Causal discovery in the presence of missing data

RUIBO TU
Causal discovery in the presence of missing data

RUIBO TU
Abstract

Missing data are ubiquitous in many domains such as healthcare. Depending on how they are missing, the (conditional) independence relations in the observed data may be different from those for the complete data generated by the underlying causal process (which are not fully observable) and, as a consequence, simply applying existing causal discovery methods to the observed data may give wrong conclusions. It is then essential to extend existing causal discovery approaches to find true underlying causal structure from such incomplete data. In this thesis, we aim at solving this problem for data that are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). With missingness mechanisms represented by the Missingness Graph, we present conditions under which addition corrected to derive conditional independence/dependence relations in the complete data. Combined with the correction method that gives closed-form, consistent tests of conditional independence, the proposed causal discovery method, as an extension of the PC algorithm, is shown to give asymptotically correct results. Experiment results illustrate that with further reasonable assumptions, the proposed algorithm can correct the conditional independence for values MCAR, MAR and rather general cases of values MNAR.
Sammanfattning

Saknade data är alltädes närvarande på många områden, t.ex. sjukvård. Beroende på hur de saknas kan de (villkorliga) oböverdeförhållanden i de observerade uppgifterna skilja sig från de för de fullständiga data som genereras av den underliggande orsaksprocessen (som inte är fullt observerbara) och som en följd av att helt enkelt tillämpa befintlig kausal upptäckt metoder för de observerade data kan ge felaktiga slutsatser. Det är då viktigt att förlänga befintliga metoder för kausala upptäckter för att hitta en sann underliggande kausalstruktur från sådana ofullständiga data. I denna avhandling strävar vi efter att lösa detta problem för data som saknas helt slumpmässigt (MCAR), saknas slumpmässigt (MAR) eller saknas inte slumpmässigt (MNAR).

Med missmekanismer representerade av Missfallsgrafen presenterar vi förhållanden under vilka tillägg korrigerade för att härleda villkorliga oböverdöende / beroendeförhållanden i de fullständiga uppgifterna. Kombinerad med korrigeringsmetoden som ger slutet form, konsekventa test av villkorligt oböverdöende, visas att den föreslagna orsaksökningsmetoden, som en förlängning av PC-algoritmen, ger asymptotiskt korrekta resultat. Experimentresultat illustrera att med ytterligare rimliga antaganden kan den föreslagna algoritmen korrigerar det villkorliga oböverdöende för värdena MCAR, MAR och ganska generella fall av värdena MNAR.
Acknowledgement

I would like to express my sincere thanks to my supervisor at KTH, Prof. Hedvig Kjellström for giving me the opportunity to pursue my master thesis with the supervision of another two wonderful cosupervisors and providing advice for organizing the project.

I am very grateful to Prof. Kun Zhang at Carnegie Mellon University and Dr. Cheng Zhang at Microsoft Research Cambridge for providing the constant support, guidance and feedback throughout the whole thesis.

I would like to thank Dr. Paul Ackermann at Karolinska University Hospital for providing and explaining the dataset. I would like to thank Prof. Joakim Gustafson at KTH for examining my thesis and organizing the public discussion.

Many people have helped me by discussing the project and providing comments, including Charles Hamesse, Bowen Kuang, Tao Wang, Shiping Song, Robin Juthberg and Marcus Bergholm.

Finally, I would like to thank my girlfriend Wen who gave me the incredible patience for my presentation and support throughout the six months.
# Contents

1 Introduction .................................................... 1
   1.1 Machine Learning in Healthcare ........................... 1
      1.1.1 Opportunities of Healthcare Data ................. 1
      1.1.2 Applications of Machine Learning in Healthcare . 2
      1.1.3 Challenges of Machine Learning in Healthcare . . 3
   1.2 Causality in Health Care .................................... 4
      1.2.1 Correlation and Causation ............................ 4
      1.2.2 Randomized Controlled Trials ....................... 4
      1.2.3 Graphical Causal Model ............................ 5
      1.2.4 Challenges of Causal Discovery in Healthcare . . 6

2 Related Works .................................................. 8
   2.1 Causal Discovery ........................................... 8
   2.2 Dealing with data with missing entries from causal perspective .................................................. 9

3 Method .......................................................... 10
   3.1 Behavior of deletion-based PC ............................ 10
      3.1.1 Missingness graph .................................. 10
      3.1.2 Assumptions on dealing with missingness .......... 11
      3.1.3 Deletion-based PC with incomplete data .......... 12
      3.1.4 Identify erroneous edges for deletion-based PC . 13
   3.2 Method ...................................................... 14
      3.2.1 Detecting causes of missingness variables .......... 14
      3.2.2 Correction for the conditional independence test 15

4 Experiment ...................................................... 18
   4.1 Baselines .................................................. 18
   4.2 Synthetic data evaluation ................................ 19
      4.2.1 Data Generation ................................. 19
4.2.2 Result ........................................ 19
4.3 The Cognition and Aging USA (CogUSA) study ....... 22
4.4 Achilles Tendon Rupture study ........................... 22

5 Discussion ......................................... 24
  5.1 Sustainability, ethics and social aspects .............. 24
  5.2 Challenges ...................................... 25
  5.3 Future work ...................................... 25

Bibliography .......................................... 27

A Unnecessary Appended Material ......................... 31
  A.1 Appendix ....................................... 31
Chapter 1

Introduction

1.1 Machine Learning in Healthcare

Machine learning has made massive achievement in different fields, such as retail, banking and finance and in the automotive industry. It is elaborated by three important factors: data, computational power and algorithms. Compared with other applications of machine learning, machine learning in health care has many distinct features and challenges, which we will now discuss.

1.1.1 Opportunities of Healthcare Data

The recent increase in healthcare data [19] opens up for new opportunities. Astonishing applications of machine learning in other fields keep reminding people in the healthcare sector that it is the time to consider the new technology to provide higher quality and efficiency of healthcare service.

A significant feature of healthcare data is the diversity. Various sources of data lead to the diversity. Healthcare data comprise clinical registries, administrative data, biometric data, patient reports, radiological imaging and the data from wearable devices [13]. The diversity of healthcare data provides to more possibility of collecting more data and combining them to gain more insights.

The governments, universities and companies are devoting themselves to the utility of machine learning in health care. The reformation of the healthcare system in the US leverages electronic health records (EHRs). According to the report of National coordinator for health
information technology, the adoption of EHRs has achieved 9-fold increases since 2008 [13]. At the same time, the EU 2020 strategy emphasised the accessibility and deployment of eHealth [11]. Pittsburgh Health Data Alliance consisted of UPMC enterprise, Carnegie Mellon University and Pittsburgh university also claims that the future of healthcare is data and aims at turning data into improved human health.

1.1.2 Applications of Machine Learning in Healthcare

Machine learning leverages massive data to predict outcomes and discover patterns of data. In the following, we discuss extraction of non-text records, decision support systems, public health and predictive system.

One application of machine learning algorithms is extracting the medical information from textual documents. According to the report of National Coordinator for Health Information Technology [3], the sources of EHR data are demographics, clinical diagnoses, laboratory tests, diagnostic studies, prescribed medications, vaccines, and selected health behaviors. Machine learning algorithms in natural language processing are used for extracting information from narrative data, like clinical diagnose. The extraction is also used for the further decision support model.

A decision support system is a helpful evidence-based tool. In many complicated situations, a disease might need many different treatments and the choice of treatments is always experience-based, such as Acute Achilles Tendon Ruptures [21]. In this case, data-driven tools are good candidates to guide clinical decisions, like IBM Watson which can provide the analysis of patients based on the EHR data.

The primary methods of public health are analysing the distribution and spread of diseases and social behaviors [28]. Recently, with the help of social media data, researchers can have a broader investigation and more profound insight into epidemiology and lifestyle diseases. [8] and [41] are trying to use images in Instagram and texts in Twitter to track lifestyle diseases, like obesity and drinking, and describe the spread of diseases, such as Influenza and Ebola.

Predictive models can be used for reducing the cost of healthcare. The possible use cases are high-cost patient, readmissions, triage, decompensation, adverse events and treatment optimization [2]. For ex-
ample, the paper [40] used a Bayesian network to predict pancreatic cancer with the combination of PubMed knowledge and EHRs. In the paper [1], they applied support vector machine, random forest and other machine learning methods to classify heart failure subtypes of patients.

1.1.3 Challenges of Machine Learning in Healthcare

In the following, we discuss the challenges in data issues, missing data, time dependencies and methodology problem.

Many distinctive properties of healthcare data are difficult to handle, such as data heterogeneity, high-dimensionality, sparsity, and high noise levels due to measurement errors and temporal dependencies. Healthcare data have different forms, like medical images, narrative reports, etc. Combining them is hard for traditional machine learning methods [31]. Treatment records and some specific data, like microarray data or next-generation sequencing data, might also have too many attributes which cause the curse of dimensionality problem for machine learning methods.

Moreover, missing data in clinical records and patients’ data are a common phenomena due to the limitation of medical resources, or high cost of acquiring them. Measurement errors and frequency of statistical mistakes are a common phenomena in the real cases. Some methods are greatly influenced by the contradictory conditional independence and dependences for variables in their intersections [37].

Furthermore, time dependencies might occur when treatments in different hospitals use different time unit when the treatments depend on the stage of disease with different symptoms [28] or when the underlying generating process changed over time [38].

As for methodology, machine learning is an observational study that finds correlations rather than causality. This is an essential factor in the decision support system [13]. Moreover, many machine learning algorithms are like black boxes and not interpretable in the general case, like deep learning methods. In this case, even if deep learning methods have the astonishing performance, the interpretability also limits the application of deep learning in health care.

Nevertheless, big data are sometimes not big enough in healthcare applications. For example, in [27], there are 57205 entries for 374 patients and 216 measurements. Data volume is big, but the number of
measurements is also big. So when it comes to measurements per patient, the data volume will be not big enough for machine learning algorithms like neural networks.

1.2 Causality in Health Care

As discussed before, machine learning can determine correlation rather causation. However, causation is necessary for evidence-based medicine [14] and proving the efficacy and safety of treatments. Causal effect between the intervention and outcome can be identified through randomized controlled trials (RCTs) that are gold standard for treatment comparison [12], but it is not always possible.

1.2.1 Correlation and Causation

Causation is not correlation [9]. Causation makes sense of data and explains why and how cause influences effect [23]. In healthcare, research is interested in causal questions. For example, does smoking cause cancer? Which is the main factor to increase outcomes of ATR measurements? However, machine learning methods and regression methods cannot infer causes from non-experimental samples. They can provide correlation information. But correlation cannot tell the differences of cause variable and result variable. So this disadvantage limits the application of correlation on interpretability of algorithms and decision support system.

Experimental methods, like Randomized Controlled Trials (RCTs), and graphical causality methods, like causal discovery and causal inference, are two common ways to infer causal relationships. Compared with experimental methods, causality methods need more assumptions to find causal information from correlation.

1.2.2 Randomized Controlled Trials

In [35], the author claims that RCTs are the best way to determine causal effect between intervention and outcome. RCTs randomly assign patients into two groups, control group and experimental group. Patients in the experimental group are given the intervention treatments, while patients in control group are given the placebo. The efficacy and safety of treatment can be analyzed by the different outcomes
of two groups [10]. Furthermore, the CONSORT (Consolidated Standards of Reporting Trials) statement figures out the requirements of RCTs which make RTCs more reliable [18]. For example, the population sampling and randomisation efficiently avoid selection bias and confounding factors’ impact.

Nevertheless, RCTs also have some practical and ethical problems that will influence the reliability of the result. In [20], the author lists issues of RCTs: 1. hard to assess rare outcomes 2. hard to follow-up, like long-term outcomes 3. effectiveness of the intervention can be influenced by participators’ preference 4. ethical and political obstacles 5. low recruitment rates and insufficient samples.

Compared with the time-consuming and expensive randomized experimental methods, causal discovery is a powerful tool that leverages the observation study to determine the causal relationships among the factors [9].

1.2.3 Graphical Causal Model

The Markov assumption and the Faithfulness assumption provide the foundation of a graphical causality framework. These assumptions connect statistics and causality and give the graphical method to find causality and make prediction.

With the assumptions, the graphical causal model combines the probability, graphs and structural equations. Structural causal model (SCM) represents the causal relationship between variables, like the causal effect of a treatment on a disease. SCM consists of variables and functions. Cause and effect can be regarded as variables, and the value of effect variables is assigned by cause variables. Furthermore, every SCM is associated with a directed acyclic graph (DAG). In the graph, variables are represented by node, and function is represented by the arrow from one variable to another [23].

In graphical causal models, causal discovery and causal inference are two main parts. Causal discovery recovers the underlying causal structure of variables from observation data. And Causal inference predicts the effects of intervention without performing experiments.

Causal discovery consists of independence-based and score-based methods. Independence-based methods assume that all conditional independences are entailed from Markov condition, and have two common steps that are an estimation of the undirected graph and make
sure the direction of edges, such as PC and fast causal inference (FCI). Both methods that are asymptotically correct start with the complete undirected graph, and give results that satisfy the same conditional dependence. However, PC fails when the variables have unmeasured common causes [9].

Score-based methods can distinguish different DAGs with equivalent classes. They are benefited with further assumptions, like the parametric models, such as linear, non-Gaussian, acyclic model, nonlinear additive noise model and post-nonlinear model [38]. With the parametric model, the typical method is Greedy Equivalence Search [5].

1.2.4 Challenges of Causal Discovery in Healthcare

Causal discovery faces many practical challenges, such as missing data, measurement error, selection bias, temporally aggregated time series, nonstationary data and deterministic case [38]. Some problems, like measurement error, nonstationary data and temporally aggregated time series, have been mentioned in the section 1.1.3.

Missing data is one of the major issues. Missing data issue arises when values for one or more variables are missing from recorded observations [33]. And It is very common that medical records or datasets have plenty of missing data. The missing data may come from imperfect data collection, various types of censoring, such as loss to follow up, various factors, such as high cost involved in measuring variables, failure of sensors, the reluctance of respondents in answering specific questions and an ill-designed questionnaire [17]. Missing data lead some harmful consequences. It can cause significant bias of research studies because the response profiles of non-respondents and respondents can be significantly different from each other. Moreover, the current causal discovery algorithms are based on complete datasets. So we need causal discovery algorithms which can handle the missing data problem [24] [30].

Moreover, most existing algorithms for causal discovery are designed for complete datasets [34, 22]; for instance, the PC [34] makes use of conditional independence constraints in the data to recover information of the underlying causal process and has been widely used. Unfortunately, missing data entries are very common in a number of domains, especially in healthcare observational data. In healthcare
data, missing entries may come from imperfect data collection, various types of compensatory medical instrument, fitness of the patients, etc. [17]. There are three types of missing mechanisms [29]: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR if the missingness mechanism is independent of any variable in the system, MAR if it depends on some fully observed variable, and MNAR if it depends on a variable with missing entries.

Although very little work has been done on developing algorithms for causal discovery from incomplete data, the missing data issue has received much attention. In particular, recoverability [17] is about whether a query (say, a conditional or joint distribution) for the complete data can be estimated from the incomplete data and the m-graph asymptotically correctly, and some sufficient conditions for testability of certain d-separation relations were given by [16]. Generally speaking, challenges remain in the case of MNAR. To the best of our knowledge, there exists only a so-called test-wise deletion FCI algorithm [36] for causal discovery with MNAR data. The algorithm is based on FCI [34], and its output is not a Directed Acyclic Graph (DAG) or its equivalence class [34], but a Partial Ancestral Graph (PAG), which may not be informative enough because it may contains edges between variables that were originally not directly causally related but produced by missingness or biased selection. For instance, if X and Y are independent in the complete data but their missingness mechanisms are related, then the output will contain an edge between them.

We aim to develop an algorithm that is able to recover the true causal graph or its equivalence class for the variables of interest from their incomplete observations, under appropriate assumptions, by extending the PC algorithm. Our main contribution includes:

- We provide theoretical analysis on the error that different missing mechanism introduce in causal discovery (Section 3.1).

- Based on this analysis, we develop a correction-based method that can handle all three types of missing mechanism - MCAR, MAR and MNAR - under mild assumptions (Section 3.2).

- In addition to experimental results on synthetic data to verify our claims, we further apply our method to two real-world healthcare datasets. The results are consistent with medical domain knowledge (Chapter 4).
Chapter 2

Related Works

We start by discussing work that is closely related to the studied problem, including traditional causal discovery algorithms and approaches to dealing with missing data from a causal perspective.

2.1 Causal Discovery

Causal discovery from observational data has been of great interests in various domains in the past decades [22, 34]. In general, causal discovery consists of two families of methods: score-based methods and constraint-based methods [25]. Score-based methods find the best DAG under certain score-based criterion, such as the Bayesian information criterion (BIC). Greedy Equivalence Search (GES) [5] is a popular method in this category. Constraint-based methods rely on conditional independent tests. It assumes that all conditional independences are entailed from causal Markov condition under faithfulness assumption. Typical constraint-based methods include PC and fast causal inference (FCI). Both methods start with the complete undirected graph and give results that satisfy the same conditional dependence. Among these two methods, PC has an advantage on produce more informative output, CPDAG, where the causal directions are shown. FCI only output PAG, however, has the advantage of handling latent variables. Our proposed method is based on PC since the interpretable causal directions are of great need in health-care applications.
2.2 Dealing with data with missing entries from causal perspective

There exist a number of studies dealing with missing data from a causal perspective since [29]. In particular, recoverability and testability have been studied in [17, 4, 16, 32]. Given an m-graph, recoverability is an important issue concerning whether a query for a complete data can be found. Testability is the key of causal discovery. Testability of the conditional independence [16] is that the condition independence can be refuted by all the underlying distribution of the partially, fully observed variables and missingness indicator variables. However, the aim of causal discovery is to find relations between the true causal graph variables that we are interested rather than missingness indicators. Even if the conditional independence between missingness indicators and variables of interest may not be testable, causal relations in the truth causal graph without missingness indicators can be recovered. Moreover, when the conditional independence is testable, only the implication of the conditional independence is not enough to test the conditional independence within the missing mechanism.

In causal discovery, there are few works for MNAR. Fast causal inference (FCI) by test-wise deletion [36] was exploited to handle missing data: by considering the missingness procedure as a particular type of selection bias, it was shown that FCI combined with test-wise deletion for conditional independence tests is still sound when once aims to estimate the PAG for the variables including the effect of missingness (a particular type of selection bias), which is usually different from the causal graph for generating the variables we are concerned with. Besides the informative problem mentioned in section 1, data missingness is usually different from selection bias because in the selection bias case we have only the distribution of the selected sample but no clue about the population, while in the missing data case we may be able to check the conditional independence relationship between two variables by making use of the available data of other variables. In the case where the missingness mechanisms are known, this problem is closely related to testability of models with missing data [16].
Chapter 3

Method

3.1 Behavior of deletion-based PC

In this section, we discuss the influence of missing data on causal discovery. In particular, we utilize the missingness graph, and give assumptions of our work. We provide naive extensions of test-wise deletion based and list-wise deletion based method on PC and discuss properties of the result produced by these algorithms in the presence of MCAR, MAR, and MNAR.

3.1.1 Missingness graph

We utilize the notation of Missingness Graph (m-graph) [17] to represent missingness mechanisms of variables and their causal relations in our work. In the original definition in [17], an m-graph is a causal DAG over variable set $V = V \cup U \cup V^* \cup R$. $U$ is the set of unobserved nodes; in this paper we assume causal sufficiency, so $U$ is an empty set. $V$ is the set of observable variables containing $V_o$ and $V_m$. $V_o \subseteq V$ is the set of fully observable variables that are observed in all records, they are denoted as white notes in our graphical representation. $V_m \subseteq V$ is the set of partially observable variables that are missing in at least one record, which are shadowed in gray. $R$ denotes missingness indicators, and $R_y \in R$ is the corresponding missingness indicator of $Y \in V_m$. Associated with $Y$, its corresponding missingness indicator $R_y$ represents the missingness mechanism, where 1 presents that the corresponding value is missing, and 0 indicating that the corresponding value for the variable is observed. Proxy variable $Y^*$ is introduced.
as an auxiliary variable for the convenience of derivation observed which can be determined by $Y$ and $R_y$. It takes the value of $Y$ if $R_y = 0$, and corresponds to a missing entry if $R_y = 1$.

Figure 3.1: Missingness graphs in MCAR, MAR and MNAR. The gray nodes denote the partially observable variables, and the white nodes are the fully observed variables. $\{R_y, R_z\} \subset \mathbb{R}$ are missingness indicators.

### 3.1.2 Assumptions on dealing with missingness

Let $\{X, Y\} \in V$ denote random variables and $Z \subseteq V \setminus \{X, Y\}$. The conditional independence (CI) between $X$ and $Y$ given $Z$ is denoted by $X \perp\!\!\!\!\!\!\!\!\perp Y|Z$. $R_K$ is any subset of $R$ and conditioning on $R_K = 0$ means conditioning on all the elements of $R_K$ indicators are equal to zero. Following the result from [16], apart from the basic assumptions for PC with fully observed data, we make the following additional assumptions for all methods that address missing data entries in this paper.

- **Faithful observability**: We assume that $X \perp\!\!\!\!\!\!\!\!\perp Y|\{Z, R_K = 0\} \iff X \perp\!\!\!\!\!\!\!\!\perp Y|\{Z, R_K\}$.

- **Non-Causal missingness indicator**: We assume that a missingness indicator cannot be deterministically related to other missingness indicators or be the cause of variables in $V$.

- **No self-missingness**: We assume that the missingness indicator $R_y \in R$ for $Y \in V_m$ can not be caused by $Y$ itself; in fact, "self-missingness" is generally untestable [16, 17]. We further note that the "self-missingness" only effects casual discovery results when $R_y$ has other parents apart from $Y$ (See discussion in the Appendix).
In this paper, we further assume that the causal relation is linear. Thus, we can utilize simple conditional independent test. However, our proposed algorithm applies to general situation. In non-linear situation, we can replace the linear independent test method by a suitable non-linear one or non-parametric one [39].

3.1.3 Deletion-based PC with incomplete data

In the presence of missing data, list-wise deletion PC deletes all records that have any missing value and then applies PC to the remaining data. In contrast, test-wise deletion PC determines records to delete when performing PC [36]: it only deletes records with missing values for variables involved in the current CI test. In this paper, the CI test on records that do not have missing values for variables involved in this test is called test-wise CI test.

Deletion-based PC gives asymptotically correct results in the case of MCAR, such as Figure 3.1a. In this case, with the non-causal missingness indicator assumption $(X, Y, Z) \perp \!\!\!\!\perp R_K$ holds. With the faithful observability assumption, the CI relation between $X$ and $Y$ given $Z$ is not influenced by conditioning on $R_K = 0$; formally, $P(X, Y|Z) = P(X, Y^*|Z, R_K = 0)$. When only including $R_y \in R_K$ in the CI test, $P(X, Y|Z) = P(X, Y^*|Z, R_y = 0)$ still holds. According to the m-graph definition, list-wise deletion can be represented as conditioning on missingness indicators with zero value. Test-wise deletion can be represented as conditioning on missingness indicators whose corresponding variables involved in the current CI test with zero value. Thus, test-wise deletion is more data efficient than list-wise deletion method.

In case of MAR and MNAR, such as Figure 3.1b and 3.1c, deletion-based PC may produce erroneous edges in the result. In Figure 3.1b, both $X$ and $Y$ have a path to $R_y$. When conditioning on $R_y$, $X$ is generally not independent with $Y$; in fact, if faithfulness is assumed for the m-graph, $X \not\perp\!\!\!\!\perp Y|R_y$; furthermore, under the faithful observability assumption, it is equivalent to $X \not\perp\!\!\!\!\perp Y^*|R_y = 0$. As for test-wise CI test, the independence relation between $X$ and $Y$ is regarded as the conditional independence relation between $X$ and $Y$ conditioning on $R_y = 0$. The wrong result of independence test between $X$ and $Y$ introduces an extra edge between them which is not included in the true causal graph. The same applies for MNAR as in Figure 3.1c. There-
fore, in the following sections, we mainly solve problems of applying deletion-based PC to the data in MAR and MNAR.

### 3.1.4 Identify erroneous edges for deletion-based PC

Since the deletion in the presence of MAR and MNAR may introduce erroneous edges in the result of PC algorithm, we identify possible errors of deletion-based PC and correct them. Firstly, we show that when we directly apply test-wise deletion PC, the skeleton (undirected graph) has no missing edges, but might contain extra edges, compared with the true causal graph. We then study under what conditions, extra edges can be produced by test-wise deletion PC. Finally, we show that we could only consider the better method, test-wise deletion PC because the set of extra edges produced by test-wise deletion PC is the subset of the set of extra edges produced by list-wise deletion PC.

**Proposition 1.** Under assumptions in Section 3.1.2 and the faithfulness assumption on the m-graph, the test-wise CI, which is \( X \perp \perp Y \mid \{Z, R_x = 0, R_y = 0, R_z = 0\} \), can imply the CI with complete data, which is \( X \perp \perp Y \mid \{Z\} \), where \( X \) and \( Y \) are two random variables and \( Z \) is a set of random variables.

The proof is given in the appendix. The skeleton search of PC is based on the results of CI tests. In the Proposition 1, we show that the CI relations in test-wise deleted dataset imply the CI relations in the complete dataset. However, when variables in the test-wise deleted dataset are not (conditionally) independent, the dependency might be caused by the test-wise deletion rather than the true causal relation. The wrong results of test-wise CI tests produce extra edges in the causal graph. Therefore, we want to detect such extra edges and correct them after the skeleton search of test-wise deletion PC. Fortunately, extra edges produced by test-wise deletion PC follow some particular graph structures.

**Proposition 2.** Suppose \( X \perp \perp Y \mid \{Z\} \), but \( X \not\perp \perp Y \mid \{Z, R_x = 0, R_y = 0, R_z = 0\} \), then under our assumptions, there exists a variable set \( Z \subseteq V \setminus \{X, Y\} \), such that \( X \perp \perp Y \mid Z \) and that for at least one variable in \( \{X\} \cup \{Y\} \cup Z \), its corresponding missingness indicator is a common descendant of \( X \) and \( Y \).

The proof is given in the appendix. Proposition 2 indicates that the extra edge between two variables can be removed by correcting
their conditional independence relation with their direct common effects which relevant missingness indicators depend on. Although we do not know their direct common effects, we can consider all variables that are adjacent to both X and Y as for such correction. Therefore, in the following section, we will introduce how to find the variable or variables adjacent to missingness indicators, and how to detect extra edges in particular graph structures by correcting them.

With these two propositions, the set of extra edges produced by list-wise deletion PC, denoted by $E_{\text{list}}$, contains the set of extra edges produced by test-wise deletion PC, denoted by $E_{\text{test}}$. We can easily get the conclusion that the CI in the list-wise deleted dataset implies the CI in the complete dataset with the similar proof of Proposition 1; moreover, the missingness indicators in each list-wise CI test contain the ones in the test-wise CI test. Thus, we have $E_{\text{test}} \subseteq E_{\text{list}}$. Therefore, we mainly discuss test-wise deletion PC that has less extra edges in the following sections.

3.2 Method

As mentioned in Section 3.1, the presence of MAR and MNAR may introduce erroneous edges in the result of list-wise deletion and test-wise deletion PC algorithm. We propose a method that can detect such erroneous edges by conducting correction of CI relations. The correction aims to test whether a particular conditional independence relation holds on the underlying complete data by analyzing data with missing entries. Then we applied our method to PC algorithm, naming Missing Value PC (MVPC), that can discover the true causal graph or its equivalence class, even in cases of MAR and MNAR.

3.2.1 Detecting causes of missingness variables

Because the correction method is based on variables which relevant missingness indicators depend on, we first introduce how to detect the variable or variables adjacent to missingness indicators. With the non-causal missingness indicator assumption, the variable or variables adjacent to missingness indicators are causes of missingness indicators.

In the case of MAR, causes of missingness indicators are all fully observed. The variable or variables adjacent to a missingness indicator, denoted by $R_x \in \mathbb{R}$, can be detected by the skeleton search of
PC algorithm which only includes (conditionally) independent relation tests between $R_x$ and $v_i \in V_o$. In fact, under assumptions in Section 3.1.2, when conditioning on the variable or variables adjacent to $R_x$, $R_x$ is conditionally independent with all the other variables. Thus, the skeleton search can remove edges for such (conditionally) independent relations and find the variable or variables adjacent to $R_x$.

In the case of MNAR, causes of missingness indicators are partially observed variables. The variable or variables adjacent to $R_x$ can be detected by the skeleton search of test-wise deletion PC including all the variables in $V \setminus \{X\}$, and the test-wise deletion will not produce the extra edge between $R_x$ and any other variables. Because the extra edge only happens if $R_x$ and an included variable have at least one common descendant, according to Proposition 2; however, with the non-causal missingness indicator assumption, $R_x$ cannot be a cause. Therefore, in this case, test-wise CI test is asymptotically correct. The algorithm for detecting the variable or variables adjacent to missingness indicators is summarized in the appendix.

After detecting causes of missingness indicators, we want to detect extra edges by correcting them only for the particular structures. As mentioned Section 3.1.4, the particular structure can be determined with Proposition 2.

### 3.2.2 Correction for the conditional independence test

We now introduce our correction method that can correct possible extra edges in the result of test-wise deletion PC.

The intuitive idea of our correction method is that the test-wise CI test is asymptotically correct by conditioning on causes of missingness indicators involved in the test-wise CI test, and then marginalizing them. In Figure 3.1b, the consistent estimate for MAR can be represented in the Equation 3.1.

$$
P(X, Y) = \sum Z P(X, Y \mid Z)P(Z) = \sum Z P(X, Y \mid Z, R_y = 0)P(Z) = \sum Z P(X, Y* \mid Z, R_y = 0)P(Z) \tag{3.1}
$$

The consistent estimate for MNAR (as in Figure 3.1c ) can be repre-
presented in the Equation 3.2.

\[
P(X, Y) = \sum_Z P(X, Y \mid Z)P(Z) \\
= \sum_{Z^*} P(X, Y^* \mid Z^*, R_y = 0, R_z = 0)P(Z^* \mid R_z = 0) \\
= \sum_{Z^*} P(X, Y^* \mid Z^*, R_y = 0, R_z = 0)P(Z^* \mid R_z = 0) \tag{3.2}
\]

Our method implements this idea with generating virtual data for the test-wise CI test. The generated virtual data keep the same CI relation between variables with the CI relation underlying the complete data and break paths from variables to missingness indicators.

**Algorithm 1** Generating data for testing that X is conditional independent with Y given Z

**Input:** data of variables in the CI relation, such as X, Y and Z, and causes of their missingness variables, denoted by W

**Output:** generated data of variables in the CI relation, such as X, Y and Z

1. Delete records containing any missing value, denoted by \(D_d\), and denote the original dataset as \(D_o\)
2. Regress X and Y on Z and W with \(D_d\) and denote linear regression models as \(M_x\) and \(M_y\)
3. Save residual parts of \(M_x\) and \(M_y\), denoted by \(RSS_x\) and \(RSS_y\)
4. Shuffle data of W in \(D_o\) and delete records containing any missing value
5. Generate data of X, Y with data of Z and shuffled W, denoted by \(\tilde{W}\).
6. Add \(RSS_x\) and \(RSS_y\) back to generated data
7. **return** the generated data of \(\hat{X}, \hat{Y}\) and Z

The correction method consists of two steps. In the first step, after getting the test-wise deleted dataset of a CI relation, we learn two linear regression models from variables in the test-wise CI test and causes of the corresponding missingness indicators. In each linear regression model, scalar responses are respectively the variables: the (conditional) independence relation between them is tested by the test-wise CI test. Explanatory variables are causes of the corresponding missingness variables and variables that are conditioned on in the test-wise
CI test. In the second step, causes of missingness indicators are permuted so that they are independent of missingness indicators. With permuted values of missingness indicators and variables that are conditioned on in the test-wise CI test, new data are generated with linear regression models and their corresponding residuals. Therefore, using the generated virtual data, we can perform a CI test that corrects the erroneous extra edge in the result of test-wise CI test. To be noticed, we use linear regression model due to our focus on linear models. In non-linear case, one can choose a proper generator based on the (conditional) independence test method.

Taking the m-graph in Figure 3.1c as an example, two regression models in Equation 3.3 are learned from data. With the permuted data $\tilde{Z}$, new data are generated with the linear regression model and the residuals in Equation 3.4. Then the independence test for $X \perp Y$ is based on the data of $\hat{X}$, and $\hat{Y}$. Algorithm 1 summarizes procedures.

\begin{align*}
X &= \alpha_1 Z + \epsilon_1 \quad (3.3) \\
Y &= \alpha_2 Z + \epsilon_2
\end{align*}

\begin{align*}
\hat{X} &= \alpha_1 \tilde{Z} + \epsilon_1 \quad (3.4) \\
\hat{Y} &= \alpha_2 \tilde{Z} + \epsilon_2
\end{align*}

In the end, we apply the correction method to the test-wise deletion PC, named missing value PC (MVPC). We first use the skeleton search of test-wise deletion PC, and then for specific graph structures, we remove extra erroneous edges with our correction method. Finally, the orientation determination method from PC to produce the final output in forms of CPDAG.
Chapter 4

Experiment

We evaluate our method MVPC on causal discovery on both synthetic datasets and real-world healthcare datasets. We first display results of experiments on synthetic data (Section 4.2) to demonstrate the behavior of our method in a controlled environment. After that, we apply our method to two healthcare datasets where data entries are severely missing. The first one is from the Cognition and Aging USA (CogUSA) study [15] (Section 4.3). The second one is about Achilles Tendon Rupture (ATR) rehabilitation research study [26, 7]. The output causal relationships using MVPC consistently demonstrate superior performance compared with multiple baseline methods.

4.1 Baselines

Our baseline methods include: PC with list-wise deletion, denoted by "list" which is the traditional way to deal with missing data entries. PC with test-wise deletion, denoted by "test", which can be seen as PC realization of [36]. Additionally, we also present PC on the oracle data (without missing data), denoted by "Ideal", for reference. In the end, to decouple the effect of sample size, we construct virtual dataset in MCAR with the same samples size of each conditional independence test. We denote it as "target" using PC with virtual MCAR data as a reference.
4.2 Synthetic data evaluation

To best demonstrate the behavior of different causal discovery method. We first perform the evaluation using synthetic data where ground truth causal graph is known. The goal is to recover the ground truth causal graph using causal discovery algorithms.

4.2.1 Data Generation

We followed procedures in [6] and [36] to randomly generate Gaussian DAG firstly and sample data based on the given DAG. Additionally, we at least include two collider structures in the random Gaussian DAG, because our analysis in Section 3.1, the erroneous result of MAR and MNAR are shown when a common effect leads to missingness. We generated two groups of synthetic data to show the scalability of our methods: one group is with 20 nodes (with 6-10 nodes are partially observed), and another group is with 50 (with 10-14 nodes are partially observed) nodes for both MAR and MCAR. To be noticed, in MNAR case, we assume that the cause of missing is partially observed. This is different from [36] which assume the cause is a hidden variable. For each group of the experiment, we generate 400 DAG with the sample size of 100, 1000, 5000 and 10000.

4.2.2 Result

We compared the result of algorithms with structural Hamming distance (SHD) from the ground truth shown in Figure 4.1, as well as, the precision and recall of their adjacencies and orientation in Figure 4.2 in all different experimental settings.
Figure 4.1: Performance comparison using structural Hamming distance (SHD). Lower value is better. Panel (a) shows the performance comparison in case of MAR with 20 variables. Panel (b) and (c) show the performance for MNAR with 20 and 50 variables separately.
CHAPTER 4. EXPERIMENT

Figure 4.2: Precision and recall for adjacencies and orientation comparison (Higher is better). All experiments above use 20 nodes with 10000 data samples.

Cross all different metrics in Figure 4.1 and Figure 4.2, we can see that our proposed algorithm consistently have superior performance comparing to both test-wise and list-wise deletion based method and are very close to the "target" performance. Similar to [36], the test-wise deletion also performs better than the list-wise deletion method in the context of PC. Additionally, our proposed method benefit from large volume of data samples as shown in Figure 4.1 in contract to [36].
4.3 The Cognition and Aging USA (CogUSA) study

In this experiment, we aim to discover the causal relationship in the cognition study as used in [36]. This is a typical survey based healthcare dataset with a large amount of variables with missing values. In this scenario, the missing mechanism is unknown and we could expect MCAR, MAR and MNAR occur.

We use the same 16 variables as [36] whose causal relationships are of interest in the cognition and aging study. Since the missing mechanism can be introduced by other partially observed variables as well, we also utilize the rest 88 numeric to find the causal relations in the 16 variables which we are interested in. We use 100 bootstrap population for the experiment and use BIC-score test for conditional independence. Figure 4.3 shows the performance evaluated using the known causal constraints as in [36]. In general, we know that the variables are in two groups where no inter-group causal relationship exist (rule 1), and we also know the causal link exists among two pairs of variables from domain expertise (rule 2). Each violation of these known relationships adds 1 in the cost shown in Figure 4.3.

Our proposed method obtains the best performance (lowest cost) comparing with deletion based PC and deletion based FCI [36]. This demonstrates the superiority of our method in real life healthcare application where missingness can be caused by other partially observed variables.

4.4 Achilles Tendon Rupture study

In the end, we perform causal discovery on Achilles Tendon Rupture (ATR) rehabilitation. ATR is one type of soft tissue injury which involves a long rehabilitation process. Understanding the causal rela-
tionship between various factors and the healing outcome is essential for practitioners. We use an ATR dataset [26, 7] collected in multiple hospitals. About 70% data entries are missing in this dataset, which means there is barely any patient data are fully observed. In this case, list-wise deletion method is not applicable due to lack of data. We apply our method and test-wise deletion PC for causal discovery here.

![Diagram](image)

(a) Consistent results  (b) Test-wise deletion PC  (c) MVPC (proposed)

Figure 4.4: Achilles Tendon Rupture (ATR) causal discovery results. The experiment were run over all variables. We demonstrate only the part of the whole causal graph. Panel (a) shows the relations among five variables discovered by MVPC. These relation is consistent with medical study. Panel (b) and (c) shows an example where MVPC is able to correct the error of test-wise deletion PC.

We ran the experiment with the full dataset with more than 100 variables. Figure 4.4a shows part of the causal graph. We see from the causal graph discovered by MVPC that age, gender, BMI (body mass index) and LSI (Limb Symmetry Index) does not effect on the final healing outcome measured by Foot Ankle Outcome Score (FAOS). This result is consistent with the medical study [26, 7]. To test the effectiveness of MNAR, we further introduce an auxiliary variable $S$ which is generated from two variables: Operation time ($OP_{time}$) and FAOS1 (a component of FAOS). This variable further cases missingness of FAOS1. Figure 4.4b and 4.4c show the results of these variables. Our proposed MVPC can correctly remove the additional erroneous edge between Operation time and FAOS1.
Chapter 5

Discussion

In this work, we address the question of causal discovery from partially observed data, which are typical in the healthcare domain. We first provide theoretical analysis on the possible error raised from different missingness mechanisms. Our study shows that erroneous causal edges are introduced when a common effect causes the missingness of a parent variable. Based on our analysis, we propose a novel algorithm MVPC which can correct this type of error under very mild assumptions. We demonstrate the effectiveness of our method in both synthetic data and real-world healthcare applications.

5.1 Sustainability, ethics and social aspects

Our project is trying to find out causal relations of Achilles Tendon Rupture rehabilitation measurements. From the sustainable perspective, causal discovery methods can avoid further experiments for potential causal relations. Since Achilles Tendon Rupture (ATR) rehabilitation is a prolonged process which takes difficulties to improve the treatment of long-term outcomes [27], causal discovery methods can conclude causal relations from observable data without further experiments and save much more time and medical resources.

Moreover, from the ethical perspective, the interruptibility of a decision-making system is necessary. In other words, the reason for results of machine learning methods in healthcare is necessary for both doctors and patients. However, the correlation from many machine learning methods is not enough for such explanation, compared with the causation from causal discovery. For example, some studies focus-
ing on surgical and non-surgical treatments show that surgical treatments are better considering the lower return risk. However, the reason for the lower return of surgical treatments might be the unknown confounder of lower return risk, surgical treatments and non-surgical treatments rather than surgical treatments cause the low return risk[21].

As for social aspects, doctors could get suggestions for the potential important factors of patients’ outcomes among intrinsic factors, such as age, sex, and body mass index (BMI), extrinsic factors, such as activity level and compliance with treatment, and unknown confounder. With suggested factors, doctors could implement experiments more efficiently. Moreover, if we can determine causal relations to guide a clinical decision, patients would avoid time-consuming, expensive and uncomfortable treatments.

5.2 Challenges

Causal discovery is of great need in healthcare applications. However, in the real world dataset, there are many challenges.

The reliability of data might influence the reliability of the result. The Data collection process and the data source are possible reasons for the unreliability of data. For example, in the ATR dataset, variables about surgery details are extracted from doctors’ reports. However, the report might not include everything and might make mistakes. The unreliable data might not satisfy the faithfulness condition. Thus, the result of causal discovery is not reliable in this case.

Another challenge is that the related variables of missingness indicators are not included in the dataset. As for our method, we need causes of missingness indicators. If the cause cannot be observed, our method can only figure out potential wrong relations rather than getting over the influence of missing data.

5.3 Future work

In the future work, we would explore the possibility to relax the assumptions further, as well as, work jointly with practitioners for practical usage of such method in the healthcare system.

In this paper, we have the non-causal missingness indicator assumption which requires that missingness indicators cannot be deter-
ministically related and be the cause of other variables. However, the missingness indicator can be the cause of other missingness indicators if missing data of a variable are the reason why another variable has missing entries. What is more, our method are based on the linear Gaussian assumption. But real datasets are not always followed this assumption and might influence causal relations of the result. Therefore, we will combine the non-linear non-Gaussian methods and our method.
Bibliography


[28] Daniele Ravì et al. “Deep Learning for Health Informatics”. In: ()


Appendix A

Unnecessary Appended Material

A.1 Appendix

\[ X \rightarrow Y \rightarrow Z \]

(a) One self-missingness case has no influence on the conditional independence test

\[ X \rightarrow Y \rightarrow W \]

(b) One self-missingness case has influence on the conditional independence test

Figure A.1: Self-missingness analysis.

Proof. Proposition 1 Because of the definition of m-graph, we just need to show that if \( X \) and \( Y \) are test-wise independent conditional on \( Z \), we have \( X \perp\!\perp Y | \{Z, R_Z = 0, R_X = 0, R_Y = 0\} \), where some of the involved missingness indicators may only take value 0 (i.e., the corresponding variables do not have missing values). With the faithful observability assumption, the above condition implies \( X \perp\!\perp Y | \{Z, R_Z, R_X, R_Y\} \). Because of the faithfulness assumption on the m-graph, we know that \( X \) and \( Y \) are d-separated by \( \{Z, R_Z, R_X, R_Y\} \); furthermore, according to our assumption, the missingness indicators can only be leaf nodes in the m-graph. Therefore, conditioning on these leaf nodes will not destroy the above d-separation relationship. That is, in the m-graph, \( X \) and \( Y \) are d-separated by \( W \). Hence, we have \( X \perp\!\perp Y | Z \). \( \square \)
Algorithm 2 Detect causes of missingness indicators

Input: a graph $\mathcal{G}$ where each missingness indicator, denoted by $R_t$, connects to all the variables in $V \setminus \{T\}$.

1: repeat
2: Select a missingness indicator $R_t \in \mathcal{R}$
3: repeat
4: Select a variable $X \in V \setminus \{T\}$
5: $l = -1$
6: repeat
7: $l = l + 1$
8: Choose a set $S \subseteq V \setminus \{T, X\}$ with $|S| = l$
9: if $X$ or $S$ is partial-observed variable then
10: Delete the data of $X$, $S$ and $R_t$ where at least one of them has a missing value
11: end if
12: if $R_t$ is conditional independent with $X$ given $S$ then
13: Delete the edge between $R_t$ and $X$ in $\mathcal{G}$
14: end if
15: until all $S \subseteq V \setminus \{T, X\}$ with $|S| = l$ has been considered
16: until all the variables in $V \setminus \{T\}$ have been considered
17: until all missingness indicators have been considered
18: return $\mathcal{G}$
Proof. Proposition 2 The premise says that there exists a variable set $Z \subseteq V \setminus \{X, Y\}$ such that $X \perp\!\perp Y \mid Z$. Moreover, we know that $X$ and $Y$ are not d-separated conditional on $Z \cup \{R_X\} \cup \{R_Y\} \cup R_Z$. Therefore, some of the non-constant elements of $\{R_X\} \cup \{R_Y\} \cup R_Z$ are common descendants of $X$ and $Y$. \qed