP-DR ($\gamma = 0.20$)	5.080 (0.15)	1.874 (0.06)	3.206 (0.17)
εDP-DR ($\gamma = 0.50$)	4.937 (0.15)	2.007 (0.06)	2.930 (0.16)

Table 4.7: Results of the NHEFS data analysis. Mean and SD are computed on 2,000 bootstrap samples.

4.9 Additional Sources for Outlier-resistant Estimator for ATE

4.9.1 Proof of Theorem 4.1

Proof.

$$\begin{aligned}\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\right] &= \mathbb{E}_g\left[\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\middle| X\right]\right] \\ &= \mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha^*)}\,\mathbb{E}_g\left[h(Y;\mu^{(1)})^\gamma(Y-\mu^{(1)})\middle| T=1,X\right]\right] \\ &= \mathbb{E}_{f_1}\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\right]\end{aligned}$$

The third equality holds because of the causal consistency and the exchangeability. Since $h(y;\mu^{(1)})$ and $f_1(y)$ are symmetric about $\mu^{(1)}$, this expectation is equal to zero:

$$\mathbb{E}_{f_1}\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\right] = \int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})f_1(y)dy = 0.$$

□

4.9.2 Proof of Theorem 4.2

Proof.

$$\begin{aligned}&\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\right] \\ &= \mathbb{E}_g\left[\mathbb{E}_{\tilde{g}}\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\middle| X\right]\right] \\ &= \int\left\{(1-\varepsilon_1(x))\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy + \varepsilon_1(x)\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})\delta_1(y|x)dy\right\}g(x)dx \\ &= \int(1-\varepsilon_1(x))\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\; g(x)dx + \nu_1(\varepsilon_1) \qquad (4.51) \\ &= -\int\varepsilon_1(x)\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)dy\; g(x)dx + \nu_1(\varepsilon_1). \qquad (4.52)\end{aligned}$$

If $\varepsilon_1(x)=\varepsilon_1$, the first term disappears:

$$-\varepsilon_1\iint h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})g(y|x)g(x)dydx = -\varepsilon_1\int h(y;\mu^{(1)})^\gamma(y-\mu^{(1)})f_1(y)dy = 0.$$

□

4.9.3 Proof of Theorem 4.3

Proof. First, we assume that the true PS is given.

$$
\begin{aligned}
&\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)}) - \frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\right] \\
&= \mathbb{E}_g\left[\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)}) - \frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\middle| X\right]\right] \\
&= \mathbb{E}_g\left[-\frac{P(T=1|X)-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\right] \\
&= 0
\end{aligned}
$$

Next, we assume that the true OR model is given.

$$
\begin{aligned}
&\mathbb{E}_g\left[\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha)}(Y-\mu^{(1)}) - \frac{T-\pi(X;\alpha)}{\pi(X;\alpha)}\left\{m_{1,\mu^{(1)}}(X;\beta^*) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta^*)\right\}\right] \\
&= \mathbb{E}_g\left[\mathbb{E}_g\left[\frac{T}{\pi(X;\alpha)}h(Y;\mu^{(1)})^\gamma(Y-\mu^{(1)})\middle| X\right]\right. \\
&\qquad\qquad \left. - \mathbb{E}_g\left[\frac{T}{\pi(X;\alpha)} - 1\middle| X\right]\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right] \\
&= \mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right. \\
&\qquad\qquad \left. - \left(\frac{P(T=1|X)}{\pi(X;\alpha)} - 1\right)\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right] \\
&= \mathbb{E}_g\left[\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right] \\
&= 0
\end{aligned}
$$

Thus, the DP-DR estimating equation has double robustness under no contamination. □

4.9.4 Proof of Theorem 4.4

Proof. If the true PS model is given, the DP-DR estimating equation yields

$$
\begin{aligned}
&\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^{\gamma}}{\pi(X;\alpha^*)}(Y-\mu^{(1)})-\frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta)-\mu^{(1)}m_{2,\mu^{(1)}}(X;\beta)\right\}\right]\\
&=\mathbb{E}_g\left[\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^{\gamma}}{\pi(X;\alpha^*)}(Y-\mu^{(1)})-\frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta)-\mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\middle| X\right]\right]\\
&=\mathbb{E}_g\left[\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^{\gamma}}{\pi(X;\alpha^*)}(Y-\mu^{(1)})\middle| X\right]\right]\\
&\quad\underbrace{-\mathbb{E}_g\left[\frac{P(T=1|X)-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta)-\mu^{(1)}m_{0,\mu^{(1)}}(X;\beta)\right\}\right]}_{=0}\\
&=-\int\varepsilon_1(x)\int h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy\ g(x)dx+\nu_1(\varepsilon_1). \qquad (4.53)
\end{aligned}
$$

If the contamination ratio is independent of X, it holds that

$$-\varepsilon_1\iint h(y;\mu^{(1)})^{\gamma}(y-\mu^{(1)})g(y|x)dy\ g(x)dx=0,$$

which is the same result as that of the DP-IPW estimating equation.

If the true OR model is given, the DP-DR estimating equation yields

$$
\begin{aligned}
&\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\mu^{(1)})^{\gamma}}{\pi(X;\alpha)}(Y-\mu^{(1)})-\frac{T-\pi(X;\alpha)}{\pi(X;\alpha)}\left\{m_{1,\mu^{(1)}}(X;\beta^*)-\mu^{(1)}m_{0,\mu^{(1)}}(X;\beta^*)\right\}\right]\\
&=\mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\Big((1-\varepsilon_1(X))\mathbb{E}_g[h(Y;\mu^{(1)})^{\gamma}(Y-\mu^{(1)})|X]+\varepsilon_1(X)\mathbb{E}_{\delta}[h(Y;\mu^{(1)})^{\gamma}(Y-\mu^{(1)})|X]\Big)\right.\\
&\quad\left.-\left(\frac{P(T=1|X)}{\pi(X;\alpha)}-1\right)\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^{\gamma}(Y^{(1)}-\mu^{(1)})|X]\right]\\
&=\mathbb{E}_g\left[-\varepsilon_1(X)\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^{\gamma}(Y^{(1)}-\mu^{(1)})|X]\right]+\nu_1(\varepsilon_1(\cdot)P(T=1|\cdot)/\pi(\cdot;\alpha)).
\end{aligned}
\qquad (4.54)
$$

When the contamination ratio is independent of X, the first term becomes

$$-\varepsilon_1\mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha)}\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^{\gamma}(Y^{(1)}-\mu^{(1)})|X]\right] \qquad (4.55)$$

Thus, we have the result of Theorem 4.4. □

4.9.5 Derivation of Influence functions in Section 4.4

DP-IPW Let $\tilde{\mu}_n^{(1)}$ denote the root of the DP-IPW estimating equation under contamination.

$$\begin{aligned}
0 = & \frac{\partial}{\partial \varepsilon_1(X_i)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\tilde{g}} \left[\frac{Th(Y;\mu_n^{(1)})^\gamma}{\pi(X_i;\alpha^*)} (Y - \tilde{\mu}_n^{(1)}) \middle| X_i \right] \right\} \Big|_{\varepsilon_1(X_i)=0} \\
= & \frac{\partial}{\partial \varepsilon_1(X_i)} \iint \frac{t}{\pi(X_i;\alpha^*)} h(y;\mu_n^{(1)})^\gamma (y - \tilde{\mu}_n^{(1)}) \{ (1-\varepsilon_1(X_i)) g(y|X_i) \\
& + \varepsilon_1(X_i) \delta_{y_0}(y) \} g(t|X_i) dy dt \big|_{\varepsilon_1(X_i)=0} \\
= & \iint \frac{\partial \psi}{\partial \mu} \Big|_{\mu=\mu_n^{(1)}} g(y|X_i) g(t|X_i) dy dt \cdot IF_{DP-IPW}(y_0) \\
& + \iint \frac{t}{\pi(X_i;\alpha^*)} h(y;\mu_n^{(1)})^\gamma (y - \mu_n^{(1)}) \delta_{y_0}(y) g(t|X_i) dy dt \\
= & \mathbb{E}_g \left[\frac{\partial \psi}{\partial \mu} \Big|_{\mu=\mu_n^{(1)}} \middle| X_i \right] \cdot IF_{DP-IPW}(y_0) + h(y_0;\mu_n^{(1)})^\gamma (y_0 - \mu_n^{(1)})
\end{aligned}$$

If $\mathbb{E}_g \left[\frac{\partial \psi}{\partial \mu} \Big|_{\mu=\mu_n^{(1)}} \middle| X_i \right]$ is invertible, we obtain the IF as

$$IF_{DP-IPW}(y_0) = - \mathbb{E}_g \left[\frac{\partial \psi}{\partial \mu} \Big|_{\mu=\mu_n^{(1)}} \middle| X_i \right]^{-1} h(y_0;\mu_n^{(1)})^\gamma (y_0 - \mu_n^{(1)}). \tag{4.56}$$

DP-DR

$$\begin{aligned}
0 = & \frac{\partial}{\partial \varepsilon_1(X_i)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\tilde{g}} \left[\frac{Th(Y;\tilde{\mu}_n^{(1)})^\gamma}{\pi(X_i;\alpha^*)} (Y - \tilde{\mu}_n^{(1)}) \middle| X_i \right] \right. \\
& \left. - \mathbb{E} \left[\frac{T - \pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)} \middle| X_i \right] \{ m_{1,\tilde{\mu}_n^{(1)}}(X_i) - \tilde{\mu}_n^{(1)} m_{0,\tilde{\mu}_n^{(1)}}(X_i) \} \right\} \Big|_{\varepsilon_1(X_i)=0} \\
= & \frac{\partial}{\partial \varepsilon_1(X_i)} \left\{ \iint \frac{t}{\pi(X_i;\alpha^*)} h(y;\tilde{\mu}_n^{(1)})^\gamma (y - \tilde{\mu}_n^{(1)}) \{ (1-\varepsilon_1(X_i)) g(y|X_i) + \varepsilon_1(X_i) \delta_{y_0}(y) \} g(t|X_i) \right. \\
& \left. - \frac{t - \pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)} \{ m_{1,\tilde{\mu}_n^{(1)}}(X_i) - \tilde{\mu}_n^{(1)} m_{0,\tilde{\mu}_n^{(1)}}(X_i) \} dy dt \right\} \Big|_{\varepsilon_1(X_i)=0} \\
= & \mathbb{E}_g \left[\frac{\partial \psi}{\partial \mu} \Big|_{\mu=\mu_n^{(1)}} \middle| X_i \right] \cdot IF_{DP-DR}(y_0) + \frac{P(T=1|X_i)}{\pi(X_i;\alpha^*)} h(y_0;\mu_n^{(1)})^\gamma (y_0 - \mu_n^{(1)}) \\
& - \frac{P(T=1|X_i) - \pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)} \{ m_{1,\mu_n^{(1)}}(X_i) - \mu_n^{(1)} m_{0,\mu_n^{(1)}}(X_i) \}
\end{aligned}$$

Then, we obtain the IF as

$$IF_{DP-DR}(y_0) = -\mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]^{-1}\left\{\frac{P(T=1|X_i)}{\pi(X_i;\alpha)}h(y_0;\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)}) - \frac{P(T=1|X_i)-\pi(X_i;\alpha)}{\pi(X_i;\alpha)}\{m_{1,\mu_n^{(1)}}(X_i;\beta)-m_{0,\mu_n^{(1)}}(X_i;\beta)\mu_n^{(1)}\}\right\}. \tag{4.57}$$

εDP-DR Suppose the expected contamination ratio is correctly specified as $\bar{\varepsilon}_1 = \sum\varepsilon_1(X_i)/n$.

$$\begin{aligned}0 = &\frac{\partial}{\partial\varepsilon_1(X_i)}\left\{\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\tilde{\mu}_n^{(1)})^\gamma}{\pi(X_i;\alpha^*)}(Y-\tilde{\mu}_n^{(1)})\middle| X_i\right]\right.\\ &\left.-\left(1-\frac{1}{n}\sum_{i=1}^n\varepsilon_1(X_i)\right)\mathbb{E}\left[\frac{T-\pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)}\middle| X_i\right]\{m_{1,\tilde{\mu}_n^{(1)}}(X_i)-\tilde{\mu}_n^{(1)}m_{0,\tilde{\mu}_n^{(1)}}(X_i)\}\right\}\Bigg|_{\varepsilon_1(X_i)=0}\end{aligned}$$

$$\begin{aligned}= &\frac{\partial}{\partial\varepsilon_1(X_i)}\left\{\iint\frac{t}{\pi(X_i;\alpha^*)}h(y;\tilde{\mu}_n^{(1)})^\gamma(y-\tilde{\mu}_n^{(1)})\{(1-\varepsilon_1(X_i))g(y|X_i)+\varepsilon_1(X_i)\delta_{y_0}(y)\}g(t|X_i)\right.\\ &\left.-\left(1-\frac{1}{n}\sum_{i=1}^n\varepsilon_1(X_i)\right)\frac{t-\pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)}\{m_{1,\tilde{\mu}_n^{(1)}}(X_i)-\tilde{\mu}_n^{(1)}m_{0,\mu_n^{(1)}}(X_i)\}dydt\right\}\Bigg|_{\varepsilon_1(X_i)=0}\\ = &\ \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\mu}\right|_{\mu=\mu_n^{(1)}}\middle| X_i\right]\cdot IF_{\varepsilon DP-DR}(y_0)+\frac{P(T=1|X_i)}{\pi(X_i;\alpha^*)}h(y_0;\mu_n^{(1)})^\gamma(y_0-\mu_n^{(1)})\\ &-\frac{n-1}{n}\frac{P(T=1|X_i)-\pi(X_i;\alpha^*)}{\pi(X_i;\alpha^*)}\{m_{1,\mu_n^{(1)}}(X_i)-\mu_n^{(1)}m_{0,\mu_n^{(1)}}(X_i)\}\end{aligned}$$

Thus, we obtain (4.24).

4.9.6 Influence Functions Under Homogeneous Contamination

Under homogeneous contamination, we can apply the ordinary IF analysis. By differentiating the estimating equations with respect to ε_1 at $\varepsilon_1 = 0$, we obtain the following results.

DP-IPW

$$\begin{aligned}0 = &\ \frac{\partial}{\partial\varepsilon_1}\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\tilde{\mu}^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\tilde{\mu}^{(1)})\right]\Bigg|_{\varepsilon_1=0}\\ 0 = &\ \frac{\partial}{\partial\varepsilon_1}\iiint\frac{th(y;\tilde{\mu}^{(1)})^\gamma}{\pi(x;\alpha^*)}(y-\tilde{\mu}^{(1)})\{(1-\varepsilon_1)g(y|t,x)+\varepsilon_1\delta_{y_0}(y|x)\}g(t|x)g(x)dydtdx\Bigg|_{\varepsilon_1=0}\\ = &\ \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]\cdot IF_{DP-IPW}(y_0)+\iiint\frac{th(y;\mu^{(1)})^\gamma}{\pi(x;\alpha^*)}(y-\mu^{(1)})\delta_{y_0}(y|x)g(t|x)g(x)dydtdx\end{aligned}$$

$$IF_{DP-IPW}(y_0) = \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]^{-1} h(y_0;\mu^{(1)})^\gamma(y_0-\mu^{(1)}) \tag{4.58}$$

Thus, DP-IPW has a redescending property under homogeneous contamination.

DP-DR

$$\begin{aligned}
0 = &\ \frac{\partial}{\partial\varepsilon_1}\mathbb{E}_{\tilde{g}}\left[\frac{Th(Y;\tilde{\mu}^{(1)})^\gamma}{\pi(X;\alpha)}(Y-\tilde{\mu}^{(1)}) - \frac{T-\pi(X;\alpha)}{\pi(X;\alpha)}\left(m_{1,\tilde{\mu}^{(1)}}(X;\beta) - \tilde{\mu}^{(1)}m_{0,\tilde{\mu}^{(1)}}(X;\beta)\right)\right]\Bigg|_{\varepsilon_1=0} \\
= &\ \frac{\partial}{\partial\varepsilon_1}\iiint\left(\frac{th(y;\tilde{\mu}^{(1)})^\gamma}{\pi(x;\alpha)}(y-\tilde{\mu}^{(1)}) - \frac{t-\pi(x;\alpha)}{\pi(x;\alpha)}\left\{m_{1,\tilde{\mu}^{(1)}}(x;\beta) - \tilde{\mu}^{(1)}m_{0,\tilde{\mu}^{(1)}}(x;\beta)\right\}\right) \\
&\qquad \times\{(1-\varepsilon_1)g(y|t,x) + \varepsilon_1\delta_{y_0}(y|x)\}g(t|x)g(x)dydtdx\big|_{\varepsilon_1=0} \\
= &\ \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]\cdot IF_{DP-DR}(y_0) + \iiint\left(\frac{th(y;\mu^{(1)})^\gamma}{\pi(x;\alpha)}(y-\mu^{(1)})\right. \\
&\left. - \frac{t-\pi(x;\alpha)}{\pi(x;\alpha)}\left\{m_{1,\mu^{(1)}}(x;\beta) - \mu^{(1)}m_{0,\mu^{(1)}}(x;\beta)\right\}\right)\delta_{y_0}(y|x)g(t|x)g(x)dydtdx
\end{aligned}$$

$$\begin{aligned}
IF_{DP-DR}(y_0) = \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]^{-1}\iint\left(\frac{th(y_0;\mu^{(1)})^\gamma}{\pi(x;\alpha)}(y_0-\mu^{(1)})\right. \\
\left. -\frac{t-\pi(x;\alpha)}{\pi(x;\alpha)}\left\{m_{1,\mu^{(1)}}(x;\beta) - \mu^{(1)}m_{0,\mu^{(1)}}(x;\beta)\right\}\right)g(t|x)g(x)dtdx
\end{aligned} \tag{4.59}$$

If the true PS model is given, this IF yields

$$IF_{DP-DR}(y_0) = \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]^{-1} h(y_0;\mu^{(1)})^\gamma(y_0-\mu^{(1)}) \tag{4.60}$$

If the true OR model is given, this IF yields

$$\begin{aligned}
IF_{DP-DR}(y_0) = \mathbb{E}_g\left[\left.\frac{\partial\psi}{\partial\varepsilon_1}\right|_{\mu=\mu^{(1)}}\right]^{-1}\int\frac{P(T=1|x)}{\pi(x;\alpha)}h(y_0;\mu^{(1)})^\gamma(y_0-\mu^{(1)}) \\
-\frac{P(T=1|x)-\pi(x;\alpha)}{\pi(x;\alpha)}\mathbb{E}[h(Y;\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|x]g(x)dx
\end{aligned} \tag{4.61}$$

Thus, DP-DR has a redescending property under homogeneous contamination in the PS-correct case.

εDP-DR

$$
\begin{aligned}
0 = \; & \frac{\partial}{\partial \varepsilon_1} \mathbb{E}_{\tilde{g}} \left[\frac{T h(Y;\tilde{\mu}^{(1)})^{\gamma}}{\pi(X;\alpha)} (Y - \tilde{\mu}^{(1)}) \right. \\
& \left. - \frac{T - \pi(X;\alpha)}{\pi(X;\alpha)} (1-\varepsilon_1) \left(m_{1,\tilde{\mu}^{(1)}}(X;\beta) - \tilde{\mu}^{(1)} m_{0,\tilde{\mu}^{(1)}}(X;\beta) \right) \right] \Bigg|_{\varepsilon_1 = 0} \\
= \; & \mathbb{E}_g \left[\left. \frac{\partial \psi}{\partial \varepsilon_1} \right|_{\mu = \mu^{(1)}} \right] \cdot IF_{\varepsilon DP-DR}(y_0) \\
& + \iiint \left(\frac{t h(y;\mu^{(1)})^{\gamma}}{\pi(x;\alpha)} (y - \mu^{(1)}) - \frac{t - \pi(x;\alpha)}{\pi(x;\alpha)} \left\{ m_{1,\mu^{(1)}}(x;\beta) - \mu^{(1)} m_{0,\mu^{(1)}}(x;\beta) \right\} \right. \\
& \left. + \frac{t - \pi(x;\alpha)}{\pi(x;\alpha)} \left\{ m_{1,\mu^{(1)}}(x;\beta) - \tilde{\mu}^{(1)} m_{0,\mu^{(1)}}(x;\beta) \right\} \right) \delta_{y_0}(y|x) g(t|x) g(x) dy dt dx
\end{aligned}
$$

$$
IF_{\varepsilon DP-DR}(y_0) = \mathbb{E}_g \left[\left. \frac{\partial \psi}{\partial \varepsilon_1} \right|_{\mu = \mu^{(1)}} \right]^{-1} \int \frac{P(T=1|x)}{\pi(x;\alpha)} h(y_0;\mu^{(1)})^{\gamma} (y_0 - \mu^{(1)}) g(x) dx \quad (4.62)
$$

Thus, under homogeneous contamination, εDP-DR has a redescending property in either the PS-correct case or the OR-correct case.

4.9.7 Regularity Conditions for Theorem 4.5

Detailed discussion is available in Chapter 5 of Van der Vaart (2000), for example.

(a) The function $S(\lambda)$ is twice continuously differentiable with respect to λ.

(b) There exists a root λ^* of $\mathbb{E}_{\tilde{g}}[S(\lambda)] = 0$.

(c) $\mathbb{E}_{\tilde{g}}[\|S(\lambda^*)\|^2] < \infty$.

(d) $\mathbb{E}_{\tilde{g}}[\partial S(\lambda^*)/\partial \lambda^T]$ exists and is nonsingular.

(e) The second-order differentials of $S(\lambda)$ with respect to μ are dominated by a fixed integrable function h in a neighborhood of λ^*.

4.9.8 Proof of Theorem 4.6

Under homogeneous contamination, we see that simpler properties hold. The matrix $\mathbf{J}^{\tilde{g}}(\lambda^*)$ is partitioned as

$$\mathbf{J}^{\tilde{g}}(\lambda^*) = \begin{pmatrix} \mathbb{E}_{\tilde{g}}\left[\frac{\partial}{\partial\mu}\psi_i(\mu^*;\alpha^*,\beta^*)\right] & \mathbb{E}_{\tilde{g}}\left[\frac{\partial}{\partial\alpha^T}\psi_i(\mu^*;\alpha^*,\beta^*)\right] & \mathbb{E}_{\tilde{g}}\left[\frac{\partial}{\partial\beta^T}\psi_i(\mu^*;\alpha^*,\beta^*)\right] \\ \mathbf{0} & \mathbb{E}_g\left[\frac{\partial}{\partial\alpha^T}s_i^{PS}(\alpha^*)\right] & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{E}_{\tilde{g}}\left[\frac{\partial}{\partial\beta^T}s_i^{OR}(\beta^*)\right] \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{J}_{11}^{\tilde{g}}(\lambda^*) & \mathbf{J}_{12}^{\tilde{g}}(\lambda^*) & \mathbf{J}_{13}^{\tilde{g}}(\lambda^*) \\ \mathbf{0} & \mathbf{J}_{22}^{g}(\lambda^*) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{33}^{\tilde{g}}(\lambda^*) \end{pmatrix}.$$

If it is nonsingular, the inverse is obtained as

$$\mathbf{J}^{\tilde{g}}(\lambda^*)^{-1} = \begin{pmatrix} \mathbf{J}_{11}^{\tilde{g}}(\lambda^*)^{-1} & -\mathbf{J}_{11}^{\tilde{g}}(\lambda^*)^{-1}\mathbf{J}_{12}^{\tilde{g}}(\lambda^*)\mathbf{J}_{22}^{g}(\lambda^*)^{-1} & -\mathbf{J}_{11}^{\tilde{g}}(\lambda^*)^{-1}\mathbf{J}_{13}^{\tilde{g}}(\lambda^*)\mathbf{J}_{33}^{\tilde{g}}(\lambda^*)^{-1} \\ \mathbf{0} & \mathbf{J}_{22}^{g}(\lambda^*)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{33}^{\tilde{g}}(\lambda^*)^{-1} \end{pmatrix}.$$

Note that $\mathbf{J}_{11}^{\tilde{g}}(\cdot)$ is a scalar value.

Then, Theorem 4.6 is proved as follows.

Proof. By Taylor's theorem, the expectation of estimating equation (5.26) is expressed as

$$0 = \mathbb{E}[S_i(\lambda^*)] = \mathbb{E}[S_i(\lambda^{**})] + \mathbf{J}^{\tilde{g}}(\lambda^\dagger)(\lambda^* - \lambda^{**}),$$

where $\lambda^\dagger$ is an intermediate value between λ^{**} and λ^*. Since $\mathbb{E}[s_i^{PS}(\alpha^*)] = \mathbb{E}[s_i^{OR}(\beta^*)] = 0$ and $(\lambda^* - \lambda^{**}) = (\mu^* - \mu^{(1)}, \mathbf{0}^T, \mathbf{0}^T)^T$, only the first element is meaningful:

$$0 = \mathbb{E}[\psi_i(\mu^{(1)};\alpha^*,\beta^*)] + \mathbf{J}_{11}^{\tilde{g}}(\lambda^\dagger)(\mu^* - \mu^{(1)}).$$

Then, since $\mathbf{J}_{11}^{\tilde{g}}(\lambda^\dagger)$ is non-zero, the latent bias of μ^* reduces to

$$\mu^* - \mu^{(1)} = -\mathbf{J}_{11}^{\tilde{g}}(\lambda^\dagger)^{-1}\mathbb{E}[\psi_i(\mu^{(1)};\alpha^*,\beta^*)]. \tag{4.63}$$

From Corollary 4.1, if either the PS or the OR model is correct, we have

$$\mathbb{E}[\psi_i(\mu^{(1)};\alpha^*,\beta^*)] = \nu_1(\phi).$$

Upon substituting it into (4.63), the statement holds. □

4.9.9 Further Discussion on Asymptotic Variance

Considering the structure of the full estimating equation, the asymptotic variance can be expressed in a more explicit form. The discussion about the asymptotic variance is provided

in the next subsection.

The matrix $\mathbf{K}_{\tilde{g}}(\lambda^*)$ is also partitioned as

$$\mathbf{K}^{\tilde{g}}(\lambda^*) = \begin{pmatrix} \mathbf{K}_{11}^{\tilde{g}}(\lambda^*) & \mathbf{K}_{12}^{\tilde{g}}(\lambda^*) & \mathbf{K}_{13}^{\tilde{g}}(\lambda^*) \\ \mathbf{K}_{12}^{\tilde{g}\,T}(\lambda^*) & \mathbf{K}_{22}^{g}(\lambda^*) & \mathbf{K}_{23}^{\tilde{g}}(\lambda^*) \\ \mathbf{K}_{13}^{\tilde{g}\,T}(\lambda^*) & \mathbf{K}_{23}^{\tilde{g}\,T}(\lambda^*) & \mathbf{K}_{33}^{\tilde{g}}(\lambda^*) \end{pmatrix}.$$

The asymptotic variance is also affected by outliers. However, under Assumption 1, the asymptotic variance can be approximated by the asymptotic variance under no contamination and contamination ratio ε_1.

Theorem 4.7. *Besides to Assumption 1, assume that* $\mathbf{J}_{1m}^{\delta}(\lambda^{**}) \approx \mathbf{0}$ *and* $\mathbf{K}_{1m}^{\delta}(\lambda^{**}) \approx \mathbf{0}$ *holds for* $m = 1, 2, 3$. *Then, under homogeneous contamination,*

$$\mathbf{V}^{\tilde{g}}(\lambda^*) \approx \mathbf{J}^{\check{g}}(\lambda^{**})^{-1} \begin{pmatrix} \frac{1}{(1-\varepsilon_1)}\mathbf{K}_{11}^{g}(\lambda^{**}) & \mathbf{K}_{12}^{g}(\lambda^{**}) & \mathbf{K}_{13}^{g}(\lambda^{**}) \\ \mathbf{K}_{12}^{g}(\lambda^{**})^T & \mathbf{K}_{22}^{g}(\lambda^{**}) & \mathbf{K}_{23}^{\tilde{g}}(\lambda^{**}) \\ \mathbf{K}_{13}^{g}(\lambda^{**})^T & \mathbf{K}_{23}^{\tilde{g}}(\lambda^{**})^T & \mathbf{K}_{33}^{\tilde{g}}(\lambda^{**}) \end{pmatrix} \{\mathbf{J}^{\check{g}}(\lambda^{**})^T\}^{-1},$$

where

$$\mathbf{J}^{\check{g}}(\lambda^{**}) = \begin{pmatrix} \mathbf{J}_{11}^{g}(\lambda^{**}) & \mathbf{J}_{12}^{g}(\lambda^{**}) & \mathbf{J}_{13}^{g}(\lambda^{**}) \\ \mathbf{0} & \mathbf{J}_{22}^{g}(\lambda^{**}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{33}^{\tilde{g}}(\lambda^{**}) \end{pmatrix}$$

.

If both the PS and the OR models are correct, the asymptotic variance of $\hat{\mu}$ has a simpler expression. From a similar discussion to Section 4.3, the following lemma holds:

Lemma 4.1. *If the PS model is correct,* $\mathbf{J}_{13}^{g}(\lambda^{**}) = \mathbf{0}$. *If the OR model is correct,* $\mathbf{J}_{12}^{g}(\lambda^{**}) = \mathbf{0}$.

Using Lemma 4.1, we can see the asymptotic variance of $\hat{\mu}$ is simply expressed.

Theorem 4.8. *Under the same assumptions of Theorem 4.7, if the PS and the OR models are both correct, then*

$$\mathbf{V}_{\tilde{g}}(\mu^*) \approx \frac{1}{1-\varepsilon_1}\mathbf{J}_{11}^{g}(\lambda^{**})^{-1}\mathbf{K}_{11}^{g}(\lambda^{**})\left\{\mathbf{J}_{11}^{g}(\lambda^{**})^T\right\}^{-1}.$$

Proof. By applying Lemma 4.1 to Theorem 4.7, the statement holds. □

This implies that the εDP-DR appropriately ignores outliers.

Proof of Theorem 4.7 If either the PS or the OR model is correct, we can say $\mu^* \approx \mu^{(1)}$. Note that the PS model is not related to the contamination distribution δ, and

the contamination in the OR model cannot be removed in general. From the assumptions,

$$
\begin{aligned}
\mathbf{J}^{\tilde{g}}(\lambda^*) &\approx \mathbf{J}^{\tilde{g}}(\lambda^{**}) \\
&= \begin{pmatrix} (1-\varepsilon_1)\mathbf{J}^g_{11} & (1-\varepsilon_1)\mathbf{J}^g_{12} & (1-\varepsilon_1)\mathbf{J}^g_{13} \\ \mathbf{0} & \mathbf{J}^g_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}^{\tilde{g}}_{33} \end{pmatrix} + \begin{pmatrix} \varepsilon_1\mathbf{J}^\delta_{11} & \varepsilon_1\mathbf{J}^\delta_{12} & \varepsilon_1\mathbf{J}^\delta_{13} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \\
&\approx \begin{pmatrix} (1-\varepsilon_1)\mathbf{J}^g_{11} & (1-\varepsilon_1)\mathbf{J}^g_{12} & (1-\varepsilon_1)\mathbf{J}^g_{13} \\ \mathbf{0} & \mathbf{J}^g_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}^{\tilde{g}}_{33} \end{pmatrix} \\
&= \begin{pmatrix} 1-\varepsilon_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_\alpha & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_\beta \end{pmatrix} \mathbf{J}^{\check{g}}(\lambda^{**})
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{K}^{\tilde{g}}(\lambda^*) &\approx \mathbf{K}^{\tilde{g}}(\lambda^{**}) \\
&= \begin{pmatrix} (1-\varepsilon_1)\mathbf{K}^g_{11} & (1-\varepsilon_1)\mathbf{K}^g_{12} & (1-\varepsilon_1)\mathbf{K}^g_{13} \\ (1-\varepsilon_1)\mathbf{K}^{g\ T}_{12} & \mathbf{K}^g_{22} & \mathbf{K}^{\tilde{g}}_{23} \\ (1-\varepsilon_1)\mathbf{K}^{g\ T}_{13} & \mathbf{K}^{\tilde{g}\ T}_{23} & \mathbf{K}^{\tilde{g}}_{33} \end{pmatrix} + \begin{pmatrix} \varepsilon_1\mathbf{K}^\delta_{11} & \varepsilon_1\mathbf{K}^\delta_{12} & \varepsilon_1\mathbf{K}^{\delta T}_{13} \\ \varepsilon_1\mathbf{K}^{\delta\ T}_{12} & \mathbf{0} & \mathbf{0} \\ \varepsilon_1\mathbf{K}^{\delta\ T}_{13} & \mathbf{0} & \mathbf{0} \end{pmatrix} \\
&\approx \begin{pmatrix} (1-\varepsilon_1)\mathbf{K}^g_{11} & (1-\varepsilon_1)\mathbf{K}^g_{12} & (1-\varepsilon_1)\mathbf{K}^g_{13} \\ (1-\varepsilon_1)\mathbf{K}^{g\ T}_{12} & \mathbf{K}^g_{22} & \mathbf{K}^{\tilde{g}}_{23} \\ (1-\varepsilon_1)\mathbf{K}^{g\ T}_{13} & \mathbf{K}^{\tilde{g}\ T}_{23} & \mathbf{K}^{\tilde{g}}_{33} \end{pmatrix}
\end{aligned}
$$

The input (λ^{**}) is dropped for notation simplicity. Thus, we have

$$
\mathbf{V}^{\tilde{g}}(\lambda^*) \approx \mathbf{J}^{\check{g}}(\lambda^{**})^{-1} \begin{pmatrix} \frac{1}{(1-\varepsilon_1)}\mathbf{K}^g_{11}(\lambda^{**}) & \mathbf{K}^g_{12}(\lambda^{**}) & \mathbf{K}^g_{13}(\lambda^{**}) \\ \mathbf{K}^g_{12}(\lambda^{**})^T & \mathbf{K}^g_{22}(\lambda^{**}) & \mathbf{K}^{\tilde{g}}_{23}(\lambda^{**}) \\ \mathbf{K}^g_{13}(\lambda^{**})^T & \mathbf{K}^{\tilde{g}}_{23}(\lambda^{**})^T & \mathbf{K}^{\tilde{g}}_{33}(\lambda^{**}) \end{pmatrix} \{\mathbf{J}^{\check{g}}(\lambda^{**})^T\}^{-1},
$$

The proof is complete.

Proof of Lemma 4.1

Proof. In the PS correct case,

$$
\begin{aligned}
\mathbf{J}^g_{13}(\lambda^{**}) &= \mathbb{E}_g\left[\frac{\partial}{\partial\beta^T}\left\{\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)}) - \frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\left\{m_{1,\mu^{(1)}}(X;\beta^*) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta^*)\right\}\right\}\right] \\
&= \mathbb{E}_g\left[\frac{P(T=1|X)}{\pi(X;\alpha^*)}h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\right. \\
&\qquad \left. -\frac{P(T=1|X)-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\frac{\partial}{\partial\beta^T}\left\{m_{1,\mu^{(1)}}(X;\beta^*) - \mu^{(1)}m_{0,\mu^{(1)}}(X;\beta^*)\right\}\right] \\
&= \mathbb{E}_g\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\right] \\
&= 0.
\end{aligned}
$$

In the OR correct case,

$$\begin{aligned}
\mathbf{J}_{12}^{g}(\lambda^{**}) &= \mathbb{E}_g\left[\frac{\partial}{\partial\alpha^T}\left\{\frac{Th(Y;\mu^{(1)})^\gamma}{\pi(X;\alpha^*)}(Y-\mu^{(1)}) - \frac{T-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\mathbb{E}_g\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X\right]\right\}\right] \\
&= \mathbb{E}_g\left[\frac{\partial}{\partial\alpha^T}\left(\frac{P(T=1|X)}{\pi(X;\alpha^*)} - \frac{P(T=1|X)-\pi(X;\alpha^*)}{\pi(X;\alpha^*)}\right)\mathbb{E}_g[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})|X]\right] \\
&= \mathbb{E}_g\left[h(Y^{(1)};\mu^{(1)})^\gamma(Y^{(1)}-\mu^{(1)})\right] \\
&= 0.
\end{aligned}$$

□

4.9.10 Remaining Results of Monte-Carlo Simulation

Remaining results of the Monte-Carlo Simulation are presented in the following pages.

- Tables 4.8 and 4.9: Gaussian covariates and Gaussian MLE with non-outliers for OR. The RMSE is presented in Tables 4.2 and 4.3 in Section 4.7.
- Tables 4.10 to 4.12: Gaussian covariates and unnormalized Gaussian modeling for OR.
- Tables 4.13 to 4.15: Uniform covariates and Gaussian MLE with non-outliers for OR.
- Tables 4.16 to 4.18: Uniform covariates and unnormalized Gaussian modeling for OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	3.004 (0.22)	3.749 (0.59)	4.493 (0.78)	5.983 (1.02)	3.766 (0.63)	4.536 (0.84)	6.070 (1.08)
	median (Firpo)	2.990 (0.26)	3.091 (0.28)	3.205 (0.30)	3.490 (0.43)	3.116 (0.28)	3.259 (0.32)	3.605 (0.47)
	median (Zhang-IPW)	2.990 (0.26)	3.091 (0.28)	3.205 (0.30)	3.490 (0.43)	3.116 (0.28)	3.259 (0.32)	3.605 (0.47)
	DP-IPW ($\gamma = 0.1$)	2.998 (0.22)	3.030 (0.27)	3.142 (0.51)	4.492 (1.70)	3.056 (0.29)	3.209 (0.57)	4.620 (1.74)
	DP-IPW ($\gamma = 0.5$)	2.986 (0.23)	2.987 (0.25)	2.989 (0.27)	3.052 (0.64)	3.011 (0.24)	3.042 (0.28)	3.173 (0.70)
	DP-IPW ($\gamma = 1.0$)	2.980 (0.26)	2.978 (0.27)	2.977 (0.27)	2.990 (0.41)	3.003 (0.26)	3.033 (0.28)	3.114 (0.48)
DR(T/T)	Naive	2.999 (0.18)	3.745 (0.60)	4.489 (0.79)	5.979 (1.04)	3.762 (0.64)	4.533 (0.86)	6.069 (1.12)
	median (Zhang-DR)	2.994 (0.24)	3.096 (0.30)	3.210 (0.33)	3.499 (0.54)	3.121 (0.31)	3.264 (0.37)	3.620 (0.66)
	median (Sued)	2.994 (0.24)	3.096 (0.30)	3.209 (0.33)	3.496 (0.48)	3.121 (0.31)	3.264 (0.36)	3.616 (0.61)
	median (TMLE)	2.994 (0.24)	3.095 (0.26)	3.208 (0.29)	3.479 (0.37)	3.120 (0.27)	3.260 (0.31)	3.587 (0.38)
	DP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.029 (0.30)	3.140 (0.55)	4.465 (1.72)	3.054 (0.31)	3.207 (0.62)	4.604 (1.78)
	DP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.997 (0.29)	3.000 (0.33)	3.060 (0.69)	3.022 (0.27)	3.053 (0.34)	3.195 (0.81)
	DP-DR ($\gamma = 1.0$)	2.992 (0.24)	2.991 (0.29)	2.992 (0.31)	3.009 (0.52)	3.017 (0.29)	3.047 (0.33)	3.137 (0.66)
	εDP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.028 (0.29)	3.138 (0.54)	4.464 (1.72)	3.054 (0.31)	3.204 (0.60)	4.604 (1.77)
	εDP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.997 (0.26)	2.999 (0.30)	3.058 (0.67)	3.022 (0.27)	3.052 (0.32)	3.190 (0.77)
	εDP-DR ($\gamma = 1.0$)	2.992 (0.24)	2.991 (0.29)	2.991 (0.30)	3.007 (0.51)	3.017 (0.29)	3.047 (0.33)	3.134 (0.63)
DR(T/F)	Naive	3.004 (0.24)	3.750 (0.60)	4.494 (0.78)	5.984 (1.03)	3.767 (0.64)	4.537 (0.85)	6.073 (1.09)
	median (Zhang-DR)	2.990 (0.27)	3.092 (0.33)	3.208 (0.35)	3.499 (0.55)	3.118 (0.33)	3.262 (0.38)	3.623 (0.67)
	median (Sued)	2.989 (0.27)	3.091 (0.33)	3.206 (0.35)	3.493 (0.50)	3.117 (0.33)	3.261 (0.38)	3.616 (0.62)
	median (TMLE)	2.999 (0.24)	3.100 (0.27)	3.214 (0.29)	3.496 (0.38)	3.125 (0.27)	3.267 (0.30)	3.607 (0.39)
	DP-DR ($\gamma = 0.1$)	2.998 (0.24)	3.033 (0.31)	3.150 (0.54)	4.490 (1.71)	3.060 (0.32)	3.218 (0.61)	4.624 (1.76)
	DP-DR ($\gamma = 0.5$)	2.986 (0.25)	2.989 (0.32)	2.992 (0.35)	3.059 (0.71)	3.014 (0.32)	3.044 (0.36)	3.196 (0.82)
	DP-DR ($\gamma = 1.0$)	2.979 (0.28)	2.978 (0.33)	2.979 (0.35)	3.001 (0.58)	3.004 (0.33)	3.035 (0.37)	3.133 (0.70)
	εDP-DR ($\gamma = 0.1$)	2.998 (0.24)	3.033 (0.31)	3.149 (0.54)	4.489 (1.71)	3.059 (0.32)	3.218 (0.60)	4.623 (1.75)
	εDP-DR ($\gamma = 0.5$)	2.986 (0.25)	2.989 (0.32)	2.992 (0.34)	3.057 (0.69)	3.013 (0.31)	3.044 (0.35)	3.192 (0.79)
	εDP-DR ($\gamma = 1.0$)	2.979 (0.28)	2.978 (0.33)	2.978 (0.34)	2.998 (0.55)	3.004 (0.33)	3.035 (0.37)	3.132 (0.70)
DR(F/T)	Naive	2.999 (0.18)	3.725 (0.50)	4.451 (0.65)	5.902 (0.86)	3.667 (0.49)	4.341 (0.65)	5.685 (0.84)
	median (Zhang-DR)	3.003 (0.24)	3.081 (0.25)	3.169 (0.27)	3.388 (0.32)	3.096 (0.25)	3.200 (0.27)	3.445 (0.32)
	median (Sued)	2.999 (0.24)	3.101 (0.25)	3.213 (0.27)	3.494 (0.34)	3.111 (0.25)	3.237 (0.28)	3.532 (0.33)
	median (TMLE)	3.003 (0.23)	3.079 (0.25)	3.164 (0.26)	3.368 (0.30)	3.093 (0.25)	3.194 (0.26)	3.424 (0.30)
	DP-DR ($\gamma = 0.1$)	2.999 (0.18)	2.997 (0.19)	3.051 (0.34)	4.326 (1.57)	3.018 (0.19)	3.077 (0.29)	4.000 (1.35)
	DP-DR ($\gamma = 0.5$)	3.001 (0.20)	2.975 (0.20)	2.950 (0.21)	2.907 (0.35)	3.000 (0.20)	3.000 (0.21)	3.008 (0.28)
	DP-DR ($\gamma = 1.0$)	3.005 (0.23)	2.978 (0.23)	2.953 (0.23)	2.895 (0.25)	3.004 (0.23)	3.006 (0.23)	3.008 (0.24)
	εDP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.020 (0.19)	3.108 (0.37)	4.541 (1.58)	3.040 (0.19)	3.131 (0.31)	4.209 (1.39)
	εDP-DR ($\gamma = 0.5$)	3.001 (0.20)	2.998 (0.20)	2.998 (0.21)	3.020 (0.38)	3.022 (0.20)	3.048 (0.21)	3.117 (0.30)
	εDP-DR ($\gamma = 1.0$)	3.005 (0.23)	3.001 (0.23)	3.001 (0.23)	3.003 (0.24)	3.027 (0.23)	3.054 (0.23)	3.116 (0.23)

Table 4.8: Mean and SD of 10,000 simulated estimates of $\mu^{(1)}$. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the Gaussian MLE using non-outliers. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.009	0.010	0.012	0.009	0.010	0.010	0.012
	median (Firpo)	2.019	2.035	2.052	2.040	2.049	2.049	2.044
	median (Zhang-IPW)	0.153	0.166	0.165	0.162	0.194	0.195	0.203
	DP-IPW ($\gamma = 0.1$)	33.713	44.646	62.069	80.462	44.256	60.934	78.925
	DP-IPW ($\gamma = 0.5$)	48.822	50.308	48.108	46.894	49.748	48.500	47.246
	DP-IPW ($\gamma = 1.0$)	65.508	64.401	62.285	55.019	63.856	61.479	54.070
DR(T/T)	Naive	0.014	0.015	0.014	0.013	0.014	0.020	0.014
	median (Zhang-DR)	0.416	0.417	0.422	0.412	0.406	0.424	0.416
	median (Sued)	1.720	1.893	1.689	1.825	1.806	1.710	1.723
	median (TMLE)	1067.234	1050.107	1021.848	1010.238	1046.995	1015.752	1000.908
	DP-DR ($\gamma = 0.1$)	20.316	55.933	131.469	209.206	57.353	141.999	202.042
	DP-DR ($\gamma = 0.5$)	54.524	47.142	47.741	39.424	50.581	43.799	37.860
	DP-DR ($\gamma = 1.0$)	83.912	81.759	74.653	58.289	78.730	73.391	60.144
	εDP-DR ($\gamma = 0.1$)	19.997	59.061	132.762	212.343	57.435	137.271	205.093
	εDP-DR ($\gamma = 0.5$)	54.806	50.210	44.774	39.729	50.144	46.977	37.970
	εDP-DR ($\gamma = 1.0$)	84.375	79.237	73.203	56.916	78.328	76.474	60.338
DR(T/F)	Naive	0.014	0.015	0.013	0.014	0.016	0.013	0.014
	median (Zhang-DR)	0.337	0.370	0.358	0.354	0.375	0.371	0.364
	median (Sued)	1.646	1.829	1.705	1.754	1.753	1.893	3.771
	median (TMLE)	999.657	991.569	983.834	991.174	997.403	994.754	991.656
	DP-DR ($\gamma = 0.1$)	18.851	63.892	137.795	198.849	57.978	138.945	195.932
	DP-DR ($\gamma = 0.5$)	52.547	49.129	45.949	40.204	53.554	47.560	39.319
	DP-DR ($\gamma = 1.0$)	80.465	81.757	72.929	60.123	79.196	73.844	56.672
	εDP-DR ($\gamma = 0.1$)	18.603	59.010	133.449	203.896	60.482	138.275	190.823
	εDP-DR ($\gamma = 0.5$)	52.892	53.978	49.416	39.773	53.920	49.583	37.347
	εDP-DR ($\gamma = 1.0$)	80.870	81.030	73.614	59.259	78.651	72.015	59.778
DR(F/T)	Naive	0.013	0.014	0.014	0.014	0.014	0.015	0.015
	median (Zhang-DR)	0.399	0.417	0.415	2.122	0.428	0.419	0.440
	median (Sued)	1.607	1.876	1.766	1.766	1.751	1.902	1.836
	median (TMLE)	951.348	975.007	970.882	986.397	970.804	979.606	988.658
	DP-DR ($\gamma = 0.1$)	20.537	58.989	143.063	218.114	54.900	123.918	248.367
	DP-DR ($\gamma = 0.5$)	52.096	54.095	45.947	39.811	56.126	48.108	40.617
	DP-DR ($\gamma = 1.0$)	85.026	84.556	74.602	59.838	84.897	80.306	65.958
	εDP-DR ($\gamma = 0.1$)	20.352	60.387	147.921	201.602	54.002	131.749	226.009
	εDP-DR ($\gamma = 0.5$)	52.473	50.362	47.613	41.028	51.504	45.609	41.540
	εDP-DR ($\gamma = 1.0$)	85.557	83.889	75.631	64.173	84.836	77.640	64.876

Table 4.9: Mean computation time (ms) of 10,000 simulations. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the Gaussian MLE using non-outliers. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.222	0.957	1.683	3.153	0.993	1.752	3.253
	median (Firpo)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
	median (Zhang-IPW)	0.257	0.294	0.367	0.649	0.306	0.409	0.769
	DP-IPW ($\gamma = 0.1$)	0.218	0.276	0.531	2.263	0.293	0.609	2.377
	DP-IPW ($\gamma = 0.5$)	0.227	0.249	0.272	0.639	0.245	0.287	0.726
	DP-IPW ($\gamma = 1.0$)	0.261	0.271	0.275	0.413	0.262	0.281	0.498
DR(T/T)	Naive	0.185	0.957	1.683	3.154	0.997	1.758	3.265
	median (Zhang-DR)	0.242	0.317	0.391	0.733	0.330	0.452	0.905
	median (Sued)	0.241	0.316	0.388	0.692	0.328	0.450	0.866
	median (TMLE)	0.237	0.280	0.359	0.600	0.295	0.401	0.696
	DP-DR ($\gamma = 0.1$)	0.183	0.301	0.563	2.262	0.317	0.649	2.395
	DP-DR ($\gamma = 0.5$)	0.202	0.290	0.326	0.692	0.274	0.349	0.839
	DP-DR ($\gamma = 1.0$)	0.239	0.287	0.307	0.530	0.287	0.335	0.669
	εDP-DR ($\gamma = 0.1$)	0.183	0.294	0.550	2.256	0.312	0.637	2.388
	εDP-DR ($\gamma = 0.5$)	0.202	0.263	0.301	0.659	0.269	0.321	0.800
	εDP-DR ($\gamma = 1.0$)	0.239	0.287	0.298	0.515	0.286	0.334	0.648
DR(T/F)	Naive	0.239	0.963	1.685	3.155	1.001	1.757	3.260
	median (Zhang-DR)	0.277	0.344	0.409	0.743	0.351	0.466	0.911
	median (Sued)	0.275	0.343	0.407	0.700	0.351	0.465	0.871
	median (TMLE)	0.242	0.285	0.364	0.624	0.297	0.406	0.725
	DP-DR ($\gamma = 0.1$)	0.240	0.315	0.563	2.267	0.331	0.645	2.391
	DP-DR ($\gamma = 0.5$)	0.251	0.322	0.353	0.720	0.321	0.365	0.840
	DP-DR ($\gamma = 1.0$)	0.284	0.337	0.349	0.588	0.332	0.380	0.709
	εDP-DR ($\gamma = 0.1$)	0.240	0.312	0.558	2.263	0.329	0.640	2.387
	εDP-DR ($\gamma = 0.5$)	0.250	0.318	0.345	0.696	0.315	0.358	0.812
	εDP-DR ($\gamma = 1.0$)	0.282	0.335	0.344	0.555	0.326	0.372	0.698
DR(F/T)	Naive	0.182	0.880	1.592	3.026	0.827	1.490	2.814
	median (Zhang-DR)	0.237	0.262	0.313	0.499	0.267	0.333	0.543
	median (Sued)	0.236	0.272	0.346	0.600	0.278	0.364	0.627
	median (TMLE)	0.235	0.259	0.306	0.470	0.264	0.324	0.513
	DP-DR ($\gamma = 0.1$)	0.183	0.193	0.346	2.063	0.192	0.302	1.687
	DP-DR ($\gamma = 0.5$)	0.200	0.209	0.221	0.365	0.205	0.211	0.282
	DP-DR ($\gamma = 1.0$)	0.230	0.234	0.242	0.278	0.231	0.234	0.244
	εDP-DR ($\gamma = 0.1$)	0.183	0.195	0.396	2.227	0.196	0.345	1.843
	εDP-DR ($\gamma = 0.5$)	0.199	0.204	0.208	0.401	0.204	0.211	0.323
	εDP-DR ($\gamma = 1.0$)	0.230	0.230	0.231	0.246	0.231	0.235	0.255

Table 4.10: Results of the comparative study. Each figure is RMSE between each method and the true value. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	3.004 (0.22)	3.749 (0.59)	4.493 (0.78)	5.983 (1.02)	3.766 (0.63)	4.536 (0.84)	6.070 (1.08)
	median (Firpo)	2.990 (0.26)	3.091 (0.28)	3.205 (0.30)	3.490 (0.43)	3.116 (0.28)	3.259 (0.32)	3.605 (0.47)
	median (Zhang-IPW)	2.990 (0.26)	3.091 (0.28)	3.205 (0.30)	3.490 (0.43)	3.116 (0.28)	3.259 (0.32)	3.605 (0.47)
	DP-IPW ($\gamma = 0.1$)	2.998 (0.22)	3.030 (0.27)	3.142 (0.51)	4.492 (1.70)	3.056 (0.29)	3.209 (0.57)	4.620 (1.74)
	DP-IPW ($\gamma = 0.5$)	2.986 (0.23)	2.987 (0.25)	2.989 (0.27)	3.052 (0.64)	3.011 (0.24)	3.042 (0.28)	3.173 (0.70)
	DP-IPW ($\gamma = 1.0$)	2.980 (0.26)	2.978 (0.27)	2.977 (0.27)	2.990 (0.41)	3.003 (0.26)	3.033 (0.28)	3.114 (0.48)
DR(T/T)	Naive	2.999 (0.18)	3.745 (0.60)	4.489 (0.79)	5.979 (1.04)	3.762 (0.64)	4.533 (0.86)	6.069 (1.11)
	median (Zhang-DR)	2.994 (0.24)	3.096 (0.30)	3.210 (0.33)	3.500 (0.54)	3.121 (0.31)	3.265 (0.37)	3.620 (0.66)
	median (Sued)	2.994 (0.24)	3.096 (0.30)	3.209 (0.33)	3.495 (0.48)	3.121 (0.31)	3.265 (0.36)	3.616 (0.61)
	median (TMLE)	2.995 (0.24)	3.094 (0.26)	3.208 (0.29)	3.476 (0.36)	3.120 (0.27)	3.260 (0.31)	3.583 (0.38)
	DP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.029 (0.30)	3.140 (0.55)	4.466 (1.72)	3.054 (0.31)	3.207 (0.62)	4.605 (1.78)
	DP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.998 (0.29)	3.000 (0.33)	3.060 (0.69)	3.022 (0.27)	3.053 (0.34)	3.197 (0.82)
	DP-DR ($\gamma = 1.0$)	2.993 (0.24)	2.991 (0.29)	2.992 (0.31)	3.010 (0.53)	3.017 (0.29)	3.047 (0.33)	3.137 (0.66)
	εDP-DR ($\gamma = 0.1$)	2.998 (0.18)	3.028 (0.29)	3.137 (0.53)	4.466 (1.72)	3.054 (0.31)	3.205 (0.60)	4.607 (1.77)
	εDP-DR ($\gamma = 0.5$)	2.996 (0.20)	2.997 (0.26)	2.999 (0.30)	3.057 (0.66)	3.022 (0.27)	3.052 (0.32)	3.192 (0.78)
	εDP-DR ($\gamma = 1.0$)	2.993 (0.24)	2.991 (0.29)	2.991 (0.30)	3.008 (0.52)	3.017 (0.29)	3.047 (0.33)	3.134 (0.63)
DR(T/F)	Naive	3.002 (0.24)	3.748 (0.61)	4.492 (0.78)	5.982 (1.03)	3.766 (0.64)	4.536 (0.85)	6.071 (1.09)
	median (Zhang-DR)	2.988 (0.28)	3.090 (0.33)	3.206 (0.35)	3.497 (0.55)	3.116 (0.33)	3.260 (0.39)	3.621 (0.67)
	median (Sued)	2.988 (0.28)	3.090 (0.33)	3.205 (0.35)	3.491 (0.50)	3.116 (0.33)	3.259 (0.39)	3.615 (0.62)
	median (TMLE)	3.000 (0.24)	3.101 (0.27)	3.216 (0.29)	3.499 (0.37)	3.126 (0.27)	3.268 (0.30)	3.610 (0.39)
	DP-DR ($\gamma = 0.1$)	2.996 (0.24)	3.032 (0.31)	3.149 (0.54)	4.490 (1.71)	3.058 (0.33)	3.217 (0.61)	4.621 (1.76)
	DP-DR ($\gamma = 0.5$)	2.984 (0.25)	2.987 (0.32)	2.990 (0.35)	3.058 (0.72)	3.012 (0.32)	3.042 (0.36)	3.194 (0.82)
	DP-DR ($\gamma = 1.0$)	2.977 (0.28)	2.976 (0.34)	2.976 (0.35)	2.999 (0.59)	3.002 (0.33)	3.033 (0.38)	3.130 (0.70)
	εDP-DR ($\gamma = 0.1$)	2.996 (0.24)	3.032 (0.31)	3.149 (0.54)	4.488 (1.71)	3.058 (0.32)	3.216 (0.60)	4.621 (1.75)
	εDP-DR ($\gamma = 0.5$)	2.984 (0.25)	2.987 (0.32)	2.990 (0.34)	3.056 (0.69)	3.011 (0.31)	3.042 (0.36)	3.190 (0.79)
	εDP-DR ($\gamma = 1.0$)	2.977 (0.28)	2.976 (0.33)	2.977 (0.34)	2.997 (0.55)	3.001 (0.33)	3.033 (0.37)	3.130 (0.69)
DR(F/T)	Naive	2.999 (0.18)	3.726 (0.50)	4.451 (0.65)	5.902 (0.86)	3.667 (0.49)	4.341 (0.65)	5.685 (0.84)
	median (Zhang-DR)	2.997 (0.24)	3.074 (0.25)	3.162 (0.27)	3.380 (0.32)	3.088 (0.25)	3.193 (0.27)	3.438 (0.32)
	median (Sued)	2.999 (0.24)	3.101 (0.25)	3.214 (0.27)	3.495 (0.34)	3.112 (0.25)	3.238 (0.28)	3.532 (0.33)
	median (TMLE)	2.997 (0.24)	3.072 (0.25)	3.157 (0.26)	3.358 (0.30)	3.086 (0.25)	3.186 (0.27)	3.414 (0.30)
	DP-DR ($\gamma = 0.1$)	2.998 (0.18)	2.996 (0.19)	3.051 (0.34)	4.333 (1.57)	3.016 (0.19)	3.077 (0.29)	4.007 (1.35)
	DP-DR ($\gamma = 0.5$)	2.995 (0.20)	2.970 (0.21)	2.944 (0.21)	2.901 (0.35)	2.994 (0.20)	2.995 (0.21)	3.003 (0.28)
	DP-DR ($\gamma = 1.0$)	2.996 (0.23)	2.969 (0.23)	2.943 (0.24)	2.885 (0.25)	2.994 (0.23)	2.996 (0.23)	2.999 (0.24)
	εDP-DR ($\gamma = 0.1$)	3.002 (0.18)	3.023 (0.19)	3.115 (0.38)	4.556 (1.59)	3.040 (0.19)	3.130 (0.32)	4.201 (1.40)
	εDP-DR ($\gamma = 0.5$)	2.999 (0.20)	2.996 (0.20)	2.997 (0.21)	3.024 (0.40)	3.017 (0.20)	3.040 (0.21)	3.104 (0.31)
	εDP-DR ($\gamma = 1.0$)	3.000 (0.23)	2.996 (0.23)	2.996 (0.23)	3.000 (0.25)	3.018 (0.23)	3.043 (0.23)	3.098 (0.24)

Table 4.11: Mean and SD of 10,000 simulated estimates of $\mu^{(1)}$. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.008	0.011	0.009	0.010	0.009	0.008	0.006
	median (Firpo)	2.028	2.038	2.009	1.992	1.591	1.604	1.610
	median (Zhang-IPW)	0.164	0.184	0.159	0.167	0.172	0.166	0.168
	DP-IPW ($\gamma = 0.1$)	33.735	44.282	60.503	79.176	39.488	54.135	70.099
	DP-IPW ($\gamma = 0.5$)	48.650	49.775	47.834	46.162	43.983	43.577	41.231
	DP-IPW ($\gamma = 1.0$)	65.725	64.715	61.072	53.993	56.336	54.092	48.227
DR(T/T)	Naive	0.028	0.013	0.015	0.015	0.014	0.017	0.015
	median (Zhang-DR)	0.412	0.442	0.429	0.435	0.359	0.371	0.370
	median (Sued)	1.736	1.806	1.758	1.836	1.579	1.624	1.605
	median (TMLE)	1055.657	1033.144	1008.369	993.124	943.045	934.975	921.709
	DP-DR ($\gamma = 0.1$)	18.983	57.100	136.473	211.645	52.968	125.900	186.668
	DP-DR ($\gamma = 0.5$)	56.355	48.725	48.968	39.064	43.681	43.085	36.285
	DP-DR ($\gamma = 1.0$)	87.300	82.622	74.404	62.842	73.313	69.041	55.489
	εDP-DR ($\gamma = 0.1$)	19.912	56.479	135.810	216.944	51.907	128.889	187.006
	εDP-DR ($\gamma = 0.5$)	57.311	52.397	50.270	39.846	44.903	46.369	35.150
	εDP-DR ($\gamma = 1.0$)	86.517	79.409	74.941	60.813	73.977	68.664	54.090
DR(T/F)	Naive	0.016	0.016	0.029	0.017	0.013	0.012	0.014
	median (Zhang-DR)	0.339	0.368	0.362	0.366	0.312	0.312	0.319
	median (Sued)	1.689	1.873	3.724	3.779	1.445	1.501	1.683
	median (TMLE)	981.209	994.937	968.273	974.131	887.897	895.797	915.279
	DP-DR ($\gamma = 0.1$)	18.828	59.230	132.483	197.779	53.452	124.888	179.470
	DP-DR ($\gamma = 0.5$)	55.630	51.825	46.049	36.762	48.728	44.227	36.664
	DP-DR ($\gamma = 1.0$)	81.111	80.063	71.932	58.312	70.519	67.554	54.077
	εDP-DR ($\gamma = 0.1$)	18.391	60.712	130.125	199.546	54.833	126.829	180.125
	εDP-DR ($\gamma = 0.5$)	56.661	52.369	47.067	40.297	49.611	42.699	37.208
	εDP-DR ($\gamma = 1.0$)	83.076	80.920	73.754	62.673	71.937	67.024	56.490
DR(F/T)	Naive	0.013	0.016	0.014	0.015	0.014	0.016	0.014
	median (Zhang-DR)	0.404	2.159	0.415	0.423	0.372	0.363	0.369
	median (Sued)	1.652	1.849	1.657	1.781	1.505	1.487	1.564
	median (TMLE)	940.593	961.615	929.511	967.157	861.292	867.211	893.940
	DP-DR ($\gamma = 0.1$)	19.768	63.865	144.046	222.438	50.837	116.164	225.446
	DP-DR ($\gamma = 0.5$)	51.923	49.406	49.673	42.449	45.033	45.493	36.107
	DP-DR ($\gamma = 1.0$)	90.062	85.686	74.602	64.006	76.707	72.257	57.556
	εDP-DR ($\gamma = 0.1$)	19.023	63.164	147.321	202.361	50.513	115.868	210.642
	εDP-DR ($\gamma = 0.5$)	57.110	50.739	46.918	42.489	47.041	43.003	36.828
	εDP-DR ($\gamma = 1.0$)	87.439	87.305	75.546	62.424	76.284	69.237	58.958

Table 4.12: Mean computation time (ms) of 10,000 simulations. The covariates X were generated from Gaussian distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.197	0.952	1.711	3.174	1.000	1.762	3.269
	median (Firpo)	0.278	0.319	0.407	0.705	0.334	0.447	0.785
	median (Zhang-IPW)	0.278	0.319	0.407	0.705	0.334	0.447	0.785
	DP-IPW ($\gamma = 0.1$)	0.199	0.230	0.567	2.329	0.253	0.609	2.429
	DP-IPW ($\gamma = 0.5$)	0.223	0.230	0.239	0.703	0.230	0.247	0.690
	DP-IPW ($\gamma = 1.0$)	0.273	0.273	0.273	0.432	0.272	0.279	0.394
DR(T/T)	Naive	0.182	0.948	1.711	3.176	1.000	1.766	3.275
	median (Zhang-DR)	0.265	0.308	0.404	0.732	0.327	0.446	0.800
	median (Sued)	0.264	0.308	0.402	0.704	0.327	0.445	0.786
	median (TMLE)	0.263	0.308	0.401	0.664	0.326	0.442	0.768
	DP-DR ($\gamma = 0.1$)	0.184	0.219	0.582	2.332	0.254	0.624	2.438
	DP-DR ($\gamma = 0.5$)	0.208	0.216	0.243	0.745	0.218	0.246	0.726
	DP-DR ($\gamma = 1.0$)	0.257	0.258	0.260	0.498	0.259	0.272	0.480
	εDP-DR ($\gamma = 0.1$)	0.184	0.218	0.579	2.330	0.251	0.618	2.435
	εDP-DR ($\gamma = 0.5$)	0.208	0.216	0.241	0.730	0.218	0.245	0.710
	εDP-DR ($\gamma = 1.0$)	0.257	0.258	0.260	0.481	0.260	0.266	0.437
DR(T/F)	Naive	0.202	0.954	1.715	3.179	1.004	1.767	3.276
	median (Zhang-DR)	0.285	0.324	0.413	0.738	0.340	0.453	0.806
	median (Sued)	0.285	0.325	0.412	0.710	0.340	0.453	0.792
	median (TMLE)	0.267	0.313	0.406	0.682	0.331	0.448	0.786
	DP-DR ($\gamma = 0.1$)	0.204	0.235	0.579	2.332	0.261	0.625	2.436
	DP-DR ($\gamma = 0.5$)	0.229	0.235	0.257	0.752	0.236	0.268	0.754
	DP-DR ($\gamma = 1.0$)	0.279	0.279	0.280	0.514	0.279	0.287	0.487
	εDP-DR ($\gamma = 0.1$)	0.204	0.234	0.577	2.329	0.262	0.623	2.433
	εDP-DR ($\gamma = 0.5$)	0.229	0.235	0.256	0.742	0.235	0.266	0.743
	εDP-DR ($\gamma = 1.0$)	0.279	0.279	0.279	0.500	0.278	0.286	0.476
DR(F/T)	Naive	0.180	0.887	1.611	3.037	0.832	1.488	2.805
	median (Zhang-DR)	0.259	0.288	0.349	0.544	0.295	0.365	0.591
	median (Sued)	0.259	0.299	0.387	0.658	0.306	0.399	0.683
	median (TMLE)	0.256	0.282	0.338	0.508	0.289	0.354	0.554
	DP-DR ($\gamma = 0.1$)	0.183	0.195	0.381	2.121	0.196	0.315	1.734
	DP-DR ($\gamma = 0.5$)	0.206	0.216	0.229	0.467	0.213	0.219	0.292
	DP-DR ($\gamma = 1.0$)	0.247	0.251	0.259	0.301	0.249	0.251	0.258
	εDP-DR ($\gamma = 0.1$)	0.183	0.198	0.426	2.273	0.202	0.358	1.888
	εDP-DR ($\gamma = 0.5$)	0.206	0.212	0.217	0.491	0.212	0.220	0.318
	εDP-DR ($\gamma = 1.0$)	0.247	0.247	0.247	0.264	0.248	0.253	0.278

Table 4.13: Results of the comparative study. Each figure is RMSE between each method and the true value. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the Gaussian MLE using non-outliers. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	3.003 (0.20)	3.759 (0.57)	4.521 (0.79)	6.004 (1.02)	3.783 (0.62)	4.556 (0.83)	6.090 (1.07)
	median (Firpo)	2.985 (0.28)	3.099 (0.30)	3.233 (0.33)	3.541 (0.45)	3.130 (0.31)	3.292 (0.34)	3.666 (0.42)
	median (Zhang-IPW)	2.985 (0.28)	3.099 (0.30)	3.233 (0.33)	3.541 (0.45)	3.130 (0.31)	3.292 (0.34)	3.666 (0.42)
	DP-IPW ($\gamma = 0.1$)	3.000 (0.20)	3.038 (0.23)	3.186 (0.54)	4.590 (1.70)	3.066 (0.24)	3.241 (0.56)	4.713 (1.72)
	DP-IPW ($\gamma = 0.5$)	2.991 (0.22)	2.990 (0.23)	2.994 (0.24)	3.075 (0.70)	3.017 (0.23)	3.048 (0.24)	3.185 (0.66)
	DP-IPW ($\gamma = 1.0$)	2.979 (0.27)	2.978 (0.27)	2.982 (0.27)	2.998 (0.43)	3.009 (0.27)	3.043 (0.28)	3.123 (0.37)
DR(T/T)	Naive	3.001 (0.18)	3.757 (0.57)	4.519 (0.79)	6.003 (1.03)	3.781 (0.62)	4.555 (0.84)	6.090 (1.08)
	median (Zhang-DR)	2.990 (0.27)	3.104 (0.29)	3.237 (0.33)	3.547 (0.49)	3.135 (0.30)	3.295 (0.33)	3.669 (0.44)
	median (Sued)	2.990 (0.26)	3.103 (0.29)	3.237 (0.32)	3.544 (0.45)	3.135 (0.30)	3.295 (0.33)	3.668 (0.41)
	median (TMLE)	2.991 (0.26)	3.104 (0.29)	3.236 (0.32)	3.530 (0.40)	3.134 (0.30)	3.292 (0.33)	3.652 (0.41)
	DP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.035 (0.22)	3.185 (0.55)	4.582 (1.71)	3.064 (0.25)	3.238 (0.58)	4.704 (1.74)
	DP-DR ($\gamma = 0.5$)	2.995 (0.21)	2.992 (0.22)	2.997 (0.24)	3.083 (0.74)	3.020 (0.22)	3.051 (0.24)	3.190 (0.70)
	DP-DR ($\gamma = 1.0$)	2.988 (0.26)	2.985 (0.26)	2.988 (0.26)	3.008 (0.50)	3.016 (0.26)	3.049 (0.27)	3.136 (0.46)
	εDP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.035 (0.22)	3.184 (0.55)	4.582 (1.71)	3.064 (0.24)	3.237 (0.57)	4.706 (1.74)
	εDP-DR ($\gamma = 0.5$)	2.995 (0.21)	2.992 (0.22)	2.996 (0.24)	3.081 (0.73)	3.020 (0.22)	3.051 (0.24)	3.188 (0.68)
	εDP-DR ($\gamma = 1.0$)	2.988 (0.26)	2.985 (0.26)	2.988 (0.26)	3.005 (0.48)	3.015 (0.26)	3.049 (0.26)	3.132 (0.42)
DR(T/F)	Naive	3.005 (0.20)	3.762 (0.57)	4.524 (0.79)	6.008 (1.03)	3.786 (0.62)	4.560 (0.83)	6.095 (1.07)
	median (Zhang-DR)	2.987 (0.28)	3.103 (0.31)	3.237 (0.34)	3.550 (0.49)	3.134 (0.31)	3.296 (0.34)	3.673 (0.44)
	median (Sued)	2.987 (0.28)	3.102 (0.31)	3.236 (0.34)	3.544 (0.46)	3.133 (0.31)	3.295 (0.34)	3.670 (0.42)
	median (TMLE)	2.996 (0.27)	3.110 (0.29)	3.243 (0.33)	3.547 (0.41)	3.140 (0.30)	3.300 (0.33)	3.671 (0.41)
	DP-DR ($\gamma = 0.1$)	3.003 (0.20)	3.042 (0.23)	3.192 (0.55)	4.590 (1.71)	3.070 (0.25)	3.249 (0.57)	4.714 (1.73)
	DP-DR ($\gamma = 0.5$)	2.994 (0.23)	2.993 (0.23)	2.997 (0.26)	3.090 (0.75)	3.020 (0.23)	3.053 (0.26)	3.204 (0.73)
	DP-DR ($\gamma = 1.0$)	2.982 (0.28)	2.981 (0.28)	2.985 (0.28)	3.008 (0.51)	3.011 (0.28)	3.046 (0.28)	3.136 (0.47)
	εDP-DR ($\gamma = 0.1$)	3.003 (0.20)	3.042 (0.23)	3.191 (0.54)	4.587 (1.70)	3.070 (0.25)	3.248 (0.57)	4.713 (1.73)
	εDP-DR ($\gamma = 0.5$)	2.994 (0.23)	2.992 (0.23)	2.997 (0.26)	3.087 (0.74)	3.020 (0.23)	3.052 (0.26)	3.201 (0.72)
	εDP-DR ($\gamma = 1.0$)	2.982 (0.28)	2.980 (0.28)	2.984 (0.28)	3.006 (0.50)	3.011 (0.28)	3.046 (0.28)	3.134 (0.46)
DR(F/T)	Naive	3.000 (0.18)	3.737 (0.49)	4.469 (0.66)	5.911 (0.87)	3.676 (0.49)	4.344 (0.64)	5.676 (0.84)
	median (Zhang-DR)	3.005 (0.26)	3.088 (0.27)	3.187 (0.29)	3.417 (0.35)	3.104 (0.28)	3.217 (0.29)	3.482 (0.34)
	median (Sued)	2.999 (0.26)	3.113 (0.28)	3.243 (0.30)	3.546 (0.37)	3.125 (0.28)	3.264 (0.30)	3.587 (0.35)
	median (TMLE)	3.004 (0.26)	3.085 (0.27)	3.180 (0.29)	3.391 (0.32)	3.101 (0.27)	3.210 (0.29)	3.454 (0.32)
	DP-DR ($\gamma = 0.1$)	3.000 (0.18)	3.007 (0.20)	3.088 (0.37)	4.422 (1.57)	3.027 (0.19)	3.104 (0.30)	4.082 (1.36)
	DP-DR ($\gamma = 0.5$)	3.000 (0.21)	2.970 (0.21)	2.943 (0.22)	2.907 (0.46)	2.997 (0.21)	2.996 (0.22)	2.999 (0.29)
	DP-DR ($\gamma = 1.0$)	3.003 (0.25)	2.970 (0.25)	2.939 (0.25)	2.868 (0.27)	2.999 (0.25)	2.998 (0.25)	2.994 (0.26)
	εDP-DR ($\gamma = 0.1$)	3.000 (0.18)	3.032 (0.20)	3.150 (0.40)	4.641 (1.57)	3.052 (0.20)	3.162 (0.32)	4.286 (1.38)
	εDP-DR ($\gamma = 0.5$)	3.000 (0.21)	2.997 (0.21)	2.999 (0.22)	3.036 (0.49)	3.024 (0.21)	3.052 (0.21)	3.121 (0.29)
	εDP-DR ($\gamma = 1.0$)	3.003 (0.25)	2.999 (0.25)	3.001 (0.25)	3.000 (0.26)	3.029 (0.25)	3.060 (0.25)	3.125 (0.25)

Table 4.14: Mean and SD of 10,000 simulated estimates of $\mu^{(1)}$. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the Gaussian MLE using non-outliers. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.006	0.008	0.008	0.010	0.009	0.008	0.011
	median (Firpo)	1.624	1.554	1.598	1.596	1.640	1.599	1.595
	median (Zhang-IPW)	0.144	0.155	0.146	0.149	0.163	0.166	0.166
	DP-IPW ($\gamma = 0.1$)	28.899	40.705	57.891	72.877	41.438	58.884	72.039
	DP-IPW ($\gamma = 0.5$)	41.124	43.104	41.639	40.541	43.319	42.599	40.527
	DP-IPW ($\gamma = 1.0$)	58.435	57.102	56.009	48.982	57.823	54.624	48.146
DR(T/T)	Naive	0.013	0.017	0.014	0.014	0.011	0.013	0.012
	median (Zhang-DR)	0.354	0.380	0.429	0.452	0.391	0.387	0.388
	median (Sued)	1.470	1.427	1.650	1.733	1.660	1.535	1.645
	median (TMLE)	1019.177	987.574	1046.361	1019.792	1012.613	973.752	953.734
	DP-DR ($\gamma = 0.1$)	14.030	65.830	156.014	221.084	65.235	150.543	206.717
	DP-DR ($\gamma = 0.5$)	42.618	39.445	42.180	33.461	39.311	39.022	30.945
	DP-DR ($\gamma = 1.0$)	80.107	73.122	75.633	54.222	74.230	66.690	52.706
	εDP-DR ($\gamma = 0.1$)	13.830	61.746	159.505	222.713	64.413	146.996	203.859
	εDP-DR ($\gamma = 0.5$)	42.832	43.615	42.286	36.057	43.612	37.242	31.352
	εDP-DR ($\gamma = 1.0$)	80.584	75.842	71.612	55.196	77.477	64.020	52.287
DR(T/F)	Naive	0.011	0.013	0.018	0.014	0.013	0.013	0.018
	median (Zhang-DR)	0.298	2.320	0.365	0.371	0.316	0.320	0.334
	median (Sued)	1.497	4.069	1.717	1.870	1.548	1.509	1.679
	median (TMLE)	908.448	921.784	963.501	973.878	904.782	896.085	914.020
	DP-DR ($\gamma = 0.1$)	12.200	61.403	166.398	223.529	62.714	154.528	201.548
	DP-DR ($\gamma = 0.5$)	42.066	38.529	41.758	32.706	40.717	34.629	31.154
	DP-DR ($\gamma = 1.0$)	76.853	73.440	75.127	59.673	74.651	63.882	52.368
	εDP-DR ($\gamma = 0.1$)	12.005	64.193	164.242	216.765	61.661	152.694	206.549
	εDP-DR ($\gamma = 0.5$)	42.311	40.010	39.141	34.312	40.385	37.031	31.584
	εDP-DR ($\gamma = 1.0$)	77.362	74.394	73.187	54.911	75.535	67.096	51.224
DR(F/T)	Naive	0.014	0.014	0.017	0.017	0.012	0.013	0.011
	median (Zhang-DR)	0.368	0.431	0.412	0.431	0.390	0.379	2.279
	median (Sued)	1.586	3.853	1.855	1.805	1.589	1.626	4.012
	median (TMLE)	896.351	970.473	976.782	987.069	909.770	907.980	930.365
	DP-DR ($\gamma = 0.1$)	13.761	74.393	177.183	239.704	60.244	143.986	246.342
	DP-DR ($\gamma = 0.5$)	44.319	46.824	40.203	37.486	42.798	40.716	34.311
	DP-DR ($\gamma = 1.0$)	85.614	85.866	75.166	60.665	80.635	71.475	59.946
	εDP-DR ($\gamma = 0.1$)	13.569	72.655	175.524	216.075	58.960	144.481	227.731
	εDP-DR ($\gamma = 0.5$)	44.594	50.019	45.107	37.493	41.513	42.198	33.163
	εDP-DR ($\gamma = 1.0$)	85.994	84.480	77.364	62.210	80.407	74.459	57.817

Table 4.15: Mean computation time (ms) of 10,000 simulations. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the Gaussian MLE over non-outliers. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.197	0.952	1.711	3.174	1.000	1.762	3.269
	median (Firpo)	0.278	0.319	0.407	0.705	0.334	0.447	0.785
	median (Zhang-IPW)	0.278	0.319	0.407	0.705	0.334	0.447	0.785
	DP-IPW ($\gamma = 0.1$)	0.199	0.230	0.567	2.329	0.253	0.609	2.429
	DP-IPW ($\gamma = 0.5$)	0.223	0.230	0.239	0.703	0.230	0.247	0.690
	DP-IPW ($\gamma = 1.0$)	0.273	0.273	0.273	0.432	0.272	0.279	0.394
DR(T/T)	Naive	0.182	0.948	1.711	3.176	1.000	1.766	3.275
	median (Zhang-DR)	0.265	0.309	0.404	0.732	0.328	0.446	0.799
	median (Sued)	0.264	0.308	0.403	0.705	0.327	0.445	0.785
	median (TMLE)	0.263	0.307	0.402	0.659	0.327	0.443	0.761
	DP-DR ($\gamma = 0.1$)	0.184	0.219	0.583	2.333	0.254	0.623	2.439
	DP-DR ($\gamma = 0.5$)	0.208	0.216	0.243	0.746	0.218	0.246	0.726
	DP-DR ($\gamma = 1.0$)	0.257	0.258	0.260	0.504	0.259	0.272	0.480
	εDP-DR ($\gamma = 0.1$)	0.184	0.218	0.578	2.329	0.251	0.618	2.434
	εDP-DR ($\gamma = 0.5$)	0.208	0.216	0.241	0.726	0.218	0.245	0.708
	εDP-DR ($\gamma = 1.0$)	0.257	0.258	0.260	0.482	0.260	0.266	0.434
DR(T/F)	Naive	0.203	0.954	1.714	3.178	1.003	1.767	3.275
	median (Zhang-DR)	0.285	0.325	0.413	0.725	0.341	0.453	0.806
	median (Sued)	0.284	0.324	0.412	0.710	0.340	0.453	0.792
	median (TMLE)	0.266	0.313	0.408	0.697	0.331	0.449	0.787
	DP-DR ($\gamma = 0.1$)	0.205	0.236	0.580	2.330	0.262	0.625	2.437
	DP-DR ($\gamma = 0.5$)	0.230	0.236	0.258	0.750	0.237	0.269	0.755
	DP-DR ($\gamma = 1.0$)	0.280	0.280	0.281	0.501	0.279	0.288	0.489
	εDP-DR ($\gamma = 0.1$)	0.205	0.235	0.577	2.328	0.261	0.622	2.434
	εDP-DR ($\gamma = 0.5$)	0.230	0.236	0.256	0.738	0.236	0.267	0.743
	εDP-DR ($\gamma = 1.0$)	0.280	0.279	0.280	0.488	0.279	0.287	0.475
DR(F/T)	Naive	0.182	0.887	1.611	3.037	0.832	1.488	2.805
	median (Zhang-DR)	0.260	0.285	0.344	0.535	0.292	0.359	0.582
	median (Sued)	0.259	0.300	0.388	0.658	0.306	0.399	0.683
	median (TMLE)	0.258	0.280	0.333	0.498	0.287	0.348	0.543
	DP-DR ($\gamma = 0.1$)	0.185	0.197	0.385	2.128	0.197	0.319	1.741
	DP-DR ($\gamma = 0.5$)	0.208	0.219	0.233	0.474	0.215	0.221	0.290
	DP-DR ($\gamma = 1.0$)	0.249	0.254	0.265	0.304	0.251	0.253	0.261
	εDP-DR ($\gamma = 0.1$)	0.185	0.202	0.449	2.294	0.204	0.368	1.897
	εDP-DR ($\gamma = 0.5$)	0.207	0.212	0.217	0.516	0.213	0.220	0.320
	εDP-DR ($\gamma = 1.0$)	0.248	0.247	0.247	0.267	0.248	0.251	0.272

Table 4.16: Results of the comparative study. Each figure is RMSE between each method and the true value. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	3.003 (0.20)	3.759 (0.57)	4.521 (0.79)	6.004 (1.02)	3.783 (0.62)	4.556 (0.83)	6.090 (1.07)
	median (Firpo)	2.985 (0.28)	3.099 (0.30)	3.233 (0.33)	3.541 (0.45)	3.130 (0.31)	3.292 (0.34)	3.666 (0.42)
	median (Zhang-IPW)	2.985 (0.28)	3.099 (0.30)	3.233 (0.33)	3.541 (0.45)	3.130 (0.31)	3.292 (0.34)	3.666 (0.42)
	DP-IPW ($\gamma = 0.1$)	3.000 (0.20)	3.038 (0.23)	3.186 (0.54)	4.590 (1.70)	3.066 (0.24)	3.241 (0.56)	4.713 (1.72)
	DP-IPW ($\gamma = 0.5$)	2.991 (0.22)	2.990 (0.23)	2.994 (0.24)	3.075 (0.70)	3.017 (0.23)	3.048 (0.24)	3.185 (0.66)
	DP-IPW ($\gamma = 1.0$)	2.979 (0.27)	2.978 (0.27)	2.982 (0.27)	2.998 (0.43)	3.009 (0.27)	3.043 (0.28)	3.123 (0.37)
DR(T/T)	Naive	3.001 (0.18)	3.757 (0.57)	4.519 (0.79)	6.003 (1.03)	3.781 (0.62)	4.555 (0.84)	6.090 (1.08)
	median (Zhang-DR)	2.990 (0.27)	3.104 (0.29)	3.238 (0.33)	3.548 (0.49)	3.135 (0.30)	3.295 (0.33)	3.668 (0.44)
	median (Sued)	2.990 (0.26)	3.104 (0.29)	3.237 (0.33)	3.544 (0.45)	3.135 (0.30)	3.295 (0.33)	3.667 (0.41)
	median (TMLE)	2.991 (0.26)	3.104 (0.29)	3.236 (0.33)	3.527 (0.40)	3.134 (0.30)	3.292 (0.33)	3.647 (0.40)
	DP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.035 (0.22)	3.185 (0.55)	4.582 (1.71)	3.064 (0.25)	3.238 (0.58)	4.706 (1.74)
	DP-DR ($\gamma = 0.5$)	2.995 (0.21)	2.992 (0.22)	2.996 (0.24)	3.083 (0.74)	3.020 (0.22)	3.051 (0.24)	3.190 (0.70)
	DP-DR ($\gamma = 1.0$)	2.988 (0.26)	2.985 (0.26)	2.988 (0.26)	3.009 (0.50)	3.016 (0.26)	3.049 (0.27)	3.136 (0.46)
	εDP-DR ($\gamma = 0.1$)	2.999 (0.18)	3.035 (0.22)	3.184 (0.55)	4.581 (1.71)	3.064 (0.24)	3.238 (0.57)	4.705 (1.74)
	εDP-DR ($\gamma = 0.5$)	2.995 (0.21)	2.992 (0.22)	2.996 (0.24)	3.080 (0.72)	3.020 (0.22)	3.051 (0.24)	3.187 (0.68)
	εDP-DR ($\gamma = 1.0$)	2.988 (0.26)	2.985 (0.26)	2.988 (0.26)	3.006 (0.48)	3.016 (0.26)	3.049 (0.26)	3.132 (0.41)
DR(T/F)	Naive	3.005 (0.20)	3.761 (0.57)	4.523 (0.79)	6.007 (1.03)	3.785 (0.62)	4.559 (0.83)	6.094 (1.07)
	median (Zhang-DR)	2.986 (0.29)	3.101 (0.31)	3.237 (0.34)	3.547 (0.48)	3.133 (0.31)	3.295 (0.34)	3.672 (0.44)
	median (Sued)	2.987 (0.28)	3.101 (0.31)	3.236 (0.34)	3.544 (0.46)	3.133 (0.31)	3.295 (0.34)	3.670 (0.42)
	median (TMLE)	2.997 (0.27)	3.111 (0.29)	3.245 (0.33)	3.550 (0.43)	3.141 (0.30)	3.302 (0.33)	3.672 (0.41)
	DP-DR ($\gamma = 0.1$)	3.002 (0.21)	3.041 (0.23)	3.192 (0.55)	4.588 (1.71)	3.070 (0.25)	3.248 (0.57)	4.715 (1.73)
	DP-DR ($\gamma = 0.5$)	2.993 (0.23)	2.992 (0.24)	2.997 (0.26)	3.088 (0.74)	3.019 (0.24)	3.052 (0.26)	3.202 (0.73)
	DP-DR ($\gamma = 1.0$)	2.981 (0.28)	2.980 (0.28)	2.984 (0.28)	3.007 (0.50)	3.010 (0.28)	3.045 (0.28)	3.134 (0.47)
	εDP-DR ($\gamma = 0.1$)	3.002 (0.20)	3.041 (0.23)	3.191 (0.54)	4.588 (1.70)	3.069 (0.25)	3.247 (0.57)	4.714 (1.73)
	εDP-DR ($\gamma = 0.5$)	2.993 (0.23)	2.992 (0.24)	2.997 (0.26)	3.085 (0.73)	3.019 (0.24)	3.052 (0.26)	3.200 (0.72)
	εDP-DR ($\gamma = 1.0$)	2.981 (0.28)	2.980 (0.28)	2.983 (0.28)	3.005 (0.49)	3.010 (0.28)	3.045 (0.28)	3.133 (0.46)
DR(F/T)	Naive	3.000 (0.18)	3.737 (0.49)	4.469 (0.66)	5.911 (0.87)	3.676 (0.49)	4.344 (0.64)	5.676 (0.84)
	median (Zhang-DR)	2.997 (0.26)	3.079 (0.27)	3.177 (0.29)	3.405 (0.35)	3.095 (0.28)	3.207 (0.29)	3.471 (0.34)
	median (Sued)	3.000 (0.26)	3.113 (0.28)	3.244 (0.30)	3.547 (0.37)	3.125 (0.28)	3.265 (0.30)	3.587 (0.35)
	median (TMLE)	2.996 (0.26)	3.076 (0.27)	3.170 (0.29)	3.378 (0.32)	3.092 (0.27)	3.200 (0.29)	3.440 (0.32)
	DP-DR ($\gamma = 0.1$)	2.999 (0.19)	3.006 (0.20)	3.088 (0.38)	4.435 (1.57)	3.026 (0.20)	3.104 (0.30)	4.092 (1.36)
	DP-DR ($\gamma = 0.5$)	2.993 (0.21)	2.964 (0.22)	2.936 (0.22)	2.901 (0.46)	2.990 (0.21)	2.990 (0.22)	2.993 (0.29)
	DP-DR ($\gamma = 1.0$)	2.991 (0.25)	2.957 (0.25)	2.927 (0.25)	2.856 (0.27)	2.987 (0.25)	2.986 (0.25)	2.982 (0.26)
	εDP-DR ($\gamma = 0.1$)	3.003 (0.18)	3.035 (0.20)	3.160 (0.42)	4.658 (1.58)	3.051 (0.20)	3.160 (0.33)	4.284 (1.40)
	εDP-DR ($\gamma = 0.5$)	2.998 (0.21)	2.995 (0.21)	2.997 (0.22)	3.041 (0.51)	3.017 (0.21)	3.042 (0.22)	3.105 (0.30)
	εDP-DR ($\gamma = 1.0$)	2.996 (0.25)	2.992 (0.25)	2.994 (0.25)	2.996 (0.27)	3.017 (0.25)	3.044 (0.25)	3.102 (0.25)

Table 4.17: Mean and SD of 10,000 simulated estimates of $\mu^{(1)}$. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

		No contam.	Homogeneous			Heterogeneous		
			$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.20$
IPW(T/-)	Naive	0.012	0.009	0.011	0.008	0.009	0.008	0.008
	median (Firpo)	2.061	2.018	2.001	2.014	1.593	1.588	1.591
	median (Zhang-IPW)	0.159	0.158	0.155	0.174	0.160	0.162	0.162
	DP-IPW ($\gamma = 0.1$)	32.502	46.715	64.496	80.991	41.595	58.327	72.898
	DP-IPW ($\gamma = 0.5$)	45.786	48.372	46.799	45.370	43.191	41.859	40.298
	DP-IPW ($\gamma = 1.0$)	67.400	64.207	61.061	54.756	57.049	54.186	47.922
DR(T/T)	Naive	0.020	0.017	0.015	0.014	0.015	0.014	0.016
	median (Zhang-DR)	0.426	0.435	0.413	0.432	0.371	0.372	0.372
	median (Sued)	1.769	1.883	1.684	1.763	1.524	1.521	1.567
	median (TMLE)	1120.595	1076.392	1039.786	1029.243	991.954	961.859	947.842
	DP-DR ($\gamma = 0.1$)	15.718	71.203	165.047	228.545	66.563	152.820	207.025
	DP-DR ($\gamma = 0.5$)	50.227	46.273	41.325	36.313	43.227	37.408	29.929
	DP-DR ($\gamma = 1.0$)	92.921	85.411	73.518	56.090	74.702	65.099	53.264
	εDP-DR ($\gamma = 0.1$)	15.929	72.355	166.803	229.476	66.591	151.084	203.590
	εDP-DR ($\gamma = 0.5$)	50.950	46.989	43.582	33.812	45.355	37.982	32.001
	εDP-DR ($\gamma = 1.0$)	93.544	82.191	72.519	58.983	74.507	67.979	54.172
DR(T/F)	Naive	0.015	0.014	0.014	0.017	0.012	0.013	0.012
	median (Zhang-DR)	0.344	0.369	0.379	0.348	0.310	0.316	0.313
	median (Sued)	1.674	1.913	1.710	1.805	1.645	1.543	1.579
	median (TMLE)	1004.081	999.738	972.337	971.670	918.025	905.992	907.104
	DP-DR ($\gamma = 0.1$)	13.940	71.371	157.230	220.831	63.078	148.782	198.135
	DP-DR ($\gamma = 0.5$)	46.461	45.985	38.650	34.250	40.784	36.915	30.849
	DP-DR ($\gamma = 1.0$)	82.531	80.473	72.655	57.973	74.487	64.093	50.140
	εDP-DR ($\gamma = 0.1$)	13.889	68.556	163.008	221.689	63.626	148.044	204.685
	εDP-DR ($\gamma = 0.5$)	46.726	44.754	39.936	35.078	40.930	37.054	29.266
	εDP-DR ($\gamma = 1.0$)	83.223	81.341	72.481	56.218	73.313	64.716	49.878
DR(F/T)	Naive	0.013	0.014	0.014	0.018	0.012	0.017	0.012
	median (Zhang-DR)	0.411	0.429	0.435	0.435	0.383	0.366	0.374
	median (Sued)	1.719	1.848	1.776	1.807	1.503	1.684	1.580
	median (TMLE)	974.546	977.795	969.097	988.810	892.363	913.748	924.284
	DP-DR ($\gamma = 0.1$)	16.894	76.806	185.199	241.377	61.562	142.078	254.437
	DP-DR ($\gamma = 0.5$)	56.429	50.977	44.908	40.217	46.906	41.295	34.944
	DP-DR ($\gamma = 1.0$)	92.542	89.220	78.322	64.042	80.361	74.055	58.842
	εDP-DR ($\gamma = 0.1$)	17.194	76.893	177.770	220.029	61.670	147.734	233.473
	εDP-DR ($\gamma = 0.5$)	55.339	51.744	45.734	37.134	46.247	41.614	32.412
	εDP-DR ($\gamma = 1.0$)	93.373	88.307	80.103	64.488	80.508	74.952	61.334

Table 4.18: Mean computation time (ms) of 10,000 simulations. The covariates X were generated from uniform distributions, and the outcome regression was obtained by the unnormalized Gaussian modeling. The characters "T" and "F" denote the correct and the incorrect modeling for PS/OR.

5 Conclusion

In this thesis, we discuss statistical inference about causal relationships. As mentioned in the introduction, causality is difficult to express in the usual statistical framework, and therefore, in order to make inferences implying causality, we need to use a framework with more expressive power. High expressiveness, in turn, is related to the difficulty of estimation. For example, LiNGAM, discussed in Chapter 3, requires constraints that are not usually required for asymptotic consistency. Besides, the semiparametric estimator of the ATE discussed in Chapter 4 requires us to build nuisance models using covariates to estimate the mean without bias, even though we just want to know the mean. In addition, since causality is a matter of great practical interest, it is essential to be able to deal with the various difficulties that appear in real data. In this thesis, we discuss LiNGAM for high-dimensional and sparse data, and IPW/DR estimators under outlier contamination. The results suggest that in order to estimate causal models under data difficulties, one should not only deal with causal difficulties and data difficulties separately, but also pay attention to the difficulties arising from the combination of both. In Chapter 3, it has been found that in order to combine the uncorrelatedness of the independent components, which is the prerequisite for consistency, with the sparsity of the recovering matrix, it is necessary to incorporate both the sparse penalty and whitening of the data matrix. This has encouraged us to use the generalized lasso type penalty and the orthogonal penalty. In Chapter 4, it has been found that the DR-M estimator should be corrected with the proportion of outliers under contamination because the DR-M estimator loses its double robustness when only the conditional model is correct. The models and estimators discussed in this thesis are relatively basic in statistical causal inference, and it goes without saying that there are various difficulties with the real data other than those dealt with in this thesis. For more advanced causal questions, more complex estimands may be required, and in such cases, we need to pay more attentions about data difficulties.

References

[1] Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[3] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Series B Stat. Methodol.*, 80(4):597–623, September 2018.

[4] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[5] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[7] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[8] A Belloni, V Chernozhukov, I Fernández-Val, and C Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[9] A Belloni, V Chernozhukov, and C Hansen. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650, April 2014.

[10] Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[11] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2016.

[12] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Now Publishers Inc, 2011.

[13] Rainer E Burkard and Eranda Cela. Linear assignment problems and extensions. In *Handbook of combinatorial optimization*, pages 75–149. Springer, 1999.

[14] Zhitang Chen and Laiwan Chan. Causality in linear nongaussian acyclic models in the presence of latent gaussian confounders. *Neural Computation*, 25(6):1605–1641, 2013.

[15] Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

[16] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[17] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[18] Iván Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

[19] Thad Dunning. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, September 2012.

[20] P Erdös and A Rényi. On random graphs i. *Publ. Math. Debrecen*, 6(290-297):18, 1959.

[21] Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007.

[22] R.A. Fisher. *The design of experiments. 1935*. Oliver and Boyd, Edinburgh, 1935.

[23] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, March 2002.

[24] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[25] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[26] Hironori Fujisawa et al. Normalized estimating equation for robust parameter estimation. *Electronic Journal of Statistics*, 7:1587–1606, 2013.

[27] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[28] David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Found. Sci.*, 3(1):151–182, January 1998.

[29] J Y Halpern. Axiomatizing causal reasoning, 2000.

[30] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[31] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC press, 2015.

[32] Miguel A Hernán and James M Robins. *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC, 2020.

[33] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[34] Takahiro Hoshino. Doubly robust-type estimation for covariate adjustment in latent variable modeling. *Psychometrika*, 72(4):535–549, 2007.

[35] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[36] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

[37] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[38] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[39] Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

[40] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

[41] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[42] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[43] MC Jones, Nils Lid Hjort, Ian R Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[44] Pearl Judea. Causality: Models, reasoning, and inference. *Cambridge University Press. ISBN 0*, 521(77362):8, 2000.

[45] Takafumi Kanamori and Hironori Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, 2015.

[46] Takayuki Kawashima and Hironori Fujisawa. Robust and sparse regression via γ-divergence. *Entropy*, 19(11):608, 2017.

[47] Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Semi-parametric counterfactual density estimation. *arXiv preprint arXiv:2102.12034*, 2021.

[48] Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 366–374, 2008.

[49] AJ Lawrance. On conditional and partial correlation. *The American Statistician*, 30(3):146–149, 1976.

[50] Yunting Liu, Xia Wu, Jiacai Zhang, Xiaojuan Guo, Zhiying Long, and Li Yao. Altered effective connectivity model in the default mode network between bipolar and unipolar depression based on resting-state fmri. *Journal of Affective Disorders*, 182:8–17, 2015.

[51] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[52] Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[53] Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 639–647, 2011.

[54] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33, 2020.

[55] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[56] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[57] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

[58] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[59] Paul R Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002.

[60] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[61] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[62] Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *J. Mach. Learn. Res.*, 21:39–1, 2020.

[63] Shohei Shimizu and Kenneth Bollen. Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-gaussian distributions. *J. Mach. Learn. Res.*, 15(1):2629–2652, 2014.

[64] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[65] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[66] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[67] Viktor Pavlovich Skitovich. Analyse générale des liasons stochastique. *Revue de l'Institut International de Statistique*, 21:2–8, 1953.

[68] Viktor Pavlovich Skitovich. Linear forms of independent random variables and the normal distribution law (in russian). *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 18(2):185–200, 1954.

[69] Viktor Pavlovich Skitovich. Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab*, 2:211–228, 1962.

[70] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT press, 2000.

[71] Jerzy Splawa-Neyman, D M Dabrowska, and T P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Stat. Sci.*, 5(4):465–480, 1923.

[72] Mariela Sued, Marina Valdora, and Víctor Yohai. Robust doubly protected estimators for quantiles with missing data. *TEST*, 63(3):819–843, 2020.

[73] Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: a causal ordering method robust against latent confounders. *Neural computation*, 26(1):57–83, 2014.

[74] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[75] Ryan J Tibshirani, Jonathan Taylor, et al. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[76] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[77] Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[78] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

[79] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[80] Alexander von Eye and Richard P DeShon. Directional dependence in developmental research. *International Journal of Behavioral Development*, 36(4):303–312, 2012.

[81] Y Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1):41–59, 2020.

[82] Michael P Windham. Robustifying model fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 599–609, 1995.

[83] Xiaojie Xu. Contemporaneous causal orderings of us corn cash prices through directed acyclic graphs. *Empirical Economics*, 52(2):731–758, 2017.

[84] Victor J Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656, 1987.

[85] K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

[86] K Zhang, J Peters, D Janzing, and B Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

[87] Kun Zhang and Lai-Wan Chan. Ica with sparse connections. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 530–537. Springer, 2006.

[88] Kun Zhang, Heng Peng, Laiwan Chan, and Aapo Hyvärinen. Ica with sparse connections: Revisited. In *International Conference on Independent Component Analysis and Signal Separation*, pages 195–202. Springer, 2009.

[89] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017.

[90] Zhiwei Zhang, Zhen Chen, James F Troendle, and Jun Zhang. Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706, 2012.

[91] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

[92] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[93] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.