UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

# INTEGRATIVE CAUSAL ANALYSIS OF HETEROGENEOUS DATA SETS

BY

SOFIA TRIANTAFILLOU

Ph.D. DISSERTATION

HERAKLION FEBRUARY 2015

# Integrative Causal Analysis of Heterogeneous Data Sets

Sofia Triantafillou

February 2015

University of Crete

Department of Computer Science

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Supervisor: Ioannis Tsamardinos

UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

**Integrative causal analysis of heterogeneous data sets**

Dissertation submitted by
**Sofia Triantafillou**
In partial fulfilment of the requirements for the
degree of Doctor of Philosophy in Computer Science

Author: ______________________________
Sofia Triantafillou, University of Crete

Examination Commitee: ______________________________
Ioannis Tsamardinos, Associate Professor, University of Crete

______________________________
Dimitris Plexousakis, Professor, University of Crete

Takis Benos

Digitally signed by Takis Benos
DN: cn=Takis Benos, o=University of Pittsburgh, SOM, ou=Department of
Computational and Systems Biology, email=benos@pitt.edu, c=US
Date: 2015.03.22 10:32:50 -05'00'

______________________________
Panagiotis Benos, Associate Professor, University of Pittsburgh

______________________________
Vassilis Christophides, Professor, University of Crete

Gregory Cooper

Digitally signed by Gregory Cooper
DN: cn=Gregory Cooper, o=University of
Pittsburgh, ou=Department of Biomedical
Informatics, email=gfc@pitt.edu, c=US
Date: 2015.03.17 00:03:42 -04'00'

______________________________
Gregory Cooper, Professor, University of Pittsburgh

Clark Glymour

Digitally signed by Clark Glymour
DN: cn=Clark Glymour, o, ou, email=cg09@andrew.cmu.edu,
c=US
Date: 2015.03.13 12:24:32 -05'00'

______________________________
Clark Glymour, Alumni Professor, Carnegie Mellon University

______________________________
Marloes Maathuis, Associate Professor, ETH Zurich

Department Approval: ______________________________
Panagiotis Tsakalides, Professor and Department Chair

Heraklion, February 2015

# Abstract

Scientific practice typically involves repeatedly studying a system, each time trying to unravel a different perspective. In each study, the scientist may take measurements under different experimental conditions (interventions, manipulations, perturbations) and measure different sets of quantities (variables). The result is a collection of heterogeneous data sets coming from different data distributions. These data sets are analyzed in isolation and results are manually synthesized by the scientific community into scientific knowledge.

This thesis argues that heterogeneous data sets measuring the same system under study must all stem from, and therefore reflect, the same underlying causal mechanism, and that they can be co-analysed based on this premise. We define the problem of identifying one or all causal models that best fit all available data sets. We call this approach Integrative Causal Analysis.

The standard assumptions of causal modelling connect the statistical properties entailed in the available data sets to the underlying causal mechanism. Particularly, multivariate statistical relations of the measured variables constrain the search space of possible underlying causal models. Thus, the problem can be recast as a constraint satisfaction problem.

We propose an efficient conversion that translates statistical constraints into a SAT instance that can be solved with state-of-the-art SAT solvers. To improve scalability of our method we employ a series of approximate or exact steps that restrict the complexity of the conversion. Additionally, we introduce a scalable method for resolving conflicts arising from statistical errors. Finally, we identify a minimal example where INCA can produce a non-trivial prediction. We then test this prediction extensively in public data sets from a wide range of scientific domains, in an attempt to test whether causally-inspired predictions are verified.

We test our methods in a variety of different data sets and conditions. Results indicate that (a) our methods are robust and behave reasonably against different input parameters (b) our methods outperform state-of-the-art alternatives and (c) while causal assumptions cannot be easily verified, they lead to statistical predictions that are largely validated in real-world data sets.

# Περίληψη

Η επαναλαμβανόμενη μελέτη ενός συστήματος υπό διαφορετικές οπτικές για την εξαγωγή ενός συμπεράσματος είναι συχνό φαινόμενο στην επιστημονική πρακτική. Σε κάθε μελέτη, ο επιστήμονας συχνά μετρά διαφορετικές παραμέτρους του ίδιου συστήματος σε διαφορετικές πειραματικές συνθήκες. Το αποτέλεσμα μίας τέτοιας διαδικασίας είναι ένα σύνολο από ετερογενή σύνολα δεδομένων, που προέρχονται από διαφορετικές κατανομές. Κάθε σύνολο δεδομένων αναλύεται αυτοτελώς, και τα αποτελέσματα των αναλύσεων συντίθενται σε επιστημονική γνώση από την επιστημονική κοινότητα.

Παρ΄ όλη την ετερογένεια, σύνολα δεδομένων που μετρούν παραμέτρους του ίδιου συστήματος θα πρέπει να προέρχονται από, και άρα να αποτυπώνουν, τον ίδιο αιτιακό μηχανισμό. Υποστηρίζουμε ότι τέτοια σύνολα δεδομένων μπορούν να αναλυθούν μαζί βάσει αυτής της αρχής. Στη διατριβή αυτή, ορίζουμε και προτείνουμε μία λύση για το πρόβλημα του προσδιορισμού ενός ή όλων των πιθανών αιτιακών μηχανισμών που ταιριάζουν σε όλα τα διαθέσιμα σύνολα δεδομένων ενός συστήματος. Ονομάζουμε αυτή την προσέγγιση ολοκληρωμένη αιτιακή ανάλυση.

Χρησιμοποιούμε τη γνωστή θεωρία της αιτιακής μοντελοποίησης, που συνδέει τις στατιστικές ιδιότητες ενός συνόλου δεδομένων με τον αιτιακό μηχανισμό που περιγράφει τις μετρούμενες μεταβλητές στο σύνολο αυτό. Πιο συγκεκριμένα, οι πολυπαραγοντικές σχέσεις των μετρούμενων μεταβλητών απότελούν περιορισμούς για τους πιθανούς αιτιακούς μηχανισμούς. Με αυτό τον τρόπο, το πρόβλημα μπορεί να διατυπωθεί σαν ένα πρόβλημα ικανοποίησης περιορισμών.

Η μεθοδος που προτείνουμε μεταφράζει τους στατιστικούς περιορισμούς που προκύπρουν από τα δεδομένα σε λογικές προτάσεις, μετατρέποντας το πρόβλημα έυρεσης πιθανού αιτιακού μηχανισμού σε ένα πρόβλημα ικανοποιησιμότητας (SAT). Περιορίζουμε την πολυπλοκότητα της μεθόδου με μία σειρά από ευριστικές ή ακριβείς βελτιώσεις. Εφόσον οι λογικές προτάσεις αντιστοιχούν σε στατιστικές σχέσεις, πιθανά αιτιακά σφάλματα οδηγούν σε μη ικανοποιήσιμες λογικές προτάσεις. Προτείνουμε μία μέθοδο για την αντιμετώπιση αυτού του προβλήματος που δεν επιβαρύνει την πολυπλοκότητα του αλγορίθμου. Τέλος, ταυτοποιούμε μία περίπτωση που η ολοκληρωμένη αιτιακή ανάλυση οδηγεί σε μία μη προφανή πρόβλεψη. Ελέγχουμε την ισχύ της πρόβλεψης αυτής σε μία ευρεία γκάμα δημόσιων δεδομένων, με στόχο να ελέγξουμε την επαληθευσιμότητα των υποθέσεων της αιτιακής μοντελοποίησης.

Δοκιμάσαμε τις μεθόδους μας σε μία πληθώρα διαφορέτικών συνθηκών και συνόλων δεδομένων. Τα αποτελέσματα δείχνουν ότι (α) οι μέθοδοί μας έχουν την αναμενόμενη συμπεριφορά για διάφορες

παραμέτρους εισόδου (β) οι μέθοδοί μας ξεπερνοούν σε απόδοση τις σύγχρονες εναλλακτικές μεθόδους και (γ) αν και οι αιτιακές υποθέσεις δεν μπορούν να επαληθευτούν εύκολα, οδηγούν σε προβλέψεις που επαληθεύονται μαζικά σε πραγματικά σύνολα δεδομένων.

# Acknowledgements

This thesis would not have been possible without the constant and enduring support of my advisor, Ioannis Tsamardinos. His mentorship has had a tremendous influence on my scientific identity and my work ethic. He has really been the best advisor, even when I was a difficult student.

I would like to thank Takis Benos and Dimitris Plexousakis for our collaboration and their guidance throughout my PhD. I would also like to thank the members of my examination committee: Vassilis Christophides, Clark Glymour, Greg Cooper and Marloes Maathuis, for their constructive comments and advise.

This work was supported by a graduate scholarship from the Institute of Computer Science (ICS) in the Foundation for Research and Technology Hellas (FORTH), as well as the STATegra EU FP7 project, No 306000.

I would like to thank all my colleagues in the bioinformatics group in ICS-FORTH, who have really improved the PhD experience. I would particularly like to thank Giorgos Borboudakis for being so talented and collaborative. Finally, special thanks go to Vincenzo Lagani, with whom I was fortunate enough to collaborate in many projects. Working with him is always constructive and fun.

I would like to thank Vassilis Papadourakis for proofreading all my texts, rehearsing all my presentations, and for making everything seem very easy. Finally, I would like to thank my mother and sister, for being so understanding during my PhD, and my dad, whose absence made grad school seem trivial by comparison.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

*Chapter 1*

# Introduction

*The abundance of studies measuring overlapping sets of variables under different experimental conditions calls for novel methods of integrative analysis. We argue that all available data sets can be co-analyzed based on a simple principle: that they should all reflect the same underlying causal model. In this introduction, we formalize the problem of integrative causal analysis of heterogeneous data sets measuring the same system under study, present a motivating scenario, and discuss the core assumptions of causal analysis: The causal Markov condition (CMC) and the causal faithfulness condition (CFC).*

Causal discovery is an abiding goal in almost every scientific field. While the concept of causality has been the subject of a long philosophical debate, reasoning and acting based on perceived causes and effects is common practice both in science and in everyday life: people *quit* smoking to *reduce* their risk of having a heart attack. This significant attribute of causality, the ability not only to describe a system but also predict the responses to possible changes, is what makes causal discovery so appealing for scientists in every discipline.

In order to discover the causal mechanisms of a system, scientists typically have to perform a series of experiments (interchangeably: manipulations, interventions, or perturbations). Each experiment adds to the existing knowledge of the system and sheds light to the sought-after mechanism from a different perspective. In addition, each measurement may include a different set of quantities (variables), when for example the technology used allows only a limited number of measured quantities.

Certain assumptions that allow causal modelling link a data set measuring a set of quantities to the causal structure of said quantities, even if the data set is observational. This thesis is concerned with (a) understanding how these causal assumptions can be used to connect data sets measuring overlapping quantities and different perturbations of same system under study to the underlying causal structure (b) using these connections to develop methods that integratively analyze these data sets and infer causal features and (c) evaluating the predictions of the proposed causal discovery methods.

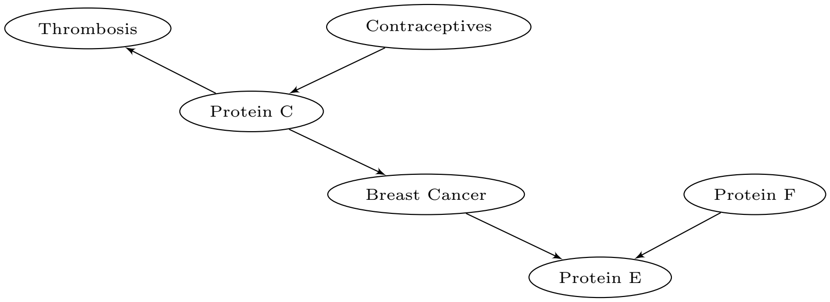


Figure 1.1: **Motivating Example**. Hypothetical causal structure of some variables relevant to breast cancer. Edges X → Y denote a probabilistic causal relationship between $X$ and $Y$, direct in the context of modeled variables.

## 1.1 Motivation

Assume that Figure 1.1 describes the causal relations of some variables related to breast cancer, where an arrow $A \rightarrow B$ denotes *direct probabilistic causality*: direct in the sense that nothing mediates this causal relation in the context of the variables present in the graph, and probabilistic in the sense that if you *intervene* and *change* the distribution of $A$, then the distribution of $B$ will also change. Several studies attempting to identify risk factors for breast cancer may involve overlapping subsets of the variables in the causal mechanism, or even intervene in this mechanism and assign values to the distributions in order to identify causal relations. Imagine, for example, the following (fictional) scenario (henceforth called the **motivating scenario**):

- Study 1: A scientist studies the relationship between contraceptives and breast cancer. Therefore, in a random sample of women, they measure whether a woman suffers from *Thrombosis*, uses *Contraceptives*, the concentration of *Protein C* in their blood and whether they have developed *Breast Cancer* by the age of 60.

- Study 2: Another researcher wants to identify the relation of some proteins in the blood, namely *Protein E* and *Protein F*, to breast cancer. They therefore measure these three variables in a random sample of women.

- Study 3: A scientist wants to check whether lowering the levels of a protein in the blood, namely *Protein C*, can prevent breast cancer. The scientist randomly samples women from the population, and then randomly assigns them into two groups: The first group is injected with high levels of the protein in their blood, while the latter is injected with enzymes that dissolve only the specific protein, effectively removing it from the blood. The scientist also measures the levels of *Protein E* in the subjects' blood, and reports whether the subjects take contraceptives and/or have thrombosis.

- Study 4: To establish whether *Thrombosis* is an effects of using contraceptives, a scientist takes a random sample of women and randomly assigns them into two groups: The first group use *Contraceptives* while the second do not. The scientist reports whether the subjects have developed *Thrombosis* a certain period.

| Variables | Thrombosis | Contraceptives | Protein C | Breast Cancer | Protein E | Protein F |
|---|---|---|---|---|---|---|
| Study | (Yes/No) | (Yes/No) | (numerical) | (Yes/No) | (numerical) | (numerical) |
| Study 1 | Yes | No | 10.5 | Yes | - | - |
|  | No | Yes | 5.3 | No | - | - |
| observational | . . . | . . . | . . . | . . . | - | - |
| data | No | No | 0.01 | No | - | - |
|  | No | Yes | 3.7 | No | - | - |
| Study 2 | No | No | **0 (Ctrl)** | No | 3.4 |  |
|  | No | Yes | **0 (Ctrl)** | No | 2.2 |  |
| experimental | . . . | . . . | . . . | . . . | . . . |  |
| data | Yes | Yes | **10 (Treat)** | Yes | 7.1 |  |
|  | Yes | Yes | **10 (Treat)** | No | 8.9 |  |
| Study 3 |  |  |  | Yes | 3.3 | 7.6 |
|  |  |  |  | No | 3.8 | 7.8 |
| observational |  |  |  | . . . | . . . | . . . |
| data |  |  |  | No | 4.7 | 9.4 |
|  |  |  |  | No | 5.1 | 10.2 |
| Study 4 | No | **No (Ctrl)** |  |  |  |  |
|  | No | **No (Ctrl)** |  |  |  |  |
| experimental | . . . | . . . |  |  |  |  |
| data | Yes | **Yes (Treat)** |  |  |  |  |
|  | No | **Yes (Treat)** |  |  |  |  |

Table 1.1: **Tabular depiction of the data sets described in the motivating example.** Different colors represent different studies, while randomized values are denoted in bold fonts. While the four studies measure the same system, the data sets cannot be pulled together due to different experimental designs and overlapping variable sets.

Typically, these data sets are analyzed in isolation, and the results are manually synthesized by the scientific community. These data sets, however, despite the differences that prevent their co-analysis, share something in common: *they stem from, and must be consistent to, the same causal mechanism.* While the studies can not be trivially combined, they must somehow reflect the common underlying causal structure illustrated in Figure 1.1.

## 1.2 Modelling Causality

Let $\mathcal{G}$ be the graph presented in Figure 1.1. $\mathcal{G}$ fully describes the direct and indirect (probabilistic) *causal* relations among the set of measured quantities $\mathbf{V}$. Given a data set measuring $\mathbf{V}$, one can estimate the **Joint Probability Distribution (JPD)** $\mathcal{P}$ over $\mathbf{V}$. A set of causal assumptions connect the graph $\mathcal{G}$ with the JPD $\mathcal{P}$.

Before presenting the assumptions, let us review some basic graph terminology. A graph $\mathcal{G}$ is an ordered pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set whose elements are called nodes (or vertices), and $\mathbf{E}$ is a set of ordered pairs of nodes, called edges. If $X \longrightarrow Y$, node $X$ is a **parent** of node $Y$ and node $Y$ is a **child** of node $X$. A path is a sequence of distinct nodes $\langle V_0, V_1, \ldots, V_n \rangle$ s.t for $0 \leq i < n$, $V_i$ and $V_{i+1}$ are adjacent in $\mathcal{G}$. A path from $V_0$ to $V_n$ is **directed** if for $0 \leq i < n$, $V_i$ is a parent $V_{i+1}$. $A$ is called an **ancestor** of $Y$ and $Y$ is called a **descendant** of $X$ in $\mathcal{G}$ if $X = Y$ in $\mathcal{G}$ or there exists a directed path from $X$ to $Y$ in $\mathcal{G}$. $\mathbf{Pa}_{\mathcal{G}}(\mathbf{X}), \mathbf{Ch}_{\mathcal{G}}(\mathbf{X}), \mathbf{An}_{\mathcal{G}}(\mathbf{X}), \mathbf{Desc}_{\mathcal{G}}(\mathbf{X})$ are used to denote the set of parents, children, ancestors and descendants of nodes $\mathbf{X}$ in $\mathcal{G}$, respectively. A **directed cycle** in

$\mathcal{G}$ occurs when $X \to Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$. A directed graph that contains no directed cycles is called a **Directed Acyclic Graph (DAG)**.

Causal Bayesian networks which constitute the foundation of graphical causal models were developed based on the assumption of **Acyclicity**: i.e., the causal structure of a system is not allowed to have cycles (causal feedback loops). While this assumption seems very weak, one could argue that in cross-sectional studies, where all measurements are taken simultaneously, feedback loops are not an issue. The theory presented in this section describes causality in acyclic systems, therefore the causal graphs are assumed to be DAGs.

A DAG $\mathcal{G}$ is connected to the JPD $\mathcal{P}$ through two conditions, the **Causal Markov Condition (CMC)** and the **Faithfulness Condition (FC)**. The notation $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ is used to denote that variables in $\mathbf{X}$ are independent from variables in $\mathbf{Y}$ given $\mathbf{Z}$.

**Definition 1.2.1 (Causal Markov Condition)** *(Spirtes et al., 2000) Let $\mathcal{G}$ be a causal graph with node set $\mathbf{V}$ and $\mathcal{P}$ be a probability distribution over the nodes in $\mathbf{V}$ generated by the causal structure represented by $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{P}$ satisfy the Causal Markov Condition if and only if for every $W$ in $\mathbf{V}$, $W$ is independent of* $\mathbf{V} \setminus (\mathbf{Descendants}(W) \cup \mathbf{Parents}(W))$ *given* $\mathbf{Parents}$*(W).*

In essence, the Causal Markov Condition states that a variable is independent its non-effects given its direct causes. Under the Causal Markov Condition, and according to the chain rule of probability, the joint probability distribution for a set $\mathbf{V}$ admits the following factorization:

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V|\mathbf{Parents}(V)) \tag{1.1}$$

where $P(V|\mathbf{Parents}(V))$ denotes the probability of $\mathbf{V}$ given the (possibly empty) set of nodes that are direct causes (parents) of $\mathbf{V}$.

If causal DAG $\mathcal{G}$ and a JPD $\mathcal{P}$ satisfy the CMC then the tuple $\langle \mathcal{G}, \mathcal{P} \rangle$ are called a **Causal Bayesian Network**. For a given causal DAG, the CMC yields a set of independence relations. For example, for the graph of Figure 1.1, the CMC yields the following conditional independencies:

$$\begin{gathered}
Contraceptives \perp\!\!\!\perp \{Thrombosis, Protein\ F\} \\
Protein\ C \perp\!\!\!\perp Protein\ F \mid Contraceptives \\
Thrombosis \perp\!\!\!\perp \{Contraceptives,\ Breast\ Cancer,\ Protein\ E,\ Protein\ F\ \} \mid Protein\ C \\
Breast\ Cancer \perp\!\!\!\perp \{Protein\ F,\ Thrombosis,\ Contraceptives\} \mid Protein\ C \\
Protein\ E \perp\!\!\!\perp \{Protein\ C,\ Thrombosis,\ Contraceptives\} \mid \{Breast\ Cancer,\ ProteinF\} \\
Protein\ F \perp\!\!\!\perp \{Breast\ Cancer, Protein\ C, Thrombosis,\ Contraceptives\}
\end{gathered}$$

The CMC also *entails* some independencies, such as:

$$Protein\ F \perp\!\!\!\perp Protein\ C \mid \{Protein\ E,\ Breast\ Cancer\}$$

The **independence model** $\mathcal{J}(\mathcal{P})$ or simply $\mathcal{J}$ is defined as the set of all conditional independencies $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{Z}$ in the joint distribution of $\mathcal{P}$, where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Z}$ are disjoint sets of variables. The Causal Markov Condition states that, if $A$ is a non-descendant of $B$ in $\mathcal{G}$, then $A \perp\!\!\!\perp B \mid \mathbf{Parents}(A) \in \mathcal{J}$. In general a probability distribution $\mathcal{P}$ satisfying the Causal Markov Condition for a causal graph $\mathcal{G}$ may include other independence relations besides those entailed by the Causal Markov Condition. If this is not the case, and the independence model $\mathcal{J}$ includes all and only the conditional independencies entailed by the Causal Markov Condition, the distribution and the graph are said to be *faithful* to each other:

**Definition 1.2.2 (Faithfulness Condition)** *(Spirtes et al., 2000) Let $\mathcal{G}$ be a causal graph and $\mathcal{P}$ a probability distribution generated by $\mathcal{G}$. $\langle \mathcal{G}, \mathcal{P} \rangle$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in $\mathcal{P}$ is entailed by the Causal Markov Condition applied to $\mathcal{G}$.*

When $\langle \mathcal{G}, \mathcal{P} \rangle$ are faithful to each other, the independencies that hold in $\mathcal{P}$ are all and only those entailed by the Causal Markov Condition. However, some of the entailed independencies are not obvious by the CMC, such as the conditional independence *Protein F* $\perp\!\!\!\perp$ *Protein C* $\mid \{$*Protein E*, *Breast Cancer*$\}$ in any probability distribution faithful to the graph in Figure 1.1. Pearl (2000) proposed a graphical criterion, called ***d*-separation**, which can identify all and only the conditional independencies stemming from applying the Causal Markov Condition to a DAG.

The definition is based on the notion of special nodes in the graph, called **colliders**: In a DAG $\mathcal{G}$ a node $Y$ is a collider on undirected path $p$ if and only if there are two distinct edges on $p$ containing $Y$ as an endpoint and both are into $Y (X \rightarrow Y \leftarrow Z)$. Otherwise $Y$ is a non-collider on $p$. In graph $\mathcal{G}$, node $Y$ is an unshielded collider on $p$ if $Y$ is a collider on $p$, $V$ is adjacent to distinct nodes $X$ and $Z$ on $p$, and $X$ and $Z$ are not adjacent in $\mathcal{G}$. The notion of collider is strongly associated with a specific triple and path, i.e. the same node may be a collider in one path and a non-collider in another.(e.g. if $\mathcal{G} = (\{X, Y, Z, W\},\ \{X \rightarrow Y,\ Z \rightarrow Y,\ Y \rightarrow W\}$, $Y$ is a collider in the path from $X$ to $Z$ and a non-collider in the path from $X$ to $W$.

**Definition 1.2.3 (*d*-separation)** *(Pearl, 1988) For a directed acyclic graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, $X, Y, \mathbf{W} \in \mathbf{V}$, $X \cap Y = \emptyset$, $X, Y \notin W$, then $X$ and $Y$ are d-separated given $\mathbf{W}$ in $\mathcal{G}$ if and only if there exists no undirected path $p$ between $X$ and $Y$, such that*

1. *Every collider on $p$ has a descendant in $\mathbf{W}$ and*
2. *no other node on $p$ is in $\mathbf{W}$.*

*If such a path exists, $X$ and $Y$ are said to be **d-connected** in $\mathcal{G}$ given $\mathbf{W}$. Any such path $p$ is called a **d-connecting** path for $X$ to $Y$ given $\mathbf{W}$.*

This definition is not very intuitive, but is based on some simple ideas regarding the flow of information through causally related events. See Pearl (1988, 2000) and Spirtes et al. (2000) for some very good examples for this graphical criterion. In any case, *d*-separation is a powerful tool for identifying conditional independencies in faithful Bayesian networks directly from the graph, and is also very useful because of the following theorem:

**Theorem 1.2.1** *(Pearl, 1988) If JPD $\mathcal{P}$ over variables $\mathbf{V}$ is faithful to some DAG, then $\mathcal{P}$ is faithful to DAG $\mathcal{G}$ with node set $\mathbf{V}$ if and only if*

1. *for all nodes $X$, $Y$ in $\mathcal{G}$, $X$ and $Y$ are adjacent if and only if $X$ and $Y$ are dependent given every set of nodes in $\mathcal{G}$ that does not include $X$ or $Y$ and*
2. *for all nodes $X$, $Y$, $Z$ such that $< X, Y, Z >$ is an unshielded triple, $X \rightarrow Y \leftarrow Z$ is a subgraph of $\mathcal{G}$ if and only if $X$, $Z$ are dependent given every set containing $Y$ but not $X$ or $Z$.*

This theorem, also known as the **local Markov condition** for graph $\mathcal{G}$, constitutes the basic tool for **constraint-based** Bayesian network learning.

Condition 1 in Theorem 1.2.1 states *that every non-adjacency in the graph $\mathcal{G}$ corresponds to a conditional independence in the JPD $\mathcal{P}$.* Imagine that one has gathered a data set over a set of variables $\mathbf{V}$, and wishes to identify the causal Bayesian network faithful to the JPD $\mathcal{P}$ over $\mathbf{V}$. According to Theorem 1.2.1-1, if there exists a set of variables $\mathbf{Z} \in \mathcal{G}$, such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in $\mathcal{P}$, then $X$ and $Y$ are not adjacent in any DAG $\mathcal{G}$ faithful to $\mathcal{P}$. Even though $\mathcal{P}$ is not *known*, the power set $P(\mathbf{V})$ can be searched for sets $\mathbf{Z}$ for which $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ according to a proper test of conditional independence. Thus, *Condition 1 in Theorem 1.2.1 enables identifying the skeleton of a causal graph given a data set and an appropriate test of independence.*

Condition 2 in Theorem 1.2.1 *enables the orientation of some edges of a causal graph (it specifically identifies unshielded colliders).* Additional orientations may be possible, based on CMC, FC and acyclicity. However, not all orientations of a causal graph $\mathcal{G}$ are distinguishable by the independence model $J$ of a faithful distribution. The set of DAGs $\mathcal{G}$ that correspond, according to the criterion of $d$-separation, to the same conditional independencies are said to form a **Markov Equivalence class**, denoted $[\mathcal{G}]$. The graphs in a Markov Equivalence class share the same skeleton and the same unshielded colliders (Verma and Pearl, 1990).

There exist two types of algorithms for learning causal Bayesian networks from data: Constraint-based methods, that constraint the search space of possible graphs according to the constraints described in Theorem 1.2.1, and score-based methods that try to identify the network whose implied factorization of the JPD is closer to the data. A comparison of these two approaches can be found in Tsamardinos et al. (2006). This thesis focuses in constraint-based methods.

In general, algorithms for learning causal Bayesian networks return a representative of all graphs that are *statistically indistinguishable*, i.e. graphs that have the edges in $[\mathcal{G}]$ and the orientations that are common to all $\mathcal{G} \in [\mathcal{G}]$. Such graphs are called **acyclic partially directed graphs** or **PDAGs** for short.

Causal Bayesian networks can be used to model interventions. For the causal structure of Figure 1.1, Study 3 describes an intervention: A scientist wants to check whether lowering *Protein C* can prevent *Breast Cancer*, and therefore samples a group from the population, and then randomly

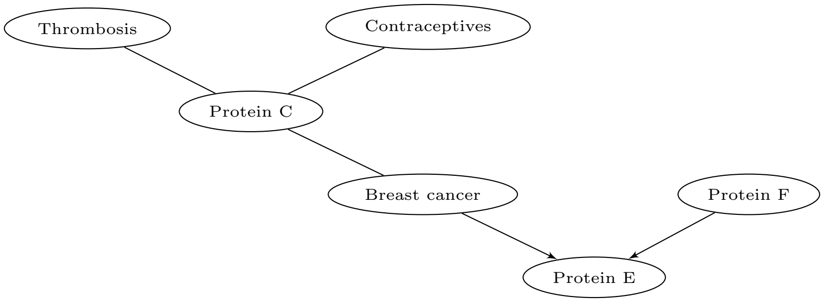


Figure 1.2: **PDAG corresponding to the equivalence class of the causal Bayesian network depicted in Figure 1.1**. Undirected edges correspond to edges that can be directed either way in different causal Bayesian networks in the Markov Equivalence class of the CBN in Figure 1.1. However, not all combinations of orientations are allowed: e.g., *Breast Cancer* and *Thrombosis* cannot be *simultaneously* into *Protein C*; this would make *Breast Cancer*→ *Protein C*← *Thrombosis* an unshielded collider, and the graph would encode a different independence model.

assigns women into two groups: The first group is injected with high levels of the protein in their blood, while the latter is injected with enzymes that dissolve only the specific protein, effectively removing it from the blood. What happens to the system under study is that the levels of *Protein C* in the graph are no longer causally dependent on the experimental design. Graphically this is equivalent to removing the edge from *Contraceptives* to *Protein C*. This procedure is called a **Randomized Control Trial (RCT)** and was first described by Peirce and Jastrow (1885).

The manipulation procedure just described is of course fictional, as is probably the existence of (ethically approved) medical procedures that can completely determine the presence of a protein in the human body. Such manipulations where the values of the manipulated variables are set solely by the experiment are called **hard manipulations**, or **ideal interventions**, and while they are typically neither possible or ethically permitted in most human studies, they are very common in molecular biology (e.g. gene knock-outs).

Such manipulations are modeled in the context of causal Bayesian networks with the removal of all edges that are incoming to the manipulated variables. This procedure is called **graph surgery** by Pearl (2000), while the resulting graph is called the **manipulated graph** in Spirtes et al. (2000). The distribution stemming from setting the values of manipulated variables is faithful to this graph. In the rest of this thesis, $\mathcal{G}^{\mathbf{I}}$ is used to denote the graph resulting from $\mathcal{G}$ after the manipulation of variables in **I**. (Pearl, 2000) has also developed the very elaborate theory of **do-caclulus** that can be used to identify post-intervention probabilities from observational causal Bayesian networks.

Predicting the effects of manipulations with the use of "graph surgery" requires a causal Bayesian network, i.e., a network where the edges denote a probabilistic direct causal relationship in the context of modeled variables. As mentioned above, some orientations in a network are not distinguishable from conditional independencies alone. Thus, when a network is learned from a data set using an asymptotically correct algorithm, some of the edges can not be oriented. For example, assuming no statistical errors, the PDAG learnt from a data set $\mathcal{D}$ measuring all variables in Figure 1.1 would be the one shown in Figure 1.2.

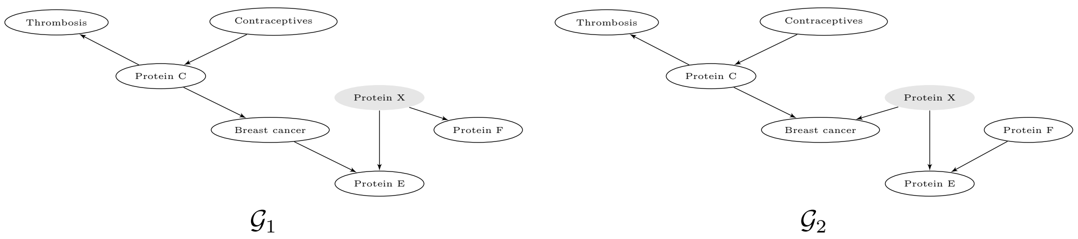


Figure 1.3: **Causal insufficiency in causal Bayesian networks.** Two different hypothetical causal structures for the variables measured in the motivating example. *Protein X* is latent (unmeasured). Graph $\mathcal{G}_1$ on the left entails the same conditional independencies as the graph in Figure 1.1, and therefore a learning algorithm would ideally learnt the PDAG in Figure 1.2, even though *Protein F* does not cause *Protein E*. No PDAG can correctly represent the conditional independencies entailed by $\mathcal{G}_2$.

Undirected edges in a PDAG are edges that can have either orientation in a faithful DAG (there exists at least one DAG faithful to the independence model where the edge is oriented as ⟶, and at least one DAG faithful to the independence model where the edge is oriented as ⟵). Thus, one cannot use undirected edges to predict the effect of interventions on adjacent nodes.

But even the directed edges in a PDAG have causal semantics that depend on an implicit, very delicate assumption: The assumption of causal sufficiency. According to the **Causal Sufficiency Condition (CSC)**, no two variables in the CBN share a latent common cause. Failure of causal sufficiency changes the semantics of the edges in a PDAG learnt from data (or any corresponding faithful DAG).

In the PDAG shown in Figure 1.2, only edges *Breast Cancer* → *Protein E* and *Protein F* → *Protein E* are directed. According to the semantics of causal Bayesian networks, this means that if one was to change the distribution of *Breast Cancer* in the population, the distribution of *Protein E* would change, as would happen if one intervened and changed the distribution of *Protein F*. However, this is not true unless the CSC holds.

For example, imagine that the true causal structure is the one shown in $\mathcal{G}_1$ in Figure 1.3 (left), where another protein, *Protein X*, causes both *Protein F* and *Protein E*. Imagine that this protein is not included in a study that measures the rest of the variables. The marginal probability distribution entails the same conditional independencies as the graph in Figure 1.1, and therefore (assuming no statistical errors) the PDAG one would learn from a this study is still the one depicted in 1.2.

Also notice that a similar scenario could not hold for the edge *Breast Cancer* → *Protein E*. imagine that the true causal structure is the one shown in $\mathcal{G}_2$ in Figure 1.3 (right). The presence of an unmeasured factor causing both *Breast Cancer* and *Protein E* would make *Breast Cancer* a collider on the path from *Protein C* to *Protein F*, requiring the orientation *Breast Cancer*⟶*Protein E*. However, since *Protein E* is also a collider on the path from *Breast Cancer* to *Protein F*, the edge must be oriented in the opposite direction: *Breast Cancer*⟵*Protein E*. Thus, there exists no DAG over the union of variables that can faithfully explain both entailed in the data.

What is described in the previous examples is the failure of causal Bayesian networks to describe situations where two of the modeled variables share a latent common cause (confounder): In some cases, there exists no faithful DAG representation of the independence model ($\mathcal{G}_2$ in Figure 1.3). In other cases ($\mathcal{G}_1$ in Figure 1.3), the entailed independence model can still be represented by a faithful causal graph, even when latent confounders are present. However, the *causal* semantics of the model cease to hold.

The causal assumptions made sofar were somewhat based on an intuitive understanding of the way causality works. Causal sufficiency, however, seems arbitrary and unrealistic for most systems. Particularly for the task of integratively analyzing data sets measuring overlapping variables, causal sufficiency can be assumed at most for the union of the overlapping variables, but certainly not for each individual data set.

## 1.3 Problem Specification and Proposed Approach

Section 1.2 presented the main assumptions that form the basis of causal discovery algorithms (at least in the context of causal graphical models), and how these assumptions are used to reverse engineer the causal structure of a set of variables, given a data set where these variables are observed.

We are interested in a more complicated problem. In the motivating scenario presented in 1.1, some aspects of the same underlying causal structure is measured in four different studies. How do CMC and FC connect the heterogeneous data sets to the underlying causal structure and how can one use them to reverse engineer the causal graph shown in Figure 1.1? *Is there a formal way to co-analyze heterogeneous studies measuring the same system under study and infer the common underlying causal structure?* This thesis *formalizes the problem* and *proposes a sound and complete algorithm* that solves it. To do so, we address the following issues:

First of all, to formalize the problem, one must *define in what ways the studies are allowed to differ.* In this thesis, we focus in two major sources of heterogeneity: Different data sets may be (a) measuring different (albeit overlapping) variable sets and (b) the data sets may be collected under different ideal interventions of some of the measured variables.

To causally co-analyse data sets with the above heterogeneities, we need causal models that allow latent confounders, and can model marginalization and interventions. Two extensions of causal Bayesian networks that allow latent confounders exist: **Maximal Ancestral Graphs (MAGs)** and **Semi-Markov Causal Models (SMCMs)**. While both causal models have been introduced for the same purpose, the correspondence between the two had not been thoroughly studied. Chapter 2 *presents the theory of both models and studies the relationship between them.*

We then propose the solving the problem by converting it into a constraint satisfaction problem. Specifically, we *introduce an efficient translation of statistical constraints entailed in the available data sets into path constraints over the possible causal graphs, and a subsequent conversion of these path constraints into SAT formulae.* The resulting SAT instance corresponds to causal graphs that fit all data sets. Apart from being able to solve the previously unresolved problem of integrative causal analysis of heterogeneous data sets, this approach has three major advantages over existing causal discovery approaches: (a) soundness and completeness is guaranteed by the inclusion of all and only the entailed constraints (b) it can be easily extended to incorporate other types of knowledge and/or experiments, as long as they can be expressed as graph properties and (c) the resulting

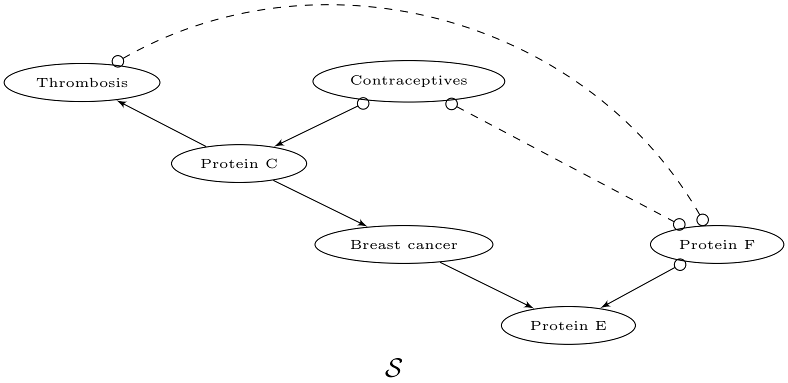


Figure 1.4: **Result of integrative causal analysis for the motivating example**. Solid lines denote edges that must be in the underlying causal structure, while dashed lines denote that the edge may or may not be in the underlying causal structure, i.e., there exist causal graphs with both configurations that are consistent with all the available studies. Circle endpoints represent ambiguity in orientation (the endpoint can be either a tail or an arrow in some consistent causal model). Notice that INCA is able to predict that no dashed or solid edge connects *Protein C* and *Protein F*, even though the two have not been measured together in any of the four studies!

instances can be solved with state-of-the-art SAT solvers, thus exploiting years of research in this field. This approach, along with some improvements aiming to tackle the (exponential) complexity of the problem, is presented in Chapter 3. This chapter introduces Algorithm COmbINE, a theoretically sound and complete algorithm that, in the absence of statistical errors, can learn causal structure from a collection of data sets that are heterogeneous in the manner stated above; The behaviour of the algorithm is shown to outperform the only similar state-of-the art literature.

Graphical (path) constraints imposed on the causal graph correspond to statistical facts entailed in the available data sets, and are obtained by the data sets using appropriate tests of independence. As a result, some erroneous constraints are expected. These constraints will typically lead to conflicts that make the SAT instance unsatisfiable. To be able to deal with conflicting constraints, COmbINE must be equipped with a strategy for selecting the best subset of non-conflicting constraints. Unfortunately, only few algorithms exist for evaluating the confidence of individual constraints, and they are computationally very expensive. In Chapter 4 we *introduce a method for estimating posterior pairwise probabilities for graph features.* The method has no significant computational overhead and is shown to perform on par, if not better, than its –computationally expensive– alternatives.

Chapter 5 shows how the aforementioned method can be used as a conflict resolution in algorithm COmbINE. The algorithm is then evaluated in several scenarios in simulated data where (a) the conflict resolution strategy is shown to outperform state-of-the-art alternatives (a) COmbINE is shown to be robust and behave reasonably against different input parameters. Finally, to showcase the availability of real problems that call for integrative causal analysis, the algorithm is employed to co-analyze a collection of public mass cytometry data sets measuring overlapping variables under three different interventions.

In principle, *integrative causal analysis of multiple, heterogeneous allows more sound causal inferences than the isolated analysis of each data set.* Graph $\mathcal{S}$, shown in Figure 1.4 shows the inferred causal structure when all four studies of the motivating example are analyzed together with COmbINE. Solid lines denote edges that must be in the underlying causal structure, while dashed lines denote that the edge may or may not be in the underlying causal structure, i.e., there exist causal graphs with both configurations that are consistent with all the available studies. Circle endpoints represent ambiguity in orientation (the endpoint can be either a tail or an arrow in some consistent causal model). Notice that no dashed or solid edge connects *Protein C* and *Protein F* in $\mathcal{S}$: Integrative causal analysis is able to predict that the two variables are independent, even though they have not been measured together in any of the four studies!

A similar inference regards the induction of the presence of an edge between two variables never measured together. This observation has significant ramifications for data analysis as it implies that additional causal relations may be inferred from already available data sets, without further studies. Moreover, *it offers a unique opportunity to test whether INCA predictions hold in practice.* Chapter 6, discusses a minimal example where such a prediction is possible: For a pair of specific input marginal structure (or independence models), each measuring two common variables ($X$ and $W$) and one distinct ($Y$ or $Z$), COmbINE predicts that $Y$ and $Z$ must be statistically dependent. This inference is based solely on the assumptions of Markov condition and faithfulness. The rule is evaluated in 20 data sets that cover a wide range of sample-sizes, dimensionality, types of variables, and domains. To the best of our knowledge, *this is the first attempt of massively validating causally-inspired assumptions in real data sets.* Computational experiments provide evidence that causally-inspired assumptions often hold to a good degree in many real systems and could be exploited for statistical inference.

*Chapter 2*

# Modelling Causally Insufficient Systems with Mixed Causal Models

*To link the available data sets to the underlying causal mechanism, we need to formalize causality. In causally insufficient systems, where latent confounders are possible, the most common causal models are Semi Markov causal modes (SMCMs) and Maximal Ancestral Graphs (MAGs). This chapter discusses the semantics of these models, their relationship to the joint probability distributions of the modelled variables based on the principles of causality, how each can be used to model experiments, and the correspondence between them.*

Causally insufficient systems are often described using Semi-Markov causal models (SMCMs) (Tian and Pearl, 2003) or Maximal Ancestral Graphs (MAGs) (Richardson and Spirtes, 2002; Richardson, 2003). Both of them are **mixed graphs**, meaning they can contain both directed ($\longrightarrow$) and bi-directed ($\longleftrightarrow$) edges. The term **mixed causal graph** is used here to denote both. In this section, their common and unique properties are briefly presented. First, a review of the basic mixed graph notation:

In a mixed graph $\mathcal{G}$, a path is a sequence of distinct nodes $\langle V_0, V_1, \ldots, V_n \rangle$ s.t for $0 \leq i < n$, $V_i$ and $V_{i+1}$ are adjacent in $\mathcal{G}$. $X$ is called a **parent** of $Y$ and $Y$ a **child** of $X$ in $\mathcal{G}$ if $X \longrightarrow Y$ in $\mathcal{G}$. A path from $V_0$ to $V_n$ is **directed** if for $0 \leq i < n$, $V_i$ is a parent $V_{i+1}$. $X$ is called an **ancestor** of $Y$ and $Y$ is called a **descendant** of $X$ in $\mathcal{G}$ if $X = Y$ in $\mathcal{G}$ or there exists a directed path from $X$ to $Y$ in $\mathcal{G}$. Notation $\mathbf{Pa}_{\mathcal{G}}(\mathbf{X}), \mathbf{Ch}_{\mathcal{G}}(\mathbf{X}), \mathbf{An}_{\mathcal{G}}(\mathbf{X}), \mathbf{Desc}_{\mathcal{G}}(\mathbf{X})$ is used to denote the set of parents, children, ancestors and descendants of nodes $\mathbf{X}$ in $\mathcal{G}$. A **directed cycle** in $\mathcal{G}$ occurs when $X \rightarrow Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$. An **almost directed cycle** in $\mathcal{G}$ occurs when $X \leftrightarrow Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$. Given a path $p$ in a mixed graph, a non-endpoint node $V$ on $p$ is called a **collider** if the two edges incident to $V$ on $p$ are both into $V$. Otherwise $V$ is called a **non-collider**. A path $p = \langle X, Y, Z \rangle$, where $X$ and $Z$ are not adjacent in $\mathcal{G}$ is called an **unshielded triple**. If $Z$ is a collider on this path, the triple is called an **unshielded collider**.

MAGs and SMCMs are graphical models that represent both causal relations and conditional independencies among a set of measured (observed) variables $\mathbf{O}$, and can be viewed as generalizations of causal Bayesian networks that can account for latent confounders. MAGs can also account for selection bias, but in this work selection bias is assumed not present.

## 2.1 Semi-Markov Causal Models

Semi-Markov causal models (SMCMs), introduced by Tian and Pearl (2003), often also reported as Acyclic Directed Mixed Graphs (ADMGs), are causal models that implicitly model hidden confounders using bi-directed edges. A directed edge $X \longrightarrow Y$ denotes that $X$ is a *direct* cause of $Y$ in the context of the variables included in the model. A bi-directed edge $X \longleftrightarrow Y$ denotes that $X$ and $Y$ are confounded by an unobserved variable. Two variables can be joined by at most two edges, one directed and one bi-directed.

Semi-Markov causal models are designed to represent marginals of causal Bayesian networks. In DAGs, the probabilistic properties of the distribution of variables included in the model can be determined graphically using the criterion of $d$-separation. The natural extension of d-separation to mixed causal graphs is called $m$-separation:

**Definition 2.1.1 ($m$-connection, $m$-separation.)** *In a mixed graph $\mathcal{G} = (\boldsymbol{E}, \boldsymbol{V})$, a path p between A and B is **m-connecting** given (conditioned on) a set of nodes $\boldsymbol{Z}$ , $\boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{A, B\}$ if*

1. *Every non-collider on p is not a member of $\boldsymbol{Z}$.*
2. *Every collider on the path is an ancestor of some member of $\mathbf{Z}$.*

*A and B are said to be m-**separated** by $\mathbf{Z}$ if there is no m-connecting path between A and B relative to $\mathbf{Z}$. Otherwise, they are said to be m-**connected** given $\mathbf{Z}$. The notation $\mathcal{J}_m(\mathcal{G})$ is used to denote the set of m-separations that hold in $\mathcal{G}$.*

Let $\mathcal{G}$ be a SMCM over a set of variables $\mathbf{O}$, $\Pi$ the joint probability distribution (JPD) over the same set of variables and $\mathcal{J}(\Pi)$ the independence model, defined as the set of conditional independencies that hold in $\Pi$. $\langle \mathbf{X}, \mathbf{Y} | \mathbf{Z} \rangle$ is used to denote the conditional independence of variables in $\mathbf{X}$ with variables in $\mathbf{Y}$ given variables in $\mathbf{Z}$. $\mathcal{J}_m(\mathcal{G})$ is used to denote the set of $m$-separations that hold in $\mathcal{G}$. Under the Causal Markov (**CMC**) and Faithfulness (**FC**) conditions (Spirtes et al., 2000), *every m-separation present in $\mathcal{G}$ corresponds to a conditional independence in $\mathcal{J}(\Pi)$ and vice-versa: $\mathcal{J}_m(\mathcal{G}) = \mathcal{J}(\Pi)$.*

In causal Bayesian networks, every missing edge in $\mathcal{G}$ corresponds to a conditional independence in $\mathcal{J}(\Pi)$ (resp. an $m$-separation in $\mathcal{G}$), meaning there exists a subset of the variables in the model that renders the two non-adjacent variables independent. Respectively, every conditional independence in $\mathcal{J}(\Pi)$ corresponds to a missing edge in the DAG $\mathcal{G}$. This is not always true for SMCMs. Figure 2.1 illustrates an example of a SMCM where two non-adjacent variables are not independent given any subset of observed variables.

Evans and Richardson (2010, 2011) deal with the factorization and parametrization of SMCMs for discrete variables. Based on this parametrization, score-based methods have also recently been explored (Richardson et al., 2012; Shpitser et al., 2013), but are still limited to small sets of discrete variables. The skeleton of a SMCM is not uniquely identifiable by the corresponding conditional independence model on the same variables (see Figure 2.1 for an example). Richardson and Spirtes (2002) overcome this obstacle by introducing a causal mixed graph with slightly different semantics, the maximal ancestral graph.

## 2.2 Maximal Ancestral Graphs

Maximal ancestral graphs (MAGs) (Richardson and Spirtes, 2002), are **ancestral** mixed graphs, meaning that they contain no directed or almost directed cycles, where an almost directed cycle occurs if $X \leftrightarrow Y$ and $X$ causes $Y$. Every pair of variables $X$, $Y$ in an ancestral graph is joined by at most one edge. The orientation of this edge represents (non) causal ancestry: A bi-directed edge $X \leftrightarrow Y$ denotes that $X$ does not cause $Y$ and $Y$ does not cause $X$, but (under the faithfulness assumption) the two share a latent confounder. A directed edge $X \rightarrow Y$ denotes causal ancestry: $X$ is a *causal ancestor* of $Y$. Thus, if $X$ causes $Y$ (not necessarily directly in the context of observed variables) and they are also confounded, there is an edge $X \rightarrow Y$ in the corresponding MAG. Undirected edges can also be present in MAGs that account for selection bias. As mentioned above, no selection bias is assumed in this work ; the theory of MAGs presented here is restricted to MAGs with no undirected edges.

Like SMCMs, ancestral graphs are also designed to represent marginals of causal Bayesian networks. Thus, under the causal Markov and faithfulness conditions for a MAG $\mathcal{M}$ and a JPD $\Pi$, $X$ and $Y$ are $m$-separated given $\mathbf{Z}$ in an ancestral graph $\mathcal{M}$ if and only if $\langle X, Y|\mathbf{Z}\rangle$ is in the corresponding independence model $\mathcal{J}(\Pi)$. Still, like in SMCMs, a missing edge does not necessarily correspond to a conditional independence. The following definition describes a subset of ancestral graphs in which every missing edge (non-adjacency) corresponds to a conditional independence:

**Definition 2.2.1 (Maximal Ancestral Graph, MAG)** *A mixed graph is called* ancestral *if it contains no directed and almost directed cycles. An ancestral graph $\mathcal{G}$ is called* maximal *if for every pair of non-adjacent nodes $(X, Y)$, there is a (possibly empty) set $\mathbf{Z}$, $X, Y \notin \mathbf{Z}$ such that $\langle X, Y|\mathbf{Z}\rangle \in \mathcal{J}_m(\mathcal{G})$.*

Figure 2.1 illustrates an ancestral graph that is not maximal, and the corresponding maximal ancestral graph. MAGs are closed under marginalization (Richardson and Spirtes, 2002). Thus, if $\mathcal{G}$ is a MAG faithful to $\Pi$, then there is a unique MAG $\mathcal{G}'$ faithful to any marginal distribution of $\Pi$.

$[_{\mathbf{L}}$ is used to denote the act of marginalizing out variables $\mathbf{L}$, thus, if $\mathcal{G}$ is a MAG over variables $\mathbf{O} \cup \mathbf{L}$ faithful to a joint probability distribution $\Pi$, $\mathcal{G}[_{\mathbf{L}}$ is the MAG over $\mathbf{O}$ faithful to the marginal joint probability distribution of $\Pi$. $\mathcal{J}[_{\mathbf{L}}$ is used to denote the *marginal independence model* of $\mathcal{J}$, i.e. the set of conditional independencies $\{X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathcal{J} : (X \cup Y \cup \mathbf{Z}) \cap \mathbf{L} = \emptyset\}$. Obviously, the DAG of a causal Bayesian network is also a MAG. For a MAG $\mathcal{G}$ over $\mathbf{O}$ and a set of variables $\mathbf{L} \subset \mathbf{O}$, the marginal MAG $\mathcal{G}[_{\mathbf{L}}$ is defined as follows:

**Definition 2.2.2 (Marginal MAG)** *(Richardson and Spirtes, 2002) MAG $\mathcal{G}[_{\mathbf{L}}$ has node set $\mathbf{O} \setminus \mathbf{L}$ and edges specified as follows: If $X$, $Y$ are s.t. $\forall \mathbf{Z} \subset \mathbf{O} \setminus \mathbf{L} \cup \{X, Y\}$, $X$ and $Y$ are $m$-connected*

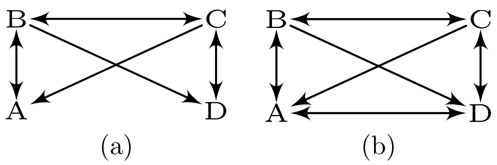


Figure 2.1: **Maximality and primitive inducing paths.**(a) Both (i) a semi Markov causal model over variables $\{A,\ B,\ C,\ D\}$; variables $A$ and $D$ are $m$-connected given any subset of observed variables, but they do not share a direct relationship in the context of observed variables and (ii) a non-maximal ancestral graph over variables $\{A,\ B,\ C,\ D\}$. (b) The corresponding MAG. $A$ and $D$ are adjacent, since they cannot be $m$-separated given any subset of $\{B, C\}$. Path $\langle A, B, C, D\rangle$ is a primitive inducing path. This example was presented in Zhang (2008b).

*given* $\mathbf{Z}$ *in* $\mathcal{G}$*, then*

$$\text{if} \left\{ \begin{array}{c} X \notin \mathbf{An}_{\mathcal{G}}(Y); Y \notin \mathbf{An}_{\mathcal{G}}(X) \\ X \in \mathbf{An}_{\mathcal{G}}(Y); Y \notin \mathbf{An}_{\mathcal{G}}(X) \\ X \notin \mathbf{An}_{\mathcal{G}}(Y); Y \in \mathbf{An}_{\mathcal{G}}(X) \end{array} \right\} \text{then} \left\{ \begin{array}{c} X \leftrightarrow Y \\ X \rightarrow Y \\ X \leftarrow Y \end{array} \right\} \text{in } \mathcal{G}[_{\mathbf{L}}$$

The following theorem was proved in Richardson and Spirtes (2002):

**Theorem 2.2.1** *If* $\mathcal{G}$ *is a MAG over* $\mathbf{V} = \mathbf{O} \cup \mathbf{L}$*, then* $\mathcal{J}_m(\mathcal{G}[_{\mathbf{L}}) = \mathcal{J}_m(\mathcal{G})[_{\mathbf{L}}$.

**Proof** See proof of Theorem 4.18 in Richardson and Spirtes (2002).

As mentioned above, every conditional independence in an independence model $\mathcal{J}$ corresponds to a missing edge in the corresponding faithful MAG $\mathcal{G}$. Conversely, if $X$ and $Y$ are dependent given every subset of observed variables, then $X$ and $Y$ are adjacent in $\mathcal{G}$. Thus, given an oracle of conditional independence it is possible to learn the skeleton of a MAG $\mathcal{G}$ over variables $\mathbf{O}$ from a data set. Still, some of the orientations of $\mathcal{G}$ are not distinguishable by mere observations. The set of MAGs $\mathcal{G}$ faithful to distributions $\Pi$ that entail a set of conditional independencies $\mathcal{J}(\Pi)$ form a **Markov equivalence class**.

It is well known that two DAGs are Markov equivalent if and only if they share the same adjacencies and unshielded colliders. Markov equivalent MAGs also share adjacencies and unshielded colliders, but this is not sufficient to characterize Markov equivalent graphs. The emergence of bi-directed edges imposes also a set of shielded colliders on the Markov equivalent MAGs. These colliders are discriminated by *discriminating paths*:

**Definition 2.2.3 (Discriminating path)** *A path* $p = \langle X, \ldots, W, V, Y\rangle$ *is called* ***discriminating*** *for* $V$ *if* $X$ *is not adjacent to* $Y$ *and every node on the path from* $X$ *to* $V$ *is a collider and a parent of* $Y$.

Discriminating paths, their properties and their connection to Markov equivalence is discussed in detail in Ali et al. (2009). Unfortunately, two Markov equivalent MAGs may not share the same discriminating paths. Moreover, a triple may be discriminated to be a collider in MAG $\mathcal{M}_1$ but not in MAG $\mathcal{M}_2$ in the same Markov equivalence class. There exists however, a subset of discriminating paths that (a) are present in all the Markov equivalent MAGs and (b) the colliders discriminated by these paths are necessary and sufficient for Markov equivalence (Ali et al., 2009). The following definition from Ali et al. (2009) is relevant:

**Definition 2.2.4 (Colliders with order)** *Let $\mathfrak{D}_i, i \geq 0$ be a set of triples of order $i$ in MAG $\mathcal{M}$, defined recursively as follows:*

- *Order 0: A triple $\langle X, Y, Z\rangle \in \mathfrak{D}_0$ if $X$ and $Z$ are not adjacent.*
- *Order i: A triple $\langle X, Y, Z\rangle \in \mathfrak{D}_{i+1}$ if,*
    1. *for all $j < i+1, \langle X, Y, Z\rangle \notin \mathfrak{D}_j$ and*
    2. *There is a discriminating path $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ such that either $\langle X, Y, Z\rangle = \langle V_n, Y, Q\rangle$ or $\langle X, Y, Z\rangle = \langle Q, Y, V_n\rangle$ and the $n$ colliders:*

$$\langle W, V_1, V_2\rangle, \ldots, \langle V_{n-1}, V_n, Y\rangle \in \bigcup_{j \leq i} \mathfrak{D}_j$$

*If $\langle X, Y, Z\rangle \in \mathfrak{D}_i$, the triple has order $i$. If the triple has order $i$ for some $i$, then the triple is said to have order. If $\langle X, Y, Z\rangle$ is a triple with order and $X {*\!\!\rightarrow} Y {\leftarrow\!\!*} Z$ is in $\mathcal{M}$, then the triple is a **collider with order** $i$ in $\mathcal{M}$. Otherwise, the triple is a **definite non-collider with order** in $\mathcal{M}$. A discriminating path $p$ has order $i$ if all colliders on the path (except from the collider $\langle V_n, Y, Q\rangle$ discriminated by the path) have order at most $i-1$, and there exists at least one collider with order $i-1$. If a discriminating path has order $i$ for some $i$, then the discriminating path is said to have order. In this work (non) colliders with order $\geq 1$ are (abusively) called **discriminating (definite non) colliders**.*

Note that not every triple on a mixed graph has order. The order (if any) of a shielded triple is the minimum of the orders of all discriminating paths with order for that triple. Triples with order 0 are the unshielded triples. Discriminating paths with order $\geq 1$ are present in all Markov equivalent MAGs, and therefore colliders with order $\geq 1$ are the triples that are colliders in all the Markov equivalent MAGs. Colliders with order, along with adjacencies, are necessary and sufficient to characterize Markov equivalent MAGs:

**Theorem 2.2.2** *Two MAGs over the same variable set are Markov equivalent if and only if they share the same edges and the same colliders with order.*

**Proof** See proof of Theorem 3.7 in Ali et al. (2009).

$[\mathcal{G}]$ is used to denote the class of MAGs that are Markov equivalent to $\mathcal{G}$. A **partial ancestral graph (PAG)** is a representative graph of this class, and has the skeleton shared by all the graphs in $[\mathcal{G}]$, and all the orientations invariant in all the graphs in $[\mathcal{G}]$. Endpoints that can be either arrows or tails in different MAGs in $\mathcal{G}$ are denoted with a circle "$\circ$" in the representative PAG. The symbol $\star$ is used as a wildcard to denote any of the three marks. The notation $\mathcal{M} \in \mathcal{P}$ is used to denote that MAG $\mathcal{M}$ belongs to the Markov equivalence class represented by PAG $\mathcal{P}$.

For a $\mathcal{M}$ and a probability distribution $\Pi$ faithful to each other, $\mathcal{J}_m(\mathcal{M}) = \mathcal{J}(\Pi)$. Thus, the set of $m$-separations entailed in $\mathcal{M}$ are exactly the conditional independencies that hold in $\Pi$. **FCI** Algorithm (Spirtes et al., 2000; Zhang, 2008a) is a sound and complete algorithm for learning the complete (maximally informative) PAG of the MAGs faithful to a distribution $\Pi$ over variables $\mathbf{O}$ in which a set of conditional independencies $\mathcal{J}(\Pi)$ hold. An important advantage of FCI is that it employs CMC, faithfulness and some graph theory to reduce the number of tests required to identify the correct PAG.

## 2.3 Correspondence between SMCMs and MAGs

Semi Markov Causal Models and Maximal Ancestral Graphs both represent causally insufficient causal structures. They both entail the conditional independence structure and the causal ancestry structure of the observed variables. Thus, under CMC and FC, the SMCM $\mathcal{G}$ and the MAG $\mathcal{M}$ over a set of variables $\mathbf{O}$ entail the same independence model: $\mathcal{J}_m(\mathcal{S}) = \mathcal{J}_m(\mathcal{M})$. They also entail the same ancestral relationships: $X$ is an ancestor of $Y$ in $\mathcal{S}$ if and only if $X$ is an ancestor of $Y$ in $\mathcal{M}$.

Nevertheless, SMCMs and MAGs also have significant differences: SMCMs describe the causal relations among observed variables, while MAGs encode independence structure with partial causal ordering. Edge semantics in SMCMs are closer to the semantics of causal Bayesian networks, whereas edge semantics in MAGs are more complicated. On the other hand, unlike in DAGs and MAGs, a missing edge in a SMCM does not necessarily correspond to a conditional independence (SMCMs do not obey a pairwise Markov property).

Figure 2.2 summarizes the main differences of SMCMs and MAGs. It shows two different DAGs, and the corresponding marginal SMCMs and MAGs over four observed variables. SMCMs have a many-to-one relationship to MAGs: For a MAG $\mathcal{M}$, there can exist more than one SMCMs that entail the same probabilistic and causal ancestry relations. On the other hand, for any given SMCM there exists only one MAG entailing the same probabilistic and causal ancestry relations. This is clear in Figure 2.2, where a unique MAG, $\mathcal{M}_1 = \mathcal{M}_2$ entails the same information as two different SMCMs, $\mathcal{S}_1$ and $\mathcal{S}_2$ in the same figure.

Directed edges in a SMCM denote a causal relation that is *direct* in the context of observed variables. In contrast, a directed edge in a MAG merely denotes causal ancestry; the causal relation is not necessarily direct. An edge $X \longrightarrow Y$ can be present in a MAG even though $X$ does not directly cause $Y$; this happens when $X$ is a causal ancestor of $Y$ and the two cannot be rendered independent given any subset of observed variables. Depending on the structure of latent variables, this edge can be either missing or bi-directed in the respective SMCM.

Figure 2.2 illustrates examples of both cases. For example, $A$ is a causal ancestor of $D$ in DAG $\mathcal{G}_1$, but not a direct cause (in the context of observed variables). Therefore, the two are not adjacent in

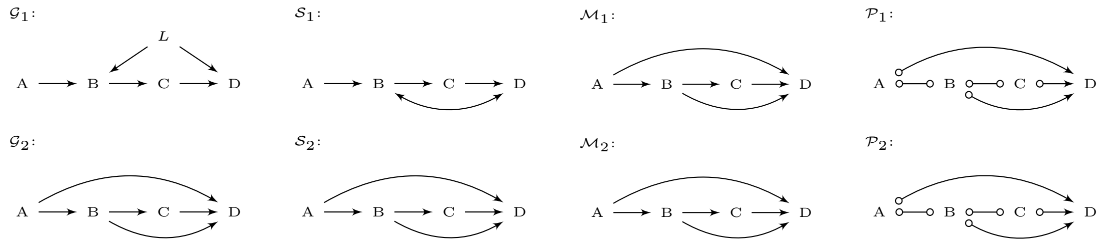


Figure 2.2: **An example two different DAGs and the corresponding mixed causal graphs over observed variables**. On the right: DAGs $\mathcal{G}_1$ over variables $\{A,\ B,\ C,\ D,\ L\}$ (top) and $\mathcal{G}_2$ over variables $\{A,\ B,\ C,\ D\}$ (bottom). From left to right, on the same row as the underlying causal DAG, the respective SMCMs $\mathcal{S}_1$ and $\mathcal{S}_2$ over $\{A,\ B,\ C,\ D\}$; the respective MAGs $\mathcal{M}_1 = \mathcal{G}_1[_L$ and $\mathcal{M}_2 = \mathcal{G}_2$ over variables $\{A,\ B,\ C,\ D\}$; finally, the respective PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$. Notice that, $\mathcal{M}_1$ and $\mathcal{M}_2$ are identical, despite representing different underlying causal structures.

the corresponding SMCM $\mathcal{S}_1$ over $\{A, B, C, D\}$. However, the two cannot be rendered independent given any subset of $\{B, C\}$, and therefore $A \longrightarrow D$ is in the respective MAG $\mathcal{M}_1$.

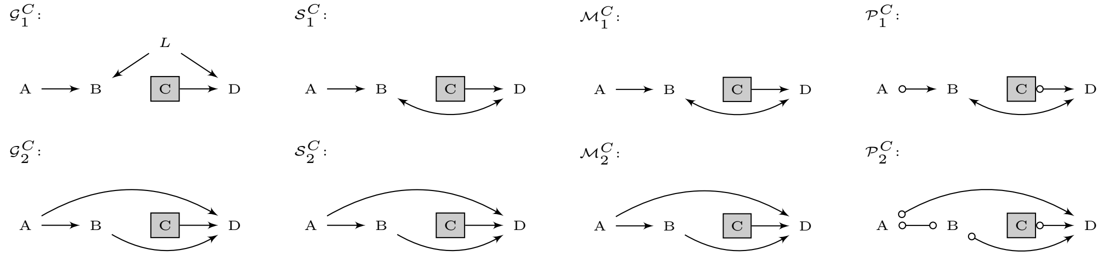


Figure 2.3: **Effect of manipulating variable $C$ on the causal graphs of Figure 2.2**. From right to left: the manipulated DAGs $\mathcal{G}_1^C$ (top) and $\mathcal{G}_2^C$ (bottom), the manipulated SMCMs $\mathcal{S}_1^C$ (top) and $\mathcal{S}_2^C$ (bottom) over variables $\{A,\ B,\ C,\ D\}$, the manipulated MAGs $\mathcal{M}_1^C = \mathcal{G}_1^C[_L$ (top) and $\mathcal{M}_2^C = \mathcal{G}_2^C$ (bottom) over the same set of variables, and the corresponding PAGs $\mathcal{P}_1^C$ (top) and $\mathcal{P}_2^C$ (bottom). Notice that edge $A \longrightarrow D$ is removed in $\mathcal{M}_1^C$, even though it is not adjacent to the manipulated variable. Moreover, on the same graph, edge $B \longrightarrow D$ is now $B \longleftrightarrow D$.

On the same DAG, $B$ is another causal ancestor (but not a direct cause) of $D$. The two variables share the common cause $L$. Thus, in the corresponding SMCM $\mathcal{S}_1$ over $\{A, B, C, D\}$ $B \longleftrightarrow D$ is present. However, a bi-directed edge between $B$ and $D$ is not allowed in MAG $\mathcal{M}_1$, since it would create an almost directed cycle. Thus, $B \longrightarrow D$ is in $\mathcal{M}_1$.

Also notice that, unlike SMCMs, MAGs only allow one edge per variable pair. Thus, if $X$ directly causes $Y$ and the two are also confounded, both edges will be in a relevant SMCM ($X \overset{\longleftrightarrow}{\longrightarrow} Y$), while the two will share a directed edge from $X$ to $Y$ in the corresponding MAG.

Overall, a SMCM has a subset of the adjacencies (but not necessarily edges) of its MAG counterpart. These extra adjacencies in MAGs correspond to pairs of variables that cannot be $m$-separated given any subset of observed variables, but neither directly causes the other, and the two are not confounded. These adjacencies can be checked in a SMCM using a special type of path, called **inducing path** (Richardson and Spirtes, 2002).

**Definition 2.3.1 (Inducing path)** *A path $p = \langle V_1, V_2, \ldots, V_n \rangle$ on a mixed causal graph $\mathcal{G}$ over a set of variables $\mathbf{V} = \mathbf{O}\dot{\cup}\mathbf{L}$ is called **inducing** with respect to $\mathbf{L}$ if every non-collider on the path is in $\mathbf{L}$ and every collider is an ancestor of either $V_1$ or $V_n$. A path that is inducing with respect to the empty set is called a **primitive** inducing path.*

Obviously, an edge joining $X$ and $Y$ is a primitive inducing path. Intuitively, an inducing path with respect to $\mathbf{L}$ is $m$-connecting given any subset of variables that does not include variables in $\mathbf{L}$. Path $A \longrightarrow B \longleftarrow L \longrightarrow D$ is an inducing path with respect to $L$ in $\mathcal{G}_1$ of Figure 2.2, and $A \longrightarrow B \longleftrightarrow D$ is an inducing path with respect to the empty set in $\mathcal{S}_1$ of the same figure. Inducing paths are extensively discussed in Richardson and Spirtes (2002), where the following theorem is proved:

**Theorem 2.3.1** *If $\mathcal{G}$ is an ancestral graph over variables $\mathbf{V} = \mathbf{O}\dot{\cup}\mathbf{L}$, and $X, Y \in \mathbf{O}$ then the following statements are equivalent:*

*i. $X$ and $Y$ are adjacent in $\mathcal{G}[_{\mathbf{L}}$.*

*ii. There is an inducing path with respect to $\mathbf{L}$ in $\mathcal{G}$.*

*iii. $\forall \mathbf{Z}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{L} \cup \{X, Y\}$, $X$ and $Y$ are $m$-connected given $\mathbf{Z}$ in $\mathcal{G}$.*

**Proof** See proof of Theorem 4.2 in Richardson and Spirtes (2002).

This theorem links inducing paths in an ancestral graph to $m$-separations in the same graph and to adjacencies in any marginal ancestral graph. The equivalence of (ii) and (iii) can also be proved for SMCMs, using the proof presented in Richardson and Spirtes (2002) for Theorem 2.3.1:

**Theorem 2.3.2** *If $\mathcal{G}$ is a SMCM over variables $\mathbf{V} = \mathbf{O}\dot{\cup}\mathbf{L}$, and $X, Y \in \mathbf{O}$ then the following statements are equivalent:*

*i. There is an inducing path with respect to $\mathbf{L}$ in $\mathcal{G}$.*

*ii. $\forall \mathbf{Z}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{L} \cup \{X, Y\}$, $X$ and $Y$ are $m$-connected given $\mathbf{Z}$ in $\mathcal{G}$.*

**Proof** See proof of Theorem 4.2 in Richardson and Spirtes (2002).

The following proposition follows from Theorems 2.3.1 and 2.3.2:

**Proposition 2.3.3** . *Let $\mathbf{O}$ be a set of variables and $\mathcal{J}$ the independence model over $\mathbf{O}$. Let $\mathcal{S}$ be a SMCM over variables $\mathbf{O}$ that is faithful to $\mathcal{J}$ and $\mathcal{M}$ be the MAG over the same variables that is faithful to $\mathcal{J}$. Let $X, Y \in \mathbf{O}$. Then there is an inducing path between $X$ and $Y$ with respect to $\mathbf{L}$, $\mathbf{L} \subseteq \mathbf{O}$ in $\mathcal{S}$ if and only if there is an inducing path between $X$ and $Y$ with respect to $\mathbf{L}$ in $\mathcal{M}$.*

**Proof** ($\Rightarrow$) Assume there exists a path $p$ in $\mathcal{S}$ that is inducing w.r.t. $\mathbf{L}$. Then by theorem 2.3.2 there exists no $\mathbf{Z} \subseteq \mathbf{O} \setminus \mathbf{L} \cup \{X, Y\}$ such that $X$ and $Y$ are $m$-separated given $\mathbf{Z}$ in $\mathcal{S}$, and since $\mathcal{S}$ and $\mathcal{M}$ entail the same $m$-separations there exists no $\mathbf{Z} \subseteq \mathbf{O} \setminus \mathbf{L} \cup \{X, Y\}$ such that $X$ and $Y$ are $m$-separated given $\mathbf{Z}$ in $\mathcal{M}$. Thus, by Theorem 2.3.1 there exists an inducing path between $X$ and $Y$ with respect to $\mathbf{L}$ in $\mathcal{M}$.
($\Leftarrow$) Similarly, assume there exists a path $p$ in $\mathcal{M}$ that is inducing w.r.t. $\mathbf{L}$. Then by theorem 2.3.1 there exists no $\mathbf{Z} \subseteq \mathbf{O} \setminus \mathbf{L} \cup \{X, Y\}$ such that $X$ and $Y$ are $m$-separated given $\mathbf{Z}$ in $\mathcal{M}$, and since $\mathcal{S}$ and $\mathcal{M}$ entail the same $m$-separations there exists no $\mathbf{Z} \subseteq \mathbf{O} \setminus \mathbf{L} \cup \{X, Y\}$ such that $X$ and $Y$ are $m$-separated given $\mathbf{Z}$ in $\mathcal{S}$. Thus, by Theorem 2.3.2 there exists an inducing path between $X$ and $Y$ with respect to $\mathbf{L}$ in $\mathcal{S}$.

Primitive inducing paths are connected to the notion of maximality in ancestral graphs: Every ancestral graph can be transformed into a maximal ancestral graph with the addition of a finite number of bi-directed edges. Such edges are added between variables $X, Y$ that are $m$-connected through a **primitive inducing path** (Richardson and Spirtes, 2002). Path $A \leftrightarrow B \leftrightarrow C \leftrightarrow D$ in Figure 2.1 is an example of a primitive inducing path.

Inducing paths are crucial in this work because adjacencies and non-adjacencies in marginal ancestral graphs can be translated into existence or absence of inducing paths in causal graphs that include some additional variables. For example, path $A \longrightarrow B \longleftarrow L \longrightarrow D$ is an inducing path w.r.t. $L$ in $\mathcal{G}_1$ in Figure 2.2, and therefore $A$ and $D$ are adjacent in $\mathcal{M}_1$. Thus, inducing paths are useful for combining causal mixed graphs over overlapping variables.

Inducing paths are also necessary to decide whether two variables in an SMCM will be adjacent in a MAG over the same variables without having to check all possible $m$-separations. Algorithm 1 describes how to turn a SMCM into a MAG over the same variables.

Algorithm 1 takes as input a SMCM $\mathcal{S}$ and adds the necessary edges to transform it into a MAG $\mathcal{M}$ by looking for primitive inducing paths. The procedure can be viewed as a special case of marginalizing out variables in DAGs, presented in Spirtes and Richardson (1996) and Zhang (2008b). Similar algorithms are also presented in Sadeghi (2012), where the relationship among different types of mixed causal graphs representing the same independence model is discussed in detail. The algorithm is sound, i.e. the output MAG shares the same causal ancestry relations and entails the same independence model as $\mathcal{S}$:

**Theorem 2.3.4** . *Let $\mathbf{O}$ be a set of variables and $\mathcal{J}$ the independence model over $\mathbf{O}$. Let $\mathcal{S}$ be a SMCM over variables $\mathbf{V}$ that is faithful to $\mathcal{J}$. Let $\mathcal{M}$ = SMCMtoMAG($\mathcal{S}$). Then $\mathcal{S}$ and $\mathcal{M}$ share the same ancestry relations and $\mathcal{J}_m(\mathcal{S}) = \mathcal{J}_m(\mathcal{M})$.*

**Proof** $\mathcal{S}$ and $\mathcal{M}$ share the same ancestry relations, since during Algorithm 1 a directed edge $X \longrightarrow Y$ is added only if $X$ is an ancestor of $Y$ in $\mathcal{S}$, and no directed edges are removed. To prove that the

**Algorithm 1:** SMCMtoMAG

```
input  : SMCM $\mathcal{S}$
output: MAG $\mathcal{M}$
$\mathcal{M} \leftarrow \mathcal{S}$;
foreach ordered pair of variables $X$, $Y$ not adjacent in $\mathcal{S}$ do
    if $\exists$ primitive inducing path from $X$ to $Y$ in $\mathcal{S}$ then
        if $X \in \mathbf{An}_{\mathcal{S}}(Y)$ then
            add $X \longrightarrow Y$ to $\mathcal{M}$;
        else if $Y \in \mathbf{An}_{\mathcal{S}}(X)$ then
            add $Y \longrightarrow X$ to $\mathcal{M}$;
        else
            add $Y \longleftrightarrow X$ to $\mathcal{M}$;
        end
    end
end
foreach $X \overset{\longleftrightarrow}{\longrightarrow} Y$ in $\mathcal{M}$ do
    remove $X \longleftrightarrow Y$;
end
```

$\mathcal{I}_m(\mathcal{S}) = \mathcal{I}_m(\mathcal{M})$, consider a DAG $\mathcal{G}$ constructed from $\mathcal{S}$ as follows: For every bi-directed edge $X \longleftrightarrow Y$, introduce a new node $L_{XY}$. Remove $X \longleftrightarrow Y$ and add $X \longleftarrow L_{XY} \longrightarrow Y$. Let $\mathbf{L}$ be the set $\{L_{V_iV_j}\}$ be the set of nodes added by this procedure. Obviously, $\mathcal{G}$ is a DAG and $\mathcal{G}$ and $\mathcal{S}$ share the same ancestry relations and the same $m$-separations for variables in $\mathbf{O}$, thus $\mathcal{I}_m(\mathcal{S}) = \mathcal{I}_m(\mathcal{G})[_{\mathbf{L}}$. If $\langle X, V_1, \ldots, V_n, Y \rangle$ is a primitive inducing path in $\mathcal{S}$, then $\langle X, L_{XV_1}, V_1, \ldots, L_{V_{n-1}V_n}, V_n, L_{V_nY}, Y \rangle$ is an inducing path with respect to $\mathbf{L}$ in $\mathcal{G}$ and vice versa. Thus, $X$ and $Y$ are adjacent in $\mathcal{G}[_{\mathbf{L}}$ if only if there exists a primitive inducing path between $X$ and $Y$ in $\mathcal{S}$, and $\mathcal{G}$ shares the same ancestry relations with $\mathcal{S}$ for variables in $\mathbf{O}$, thus by Definition 2.2.2, $\mathcal{G}[_{\mathbf{L}} = \mathcal{M}$. By Theorem 2.2.1 (Richardson and Spirtes, 2002) $\mathcal{I}_m(\mathcal{M}) = \mathcal{I}_m(\mathcal{G}[_{\mathbf{L}}) = \mathcal{I}_m(\mathcal{G})[_{\mathbf{L}} = \mathcal{I}_m(\mathcal{S})$.

The algorithm is also complete, since there only exists one such MAG. The inverse procedure, converting a MAG into the underlying SMCM, is not possible, since one cannot know in general which of the edges correspond to direct causation or confounding and which are there because of a (non-trivial) primitive inducing path. Note though that, there exist sound and complete algorithms that identify all edges for which such a determination is possible (Borboudakis et al., 2012). In addition, the next section shows that co-examining manipulated distributions can indicate that some edges stand for indirect causality (or indirect confounding).

## 2.4 Manipulations under causal insufficiency

An important motivation for using causal models is to predict causal effects. This work focus on hard manipulations, where the value of the manipulated variables is set exclusively by the manipulation procedure. The assumption of locality, denoting that the intervention of each manipulated variable should not directly affect any variable other than its direct target, and more importantly, local mechanisms for other variables should remain the same as before the intervention (Zhang, 2006), is also adopted. Thus, the intervention is merely a local surgery with respect to causal mechanisms.

These assumptions may seem a bit restricting, but this type of experiment is fairly common in several modern fields where the technical capability for precise interventions is available, such as, for example, molecular biology. Finally, faithfulness of the manipulated models to the corresponding manipulated distributions is assumed.

In the context of causal Bayesian networks, hard interventions are modeled using what is referred to as "graph surgery", in which all edges incoming to the manipulated variables are removed from the graph. The resulting graph is referred to as the **manipulated graph**. Naturally, DAGs are closed under manipulation. The term **intervention target** is used to denote a set of manipulated variables. For a DAG $\mathcal{G}$ and an intervention target $\mathbf{I}$, $\mathcal{G}^{\mathbf{I}}$ is used to denote the manipulated DAG. Parameters of the distribution that refer to the probability of manipulated variables given their parents are replaced by the parameters set by the manipulation procedure, while all other parameters remain intact. $\Pi^{\mathbf{I}}$ is used to denote this **manipulated joint probability distribution**, and $\mathcal{J}^{\mathbf{I}}$ to denote the corresponding **manipulated independence model**.

Graph surgery can be easily extended to SMCMs: One must simply remove edges into the manipulated variables. Again, the notation $\mathcal{S}^{\mathbf{I}}$ is used to denote the graph resulting from a SMCM $\mathcal{S}$ after the manipulation of variables in $\mathbf{I}$. In contrast, predicting the effect of manipulations in MAGs is not trivial. Due to the complicated semantics of the edges, the manipulated graph is usually not unique.

This becomes more obvious by looking at Figures 2.2 and 2.3. Figure 2.2 shows two different causal DAGs and the corresponding SMCMs and MAGs, and Figure 2.3 shows the effect of a manipulation on the same graphs. In Figure 2.2 the marginals of DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are represented by the same MAG $\mathcal{M}_1 = \mathcal{M}_2$. However, after manipulating variable $C$, the resulting manipulated MAGs $\mathcal{M}_1^C$ and $\mathcal{M}_2^C$ do not belong to the same equivalence class (they do not even share the same skeleton). The indistinguishability of $\mathcal{M}_1$ and $\mathcal{M}_2$ refers to $m$-separation only; the absence of a direct causal edge between $A$ and $D$ could be detected using other types of tests, like the Verma constraint (Verma and Pearl, 1990). Moreover, predicting the effects of manipulations is possible for *some* cases (Zhang, 2008b; Borboudakis et al., 2012; van der Zander et al., 2014; Maathuis and Colombo).

While the effect of manipulations on a MAG $\mathcal{M}$ cannot be predicted, given a data set measuring variables $\mathbf{O}$ when variables in $\mathbf{I} \subset \mathbf{O}$ are manipulated, the PAG representative of the actual manipulated MAG $\mathcal{M}^{\mathbf{I}}$ can be obtained (assuming an oracle of conditional independence) . $\mathcal{P}^{\mathbf{I}}$ is used to denote this PAG.

Notice that, $\mathcal{P}^{\mathbf{I}}$ is used to denote the representative of the Markov equivalence class of models that are faithful to the manipulated conditional independence model $\mathcal{J}(\Pi^{\mathbf{I}})$, as opposed to the representative of the *interventional Markov equivalence class* of manipulated MAGs. The information on manipulations, not included in the present use of $\mathcal{P}^{\mathbf{I}}$, defines a smaller Markov equivalence class: For example, in Figure 2.3, MAGs in the interventional Markov equivalence class of $\mathcal{M}_1^C$ share the additional invariant characteristic of a tail into $C$ on the edge $C \circ\!\!\rightarrow D$. This invariant feature however is not oriented in $\mathcal{P}_1^C$. To the best of our knowledge, no sound and complete algorithm for identifying the maximally informative PAG for the *interventional Markov equivalence class of* $\mathcal{M}^{\mathbf{I}}$ exists (however, orienting all edges out of the manipulated variables is a trivially sound method).

By observing PAGs $\{\mathcal{P}^{\mathbf{I}_i}\}$ that stem from known, different manipulations of the same underlying distribution, some refined information for the underlying causal model can be inferred. Let's suppose, for example, that $\mathcal{G}_1$ in Figure 2.2 is the true underlying causal graph for variables $\{A, B, C, D, L\}$ and that we have the learnt PAGs $\mathcal{P}_1^A$ and $\mathcal{P}_1^C$ from relevant data sets. Graph $\mathcal{P}_1^A$ is not shown, but is identical to $\mathcal{P}_1$ in Figure 2.2 since $A$ has no incoming edges in the underlying DAG (and SMCM). $\mathcal{P}_1^C$ is illustrated in Figure 2.3. Edge $A \circ\!\!-\!\!\circ D$ is present in $\mathcal{P}_1^A$, but is missing in $\mathcal{P}_1^C$ even though neither $A$ nor $D$ are manipulated in $\mathcal{P}_1^C$. By reasoning on the basis of both graphs, we can infer that edge $A \longrightarrow D$ in $\mathcal{P}_1^A$ cannot denote a *direct* causal relation among the two variables, but must be the result of a primitive, non-trivial inducing path.

Chapter 3

# Learning Causal Structure from Overlapping Manipulations

*Given the core assumptions of causality (i.e., the causal Markov and causal Faithfulness conditions), the statistical relations of the variables entailed by the available data sets constraint the configurations of possible paths in the underlying causal mechanism, represented here with a SMCM. We propose algorithm COmbINE, that efficiently translates statistical constraints into path constraints and encodes them into a SAT instance, whose truth-setting assignments correspond to possible underlying causal mechanisms. To tackle the complexity of the problem, the algorithm utilizes the relationship between MAGs and SMCMs described in the previous chapter to reduce the number of initial constraints. A series of additional heuristic or exact improvements are employed to further increase scalability. The algorithm is asymptotically sound and complete, and its oracle version is shown to outperform SBCSD, a brute-force algorithm that can perform the same inference task.*

In the previous section we described the effect of manipulation on MAGs and saw an example of how co-examining PAGs faithful to different manipulations of the same underlying distribution can help classify an edge between two variables as not direct.

In this section, we expand this idea and present a general, constraint-based algorithm for learning causal structure from overlapping manipulations. The algorithm takes as input a set of data sets measuring overlapping variable sets $\{\mathbf{O}_i\}_{i=1}^N$; in each data set, some of the observed variables can be manipulated. The set of manipulated variables in experiment $i$ is also provided and is denoted with $\mathbf{I}_i$.

In the rest of this work, we make the following assumptions:

**A1** We assume that there exists an underlying causal mechanism over a set of variables $\mathbf{O}$ that can be described with a semi Markov causal model $\mathcal{G}$ over $\mathbf{O}$. If $\Pi$ is the joint probability distribution over $\mathbf{O}$, we assume that $\Pi$ and $\mathcal{G}$ are faithful to each other, i.e. $\mathcal{J}_m(\mathcal{G}) = \mathcal{J}(\Pi)$. We also say that $\mathcal{G}$ is faithful to the independence model $graphJ(\Pi)$.

**A2** We assume that we collect data sets in $N$ different experiments, where in experiment $i$ we observe variables $\mathbf{O}_i \subseteq \mathbf{O}$, while variables $\mathbf{L}_i = \mathbf{O} \setminus \mathbf{O}_i$ are latent and variables $\mathbf{I}_i \subset \mathbf{O}$ are manipulated. We also assume $\mathbf{O} = \bigcup_{i=1}^{N} \mathbf{O}_i$. We assume that manipulations are ideal hard interventions and that they result in removal of all edges in $\mathcal{G}$ that are incoming to the manipulated variables.

**A3** We assume faithfulness for the manipulated SMCMs and distributions, i.e. $\mathcal{J}_m(\mathcal{G}^{\mathbf{I}_i}) = \mathcal{J}(\Pi^{\mathbf{I}_i})$.

Unless mentioned otherwise, the following notation is used:

- $\mathbf{O}_i$ denotes the set of observed variables in experiment $i$.
- $\mathbf{I}_i$ denotes the set of manipulated variables in experiment $i$.
- $\mathbf{O} = \cup_i \mathbf{O}_i$ denotes the union of observed variables.
- $\mathbf{L}_i = \mathbf{O} \setminus \mathbf{O}_i$ denotes the set of latent variables (with respect to the union of observed variables) in experiment $i$.
- $\mathbf{D}_i$ denotes a data set for experiment $i$, sampled from the mechanism described by $(\mathcal{G}^{\mathbf{I}_i}, \Pi^{\mathbf{I}_i})$, measuring variables in $\mathbf{O}_i$.
- $\mathcal{J}_i$ denotes the independence model that holds in data set $i$. In the sample limit, $\mathcal{J}_i$ is equal to the set of $m$-separations that hold for sets of variables in $\mathbf{O}_i$ after manipulating $\mathbf{I}_i$ in the underlying causal model: $\mathcal{J}_i = \mathcal{J}(\Pi^{\mathbf{I_i}})[_{\mathbf{L_i}} = \mathcal{J}_m(\mathcal{G}^{\mathbf{I_i}})[_{\mathbf{L_i}}$.
- $\mathcal{P}_i$ denotes the maximally informative PAG for the (observational) Markov equivalence class of MAGs faithful to $\mathcal{J}_i$. Thus, for any MAG $\mathcal{M}_i \in \mathcal{P}_i$, $\mathcal{J}_m(\mathcal{M}_i) = \mathcal{J}_i$. Notice that, since SMCMs and MAGs over the same variables represent the same independence model, for an oracle of conditional independence, $\mathcal{P}_i = [\text{SMCMtoMAG}(\mathcal{G}^{\mathbf{I}_i})[_{\mathbf{L}_i}]$.

Under the assumptions described above, we are interested in combining information across data sets collected from different manipulations and marginalizations of the same system under study, to identify features of the possible underlying causal mechanism. If $\mathcal{S}$ is a SMCM that describes this underlying causal mechanism, then this SMCM must agree with all the observed independence models $\{\mathcal{J}_i\}_{i=1}^{N}$. This means that for each experiment, the respective manipulated $\mathcal{S}^{\mathbf{I}_i}$ must entail all and only the conditional independencies that hold in data set $\mathbf{D}_i$ (in the sample limit $\mathcal{J}_i$ can be obtained correctly from the data). For the family of independence models $\{\mathcal{J}_i\}_{i=1}^{N}$, and a family of intervention targets $\{\mathbf{I}_i\}_{i=1}^{N}$ a **possibly underlying SMCM** is defined as follows:

**Definition 3.0.1 (Possibly underlying SMCM)** *If $\{\mathcal{J}_i\}_{i=1}^{N}$ is a family of independence models over variable sets $\{\mathbf{O}_i\}_{i=1}^{N}$ and $\{\mathbf{I}_i\}_{i=1}^{N}$ is a family of intervention targets such that $\mathbf{I}_i \subseteq \mathbf{O}_i \quad \forall i$, then a SMCM $\mathcal{S}$ is a* ***possibly underlying SMCM*** *for $\{\mathcal{J}_i\}_{i=1}^{N}$ and $\{\mathbf{I}_i\}_{i=1}^{N}$ iff:*

$$\forall X, Y, \mathbf{Z} \subseteq \mathbf{O}_i,\ [X \text{ is } m\text{-separated from } Y \text{ given } \mathbf{Z} \text{ in } \mathcal{S}^{\mathbf{I}_i}] \Leftrightarrow X \perp\!\!\!\perp Y \mid \mathbf{Z} \in \mathcal{J}_i,$$

Intuitively, $\mathcal{S}$ is a SMCM such that once the effects of manipulations are modeled (i.e. $\mathcal{S}^{\mathbf{I}_i}$ is constructed), it entails all and only the independencies $\mathcal{J}_i$ observed in the corresponding data set. Thus, $\mathcal{S}$ is a possible causal model that explains all data. Since each independence model $\mathcal{J}_i$ can be graphically represented by a PAG $\mathcal{P}_i$, one can recast this definition in graph-theoretic terms: $\mathcal{S}$ is a possibly underlying SMCM if, after graph surgery, results in a marginal MAG that belongs in $\mathcal{P}_i$:

**Theorem 3.0.1** *If $\mathcal{S}$ is a SMCM, $\{\mathcal{J}_i\}_{i=1}^N$ is a family of independence models, $\{\mathbf{I}_i\}_{i=1}^N$ is a family of intervention targets and $\mathcal{P}_i$ is the PAG of the Markov equivalence class of MAGs faithful to $\mathcal{J}_i$, the following statements are equivalent:*

- *$\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.*
- $\forall i, \mathcal{M}_i \in \mathcal{P}_i,$ where $\mathcal{M}_i = \text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i}.$

**Proof**

$\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N \Leftrightarrow \mathcal{J}_m(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} = \mathcal{J}_i \quad \forall i \quad$ (by definition).

$$\mathcal{J}_m(\text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i}))[_{\mathbf{L}_i} = \mathcal{J}_m(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} = \mathcal{J}_i \quad \forall i \quad \text{(by Theorem 2.3.4)}$$
$$\mathcal{J}_m(\text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i}) = \mathcal{J}_m(\text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i}))[_{\mathbf{L}_i} = \mathcal{J}_i \quad \forall i \quad \text{(by Theorem 2.2.1)}$$
$$\mathcal{J}_m(\mathcal{M}_i) = \mathcal{J}_i \quad \forall i, \text{ and by definition of } \mathcal{P}_i, \quad \mathcal{M}_i \in \mathcal{P}_i \quad \forall i.$$

As mentioned above, PAGs $\mathcal{P}_i$ here denote the maximally informative representatives of the Markov equivalence class of MAGs that entail independence models $\mathcal{J}_i$, instead of the *interventional* Markov equivalence class of MAGs that entail both $\mathcal{J}_i$ and the interventional constraints following the manipulation of targets $\mathbf{I_i}$. Hence, this graphical criterion may seem incomplete, since the actual MAGs belong to thinner equivalence classes, which include some additional orientations: tails towards any manipulated variable and additional orientations stemming from the combination of $m$-separation and acyclicity with these aforementioned tails. However, MAGs $\mathcal{M}_i = \text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i}$ are constructed after graph surgery has been applied to the (candidate) possibly underlying SMCM and abide by definition the constraints that correspond to interventional information (i.e. tail orientations towards manipulated variables), since $\mathcal{S}^{\mathbf{I}_i}$ and $\text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})$ share the same ancestral relations. Thus, the resulting MAGs $\mathcal{M}_i$ belong (by construction) to the thinner *interventional* Markov equivalence class of MAGs, and testing Markov equivalence in the observational sense is a sound and complete graphical criterion to determine whether a SMCM is possibly underlying for a family of independence models coupled with a family of intervention targets.

Notice that PAG $\mathcal{P}_i$ can be learnt with a sound and complete algorithm such as FCI. We can now benefit by the compact representation of Markov equivalence classes of MAGs described in Theorem 2.2.2, to check whether a SMCM $\mathcal{S}$ is possibly underlying for a family of independence models $\{\mathcal{J}_i\}_{i=1}^N$ and a family of intervention targets $\{\mathbf{I}_i\}_{i=1}^N$: Instead of checking *all* conditional dependencies (resp. independencies) in $\mathcal{J}_i$ to be $m$-connections (resp. $m$-separations) in the corresponding SMCM $\mathcal{S}^{\mathbf{I}_i}$, we can construct the corresponding MAGs $\mathcal{M}_i = \text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i}$ for each experiment and check whether they belong to the Markov equivalence class represented by $\mathcal{P}_i$. By Theorem 2.2.2, we only need to check adjacencies and colliders with order.

In the next section, we present an algorithm that converts the problem of identifying a SMCM $\mathcal{S}$ that is possibly underlying for a family of observed independence models $\{\mathcal{J}_i\}_{i=1}^N$ and a family of intervention targets $\{\mathbf{I}_i\}_{i=1}^N$ into a constraint satisfaction problem. Specifically, we will create a satisfiability instance s.t. a SMCM is possibly underlying for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$ if and only if it corresponds to a truth-setting assignment for the SAT instance. For a family of independence models $\{\mathcal{J}_i\}_{i=1}^N$ and a family of intervention targets $\{\mathbf{I}_i\}_{i=1}^N$, several SMCMs may be possibly underlying. We can then use the equivalent SAT instance to query properties shared by all possibly underlying SMCMs, or to identify a single possibly underlying SMCM with some desirable characteristics. In this work, we use the equivalent SAT instance to identify all edges and endpoints that are invariant in all possibly underlying SMCMs.

## 3.1 Conversion to SAT

Theorem 3.0.1 implies that $\mathcal{M}_i$ has the same edges (adjacencies), and the same colliders with order (unshielded colliders and discriminating colliders with order) as any MAG in $\mathcal{P}_i$, for all $i$. We impose these constraints on $\mathcal{S}$ by converting them to a SAT instance. We express the constraints in terms of the following **core** variables, denoting edges and orientations in any possibly underlying SMCM $\mathcal{S}$.

- edge($X$, $Y$): true if $X$ and $Y$ are adjacent in $\mathcal{S}$, false otherwise.
- tail($X$, $Y$): true if there exists an edge between $X$ and $Y$ in $\mathcal{S}$ that is out of $Y$, false otherwise.
- arrow($X$, $Y$): true if there exists an edge between $X$ and $Y$ in $\mathcal{S}$ that is into $Y$, false otherwise.

Variables tail and arrow are not mutually exclusive, enabling us to represent $X \overset{\leftrightarrow}{\rightarrow} Y$ edges when $tail(Y, X) \wedge arrow(Y, X)$. Each independence model $\mathcal{J}_i$ is entailed by the (non) adjacencies and (non) colliders in each observed PAG $\mathcal{P}_i$. These structural characteristics correspond to paths in any possibly underlying SMCM as follows:

1. $\forall X, Y \in \mathbf{O}_i$, $X$ and $Y$ are adjacent in $\mathcal{P}_i$ if and only if there exists an inducing path between $X$ and $Y$ with respect to $\mathbf{L_i}$ in $\mathcal{S}^{\mathbf{I}_i}$ (by Theorems 2.3.1 and 2.3.2 and Proposition 2.3.3).
2. If $\langle X, Y, Z\rangle$ is an unshielded definite non collider in $\mathcal{P}_i$, then $\langle X, Y, Z\rangle$ is an unshielded triple in $\mathcal{P}_i$ and $Y$ is an ancestor of either $X$ or $Z$ in $\mathcal{S}^{\mathbf{I}_i}$ (by the semantics of edges in MAGs).
3. If $\langle X, Y, Z\rangle$ is an unshielded collider in $\mathcal{P}_i$, then $\langle X, Y, Z\rangle$ is an unshielded triple in $\mathcal{P}_i$ and $Y$ is not an ancestor of $X$ nor $Z$ in $\mathcal{S}^{\mathbf{I}_i}$ (by the semantics of edges in MAGs).
4. If $\langle W, \dots, X, Y, Z\rangle$ is a discriminating collider in $\mathcal{P}_i$, then $\langle W \dots, X, Y, Z\rangle$ is a discriminating path for $Y$ in $\mathcal{P}_i$ and $Y$ is not an ancestor of $X$ nor $Z$ in $\mathcal{S}^{\mathbf{I}_i}$ (by the semantics of edges in MAGs).
5. If $\langle W, \dots, X, Y, Z\rangle$ is a discriminating definite non collider in $\mathcal{P}_i$, then $\langle W \dots, X, Y, Z\rangle$ is a discriminating path for $Y$ in $\mathcal{P}_i$ and $Y$ is an ancestor either $X$ or $Z$ in $\mathcal{S}^{\mathbf{I}_i}$ (by the semantics of edges in MAGs).

$\mathbf{adjacent}(X, Y, \mathcal{P}_i) \leftrightarrow \exists p_{XY} : \mathit{inducing}(p_{XY}, i)$

/$^\star X$ and $Y$ are **adjacent** in $\mathcal{P}_i$ iff
there exists an inducing path from $X$ to $Y$ with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$. $_\star$/

$\mathbf{collider}(\langle X, Y, Z\rangle, \mathcal{P}_i) \leftrightarrow \neg \mathit{ancestor}(Y, X, i) \wedge \neg \mathit{ancestor}(Y, Z, i)$

/$^\star$ Triple $\langle X, Y, Z\rangle$ is **collider** in $\mathcal{P}_i$ iff
$Y$ is not an ancestor of $X$ or $Z$ in $\mathcal{S}^{\mathbf{I}_i}$. $_\star$/

$\mathbf{unshielded}(\langle X, Y, Z\rangle, \mathcal{P}_i) \leftrightarrow$
$\mathit{adjacent}(X, Y, \mathcal{P}_i) \wedge \mathit{adjacent}(Y, Z, \mathcal{P}_i) \wedge \neg \mathit{adjacent}(X, Z, \mathcal{P}_i)$

/$^\star$ Triple $\langle X, Y, Z\rangle$ is **unshielded** in $\mathcal{P}_i$ iff
$(X,\ Y)$, $(Y,\ Z)$ are adjacent in $\mathcal{P}_i, (X, Z)$ are not adjacent in $\mathcal{P}_i$. $_\star$/

$\mathbf{discriminating}(\langle V_0, \ldots, V_{n-1}, V_n, V_{n+1}\rangle, V_n, \mathcal{P}_i) \leftrightarrow$
$\neg \mathit{adjacent}(V_0, V_{n+1}, \mathcal{P}_i) \wedge \forall j \in [0, \ldots, n] \mathit{adjacent}(V_j, V_{j+1}, \mathcal{P}_i) \wedge$
$\forall j \in [1, \ldots, n-1] \big(\mathit{collider}(\langle V_{j-1}, V_j, V_{j+1}\rangle, \mathcal{P}_i)$
$\wedge\, \mathit{adjacent}(V_j, V_{n+1}, \mathcal{P}_i) \wedge \mathit{ancestral}(V_j, V_{n+1}, i)\big)$

/$^\star$ Path $\langle V_0, \ldots, V_{n+1}\rangle$ is **discriminating** for $V_n$ in $\mathcal{P}_i$ iff
$V_0, V_{n+1}$ are not adjacent in $\mathcal{P}_i$, $V_0, \ldots, V_{n+1}$ is a path in $\mathcal{P}_i$,
every node between $V_0$ and $V_n$ is a collider on the path
and a parent of $V_{n+1}$ in $\mathcal{P}_i$. $_\star$/

Figure 3.1: **Formulae relating properties of observed PAGs to the underlying SMCM $\mathcal{S}$.** In each PAG, all features that are necessary and sufficient for Markov equivalence impose constraints on possibly underlying SMCMs. Constraints are expressed using the literals and formulae introduced here. Index i is used to denote properties of an underlying SMCM in experiment i, where variables $\mathbf{L}_i$ are latent and variables $\mathbf{I}_i$ are manipulated. We use use $p_{XY}$ to denote a path between $X$ and $Y$ in $\mathcal{S}$. Conjunction and disjunction are assumed to have precedence over implication with regard to bracketing. Each formula is followed by an explanation in in natural language (in star-slash comments).

These constraints are expressed using the core variables (edges, tails and arrows), as described in Figures 3.1 and 3.2. Figure 3.1 describes how features of a PAG are imposed as path constraints in a possibly underlying SMCM. More specifically, an adjacency, a tail and an arrowhead in a PAG $\mathcal{P}_i$ correspond to an inducing path, a causal ancestry and the lack of causal ancestry on any possibly underlying SMCM, respectively. Unshielded triples and discriminating paths are expressed on the basis of these basic PAG features. In each PAG, the observed features depend on the latent and manipulated variables. When constraints are imposed on the candidate underlying SMCMs, the latent and manipulated variables in the experiment are taken under consideration: If an adjacency is observed in $\mathcal{P}_i$ in experiment i, where variables $\mathbf{L}_i$ are latent and $\mathbf{I}_i$ are manipulated, then any path on $\mathcal{S}$ that explains this adjacency must be inducing with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$. Any truth-assignment

$\textbf{inducing}(\langle V_0, \ldots, V_{n+1}\rangle, i) \leftrightarrow$
$(n = 0 \rightarrow \mathit{edge}(V_0, V_{n+1}))\wedge$
$\big(n > 0 \rightarrow (\forall j \in [1, \ldots, n]\ \mathit{unblocked}(\langle V_{j-1}, V_j, V_{j+1}\rangle, V_0, V_{n+1}, i))\big)\wedge$
$(V_0 \in \mathbf{I}_i \rightarrow \mathit{tail}(V_1, V_0)) \wedge (Y \in \mathbf{I}_i \rightarrow \mathit{tail}(V_n, V_{n+1}))$

/$^\star$ Path $\langle V_0, \ldots, V_{n+1}\rangle$ is **inducing** with respect to $\mathbf{L_i}$ in $\mathcal{S}^{\mathbf{I}_i}$ iff
if the path has only two variables, $V_0$ is adjacent to $V_n$ in $\mathcal{S}$
else each triple is **unblocked** for the endpoints with respect to $\mathbf{L_i}$,
if $V_0$ $(V_{n+1})$ is manipulated in i then the path is out of $V_0$ $(V_{n+1})$ in $\mathcal{S}$. $_\star$/

$\textbf{unblocked}(\langle Z, V, W\rangle, X, Y, i) \leftrightarrow$
$\mathit{edge}(Z, V) \wedge \mathit{edge}(V, W)\wedge$
$[V \in \mathbf{L}_i \rightarrow \neg \mathit{head2head}(\langle Z, V, W\rangle, i) \vee \mathit{ancestor}(V, X, i) \vee \mathit{ancestor}(V, Y, i)]\wedge$
$[V \notin \mathbf{L}_i \rightarrow \mathit{head2head}(\langle Z, V, W\rangle, i) \wedge (\mathit{ancestor}(V, X, i) \vee \mathit{ancestor}(V, Y, i))]$

/$^\star$ Triple $\langle Z, V, W\rangle$ is **unblocked** for $X$, $Y$ with respect to $\mathbf{L_i}$ iff
$(Z,\ V)$ $(V,\ W)$ are adjacent in $\mathcal{S}$
if $V$ is latent, if $V$ is **head2head** then it is an ancestor of $X$ or $Y$ in $\mathcal{S}^{\mathbf{I}_i}$
if $V$ is not latent, $V$ is a **head2head** and an ancestor of $X$ or $Y$ in $\mathcal{S}^{\mathbf{I}_i}$. $_\star$/

$\textbf{head2head}(\langle X, Y, Z\rangle, i) \leftrightarrow Y \notin \mathbf{I}_i \wedge \mathit{arrow}(X, Y) \wedge \mathit{arrow}(Z, Y)$
/$^\star$ Triple $\langle X, Y, Z\rangle$ is **head2head** in $\mathcal{S}^{\mathbf{I}_i}$ iff
$Y$ is not manipulated in experiment i, $X$ is into $Y$, $Z$ is into $Y$ in $\mathcal{S}$. $_\star$/

$\textbf{ancestor}(X, Y, i) \leftrightarrow \exists p_{XY} : \mathit{ancestral}(p_{XY}, i)$
/$^\star$ $X$ is an **ancestor** of $Y$ in experiment i iff
there exists an ancestral path from $X$ to $Y$ in $\mathcal{S}^{\mathbf{I}_i}$. $_\star$/

$\textbf{ancestral}(\langle V_0, \ldots, V_{n+1}\rangle, i) \leftrightarrow$
$\forall j \in [1, \ldots, n+1]\big(V_j \notin \mathbf{I}_i \wedge (\mathit{edge}(V_{j-1}, V_j) \wedge \mathit{tail}(V_j, V_{j-1}) \wedge \mathit{arrow}(V_{j-1}, V_j))\big)$
/$^\star$ Path $\langle V_0, \ldots, V_{n+1}\rangle$ is **ancestral** in experiment i iff
every variable (apart from possibly $V_0$) is not manipulated in $\mathcal{S}^{\mathbf{I}_i}$
every variable is a parent of the next in $\mathcal{S}$. $_\star$/

Figure 3.2: **Formulae reducing path properties of the graphs $\mathcal{S}^{\mathbf{I}_i}$ to the core variables:** Graph properties of $\mathcal{S}$ in each experiment, inferred by the observed PAGs using the formulae in Figure 3.1, are now expressed as boolean formulae using the "core" variables *edge*, *arrow* and *tail*. Index i is used to denote properties of an underlying SMCM in experiment i, where variables $\mathbf{L}_i$ are latent and variables $\mathbf{I}_i$ are manipulated. Conjunction and disjunction are assumed to have precedence over implication with regard to bracketing. Each formula is followed by an explanation in in natural language (in star-slash comments).

to the core variables that does not entail the presence of such an inducing path should not satisfy the

**Algorithm 2:** COmbINE

```
input : data sets $\{\mathbf{D}_i\}_{i=1}^N$, sets of intervention targets $\{\mathbf{I}_i\}_{i=1}^N$, FCI parameters $params$,
        maximum path length $mpl$, conflict resolution strategy $str$
output: Summary graph $\mathcal{H}$
foreach $i$ do $\mathcal{P}_i \leftarrow$ FCI($\mathbf{D}_i$, $params$);
$\mathcal{H}_{in} \leftarrow$ initializeSMCM ($\{\mathcal{P}_i\}_{i=1}^N$);
$(\Phi, \mathcal{F}) \leftarrow$ addConstraints ($\mathcal{H}$, $\{\mathcal{P}_i\}_{i=1}^N$, $\{\mathbf{I}_i\}_{i=1}^N$, $mpl$);
$\mathcal{F}' \leftarrow$ select a subset of non-conflicting literals $\mathcal{F}'$ according to strategy $str$;
$\mathcal{H} \leftarrow$ backBone ($\Phi \wedge \mathcal{F}'$)
```

SAT instance. The following constraints are added to ensure that the graphs satisfying constraints 1-5 above are SMCMs:

6. $\forall X, Y \in \mathbf{O}$, either $X$ is not an ancestor of $Y$ or $Y$ is not an ancestor of $X$ in $\mathcal{S}$ (no directed cycles).
7. $\forall X, Y \in \mathbf{O}$, at most one of $tail(X, Y)$ and $tail(Y, X)$ can be true (no selection bias).
8. $\forall X, Y \in \mathbf{O}$, at least one of $tail(X, Y)$ and $arrow(Y, X)$ must be true.

Naturally, Constraints 7 and 8 are meaningful only if $X$ and $Y$ are adjacent (if edge(X, Y) is true), and redundant otherwise.

## 3.2 Algorithm COmbINE

We now present algorithm **COmbINE** (Causal discovery from Overlapping INtErventions) that learns causal features from multiple, heterogenous data sets. The algorithm takes as input a set of data sets $\{\mathbf{D}_i\}_{i=1}^N$ over a set of overlapping variable sets $\{\mathbf{O}_i\}_{i=1}^N$. In each data set, a (possibly empty) subset of the observed variables $\mathbf{I}_i \subset \mathbf{O}_i$ may be manipulated. Each data set entails an independence model $\mathcal{J}_i$. FCI is run on each data set and the corresponding PAGs $\{\mathcal{P}_i\}_{i=1}^N$ are produced. The algorithm then creates a candidate underlying SMCM $\mathcal{H}_{in}$. Subsequently, for each PAG $\mathcal{P}_i$, the features of $\mathcal{P}_i$ are translated into constraints, expressed in terms of edges and endpoints in $\mathcal{H}_{in}$, using the formulae in Figures 3.1 and 3.2 . In the sample limit (and under the assumptions discussed above), the SAT formula $\Phi \wedge \mathcal{F}$ produced by this procedure is satisfied by all and only the possibly underlying SMCMs for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}$. In the presence of statistical errors, however, $\Phi \wedge \mathcal{F}$ may be unsatisfiable. To handle conflicts, the algorithm takes as input a strategy for selecting a non-conflicting subset of constraints $\mathcal{F}'$ and ignores the rest. Finally, COmbINE queries the SAT formula for variables that have the same truth-value in all satisfying assignments, translates them into graph features, and returns a graph that summarizes the invariant edges and orientations of all possibly underlying SMCMs. We call the graphical output of COmbINE a **summary graph**.

The pseudocode for COmbINE is presented in Algorithm 2. Apart from the set of data sets described above, COmbINE takes as input the chosen parameters for FCI (threshold $\alpha$, maximum conditioning set $maxK$), the maximum length of paths to consider and a strategy for selecting a subset of non-conflicting constraints.

Initially, the algorithm runs FCI on each data set $\mathbf{D}_i$ and produces the corresponding PAG $\mathcal{P}_i$. Then the candidate SMCM $\mathcal{H}_{in}$ is initialized: $\mathcal{H}_{in}$ is the graph upon which all path constraints will be imposed. Path constraints are realized on the basis of the *plausible configurations* of $\mathcal{H}_{in}$. We say that a path $p$ in $\mathcal{H}_{in}$ is **possibly inducing with respect to $\mathbf{L}$**, if we can create a graph $\mathcal{H}'_{in}$ by orienting circle endpoints in $\mathcal{H}_{in}$ such that path $p$ is inducing with respect to $\mathbf{L}$ in $\mathcal{H}'_{in}$. We say that a path $p$ in $\mathcal{H}_{in}$ is **possibly ancestral**, if we can create a graph $\mathcal{H}'_{in}$ by orienting circle endpoints in $\mathcal{H}_{in}$ such that path $p$ is ancestral $\mathcal{H}'_{in}$. To ensure the soundness of the algorithm, if $p$ is an inducing (ancestral) path in $\mathcal{S}$, it must be a possibly inducing (ancestral) path in $\mathcal{H}_{in}$. Thus, $\mathcal{H}_{in}$ must have at least a superset of edges and at most a subset of orientations of any possibly underlying SMCM $\mathcal{S}$.

An obvious–yet not very smart–choice for $\mathcal{H}_{in}$ would be the complete unoriented graph. However, looking for possibly inducing and possibly ancestral paths on the complete unoriented graph over the union of variables could make the problem intractable even for small input sizes. To reduce the number of possibly inducing and possibly ancestral paths, we use Algorithm 3 to construct $\mathcal{H}_{in}$.

Algorithm 3 constructs a graph $\mathcal{H}_{in}$ that has all edges observed in any PAG $\mathcal{P}_i$ as well as some additional edges that would not have been observed even if they existed: Edges connecting variables that have never been observed together, and edges connecting variables that have been observed together, but at least one of them was manipulated in each joint appearance in a data set. For example, variables $X9$ and $X15$ in Figure 3.4 are measured together in two data sets: $\mathbf{D}_2$ and $\mathbf{D}_3$. If $X9 \longrightarrow X15$ in the underlying SMCM, this edge would be present in $\mathcal{P}_3$. Similarly, if $X15 \longrightarrow X9$ in the underlying SMCM, the variables would be adjacent in $\mathcal{P}_2$. We can therefore rule out the possibility of a directed edge between the two variables in $\mathcal{S}$. However, it is possible that $X15$ and $X9$ are confounded in $\mathcal{S}$, and the edge disappears by the manipulation procedure in both $\mathcal{P}_2$ and $\mathcal{P}_3$. Thus, Algorithm 3 will add these possible edges in $\mathcal{H}_{in}$. In addition, in Line 5, Algorithm 3 adds all the orientations found so far in all $\mathcal{P}_i$'s that are invariant[1]. The resulting graph has, in the sample limit, a superset of edges and a subset of orientations compared to the actual underlying SMCM. Lemma 3.3.1 formalizes and proves this property.

Having initialized the search graph, Algorithm 2 proceeds to generate the constraints. This procedure is described in detail in Algorithm 4, that is the core of COmbINE. These are: (i) the bi-conditionals regarding the presence/absence of edges (Line 4), (ii) conditionals regarding unshielded and discriminating colliders (Lines 14, 13, 20 and 19), (iii) constraints that ensure that any truth-setting assignment is a SMCM, i.e., it has no directed cycles and that every edge has at least one arrowhead (Lines 8 and 9 respectively). Literal *col* (resp. *dnc*) is used to represent both unshielded and discriminating colliders (resp. unshielded and discriminating non colliders).

The constraints are realized on the basis of the *plausible* configurations of $\mathcal{H}_{in}$: Thus, for the constraints corresponding to $adjacent(X, Y, i)$ the algorithm finds all paths between $X$ and $Y$ in $\mathcal{H}_{in}$ that are possibly inducing. Then, for the literal $adjacent(X, Y, i)$ to be true, at least one of these paths is constrained to be inducing; for the opposite, none of these paths is allowed to be

[1] Other options would be to keep all non-conflicting arrows, or keep non-conflicting arrows and tails after some additional analysis on definitely visible edges (see Zhang, 2008b; Borboudakis et al., 2012, for more on this subject). These options are asymptotically correct and would constrain search even further. Nevertheless, orientation rules in FCI seem to be prone to error propagation and we chose a more conservative strategy giving a chance to the conflict resolution strategy to improve the learning quality. Naturally, if an oracle of conditional independence is available or there is a reason to be confident on certain features, one can opt to make additional orientations.

**Algorithm 3:** initializeSMCM

```
input  : PAGs $\{\mathcal{P}_i\}_{i=1}^N$
output: initial graph $\mathcal{H}_{in}$
$\mathcal{H}_{in} \leftarrow$ empty graph over $\cup\mathbf{O}_i$;
foreach $i$ do
    $\mathcal{H}_{in} \leftarrow$ add all edges in $\mathcal{P}_i$ unoriented;
end
Orient only arrowheads that are present in every $\mathcal{P}_i$;
/* Add edges between variables never measured unmanipulated together */
foreach pair $X, Y$ of non-adjacent nodes do
    if $\nexists i$ s.t. $X, Y \in \mathbf{O}_i \setminus \mathbf{I}_i$ then
        add $X{\circ\!\!-\!\!\circ}Y$ to $\mathcal{H}_{in}$;
        if $\exists i$ s.t. $X, Y \in \mathbf{O}_i$, $X \in \mathbf{I}_i$, $Y \notin \mathbf{I}_i$ then add arrow into $X$;
        if $\exists i$ s.t. $X, Y \in \mathbf{O}_i$, $Y \in \mathbf{I}_i$, $X \notin \mathbf{I}_i$ then add arrow into $Y$;
    end
end
```

inducing. This step is the most computationally expensive part of the algorithm. The parameter *mpl* controls the length of the possibly inducing paths; instead of finding *all* paths between $X$ and $Y$ that are possibly inducing, the algorithm looks for all paths of length at most *mpl*. This plays a major part in the ability of the algorithm to scale up, since finding all possible paths between every pair of variables can blow up even in relatively small networks, particularly in the presence of unoriented cliques or in relatively dense networks.

Notice that the information on manipulations is included in the satisfiability instance through the encoding of the constraints: For every adjacency between $X$ and $Y$ observed in $\mathcal{P}_i$, the plausible inducing paths are consistent with the respective intervention targets: No inducing path is allowed to include an edge that is incoming to a manipulated variable.

As an example, consider the following variation of the instance presented in Figure 3.5. Assume that variable $X$ is manipulated in experiment 1, and no variable is manipulated in experiment 2. Since no information concerning experiments is employed up to the initialization of the search graph, the resulting PAGs are the $\mathcal{P}_1$ and $\mathcal{P}_2$ shown in Figure 3.5. Thus, in the initial search graph $\mathcal{H}_{in}$, $X{\circ\!\!-\!\!\circ}Y$ and $X{\circ\!\!-\!\!\circ}Z{\circ\!\!-\!\!\circ}Y$ are the two possibly inducing paths for $X$ and $Y$ in experiment $i$. Then the following constraint will be imposed:

$$adjacent(X, Y, 1) \leftrightarrow inducing(\langle X, Y\rangle, 1) \vee inducing(\langle X, Z, Y\rangle, 1)$$

For path $\langle X, Y\rangle$, the corresponding constraint is reduced to the properties of $\mathcal{S}$ as follows:

$$\begin{aligned}&inducing(\langle X, Y\rangle, 1) \leftrightarrow \\ &\quad (X \in \mathbf{I}_1 \rightarrow tail(Y, X)) \wedge (Y \in \mathbf{I}_1 \rightarrow tail(X, Y)) \wedge edge(X, Y)\end{aligned}$$

**Algorithm 4:** addConstraints

```
input  : $\mathcal{H}_{in}$, $\{\mathcal{P}_i\}_{i=1}^N$, $\{\mathbf{I}_i\}_{i=1}^N$, $mpl$
output: $\Phi$, list of literals $\mathcal{F}$
$\Phi \leftarrow \emptyset$ foreach $X, Y$ do
    foreach $i$ do
        posIndPaths$\leftarrow$ paths in $\mathcal{H}_{in}$ of maximum length $mpl$ that are possibly inducing with respect to $\mathbf{L}_i$;
        $\Phi \leftarrow \Phi \wedge \big[adjacent(X, Y, \mathcal{P}_i) \leftrightarrow \exists p_{XY} \in$posIndPaths s. t. $inducing(p_{XY}, i)\big]$;
        if $X$, $Y$ are adjacent in $\mathcal{P}_i$ then add $adjacent(X, Y, \mathcal{P}_i)$ to $\mathcal{F}$;
        else add $\neg adjacent(X, Y, \mathcal{P}_i)$ to $\mathcal{F}$;
    end
    $\Phi \leftarrow \Phi \wedge \big[\neg ancestor(X, Y) \vee \neg ancestor(Y, X)\big]$;
    $\Phi \leftarrow \Phi \wedge \big[\neg tail(X, Y) \vee \neg tail(Y, X)\big] \wedge \big[(arrow(X, Y) \vee tail(X, Y)\big]$;
end
foreach $i$ do
    foreach unshielded triple in $\mathcal{P}_i$ do
        $\Phi \leftarrow \Phi \wedge \big[col(X, Y, Z, \mathcal{P}_i) \rightarrow unshielded(X, Y, Z, \mathcal{P}_i) \wedge collider(X, Y, Z, \mathcal{P}_i)\big]$;
        $\Phi \leftarrow \Phi \wedge \big[dnc(X, Y, Z, \mathcal{P}_i) \rightarrow unshielded(X, Y, Z, \mathcal{P}_i) \wedge \neg collider(X, Y, Z, \mathcal{P}_i)\big]$;
        if $\langle X, Y, Z\rangle$ is a collider in $\mathcal{P}_i$ then add $col(X, Y, Z, \mathcal{P}_i)$ to $\mathcal{F}$;
        else add $dnc(X, Y, Z, \mathcal{P}_i)$ to $\mathcal{F}$;
    end
    foreach discriminating path $p_{WZ} = \langle W, \ldots, X, Y, Z\rangle$ do
        $\Phi \leftarrow \Phi \wedge \big[col(X, Y, Z, \mathcal{P}_i) \rightarrow discriminating(p_{WZ}, Y, \mathcal{P}_i) \wedge collider(X, Y, Z, \mathcal{P}_i)\big]$;
        $\Phi \leftarrow \Phi \wedge \big[dnc(X, Y, Z, \mathcal{P}_i) \rightarrow discriminating(p_{WZ}, Y, \mathcal{P}_i) \wedge \neg collider(X, Y, Z, \mathcal{P}_i)\big]$;
        if $X$, $Y$, $Z$ is a collider in $\mathcal{P}_i$ then add $col(X, Y, Z, \mathcal{P}_i)$ to $\mathcal{F}$;
        else add $dnc(X, Y, Z, \mathcal{P}_i)$ to $\mathcal{F}$;
    end
end
```

which is then added in $\Phi$ as $inducing(\langle X, Y\rangle, 1) \leftrightarrow tail(Y, X) \wedge edge(X, Y)$ since $X \in \mathbf{I}_1$ is true and $Y \in \mathbf{I}_1$ is false. For the path $\langle X, Z, Y\rangle$ the corresponding constraint finally added in $\Phi$ is

$$\begin{aligned}&inducing(\langle X, Z, Y\rangle) \leftrightarrow\\ &\quad tail(Z, X) \wedge [\neg head2head(\langle X, Z, Y\rangle) \vee ancestral(Z, X) \vee ancestral(Z, Y)]\end{aligned}$$

Thus, in a SMCM that satisfies the final formula, if $inducing(\langle X, Y\rangle, i)$ is true, there will be an inducing path from $X$ to $Y$ consistent with the manipulation information.

Also notice how this constraint is *propagated* in the SAT: For example, $X{*\!\!-\!\!*}Z{*\!\!-\!\!*}Y{*\!\!-\!\!*}W$ is a plausible skeleton for a possibly underlying SMCM. By the constraints mentioned above, $X \rightarrow Z{*\!\!-\!\!*}Y$ is the inducing path for $X$ and $Y$ with respect to $L_1 = Z$. By the constraints added for the definite non collider $\langle X, Z, W\rangle$ for $\mathcal{P}_2$, $Z$ has to be an ancestor of either $X$ or $Y$ in $\mathcal{S}^{\emptyset}$. Therefore, the path $Z{*\!\!-\!\!*}Y{*\!\!-\!\!*}W$ has to be an ancestral path in $\mathcal{S}$, which implies that $Y \longrightarrow Z$ in $\mathcal{S}$. Thus, the orientation $Y \longrightarrow Z$ is imposed by a combination of constraints stemming from different PAGs, for two variables never jointly measured.

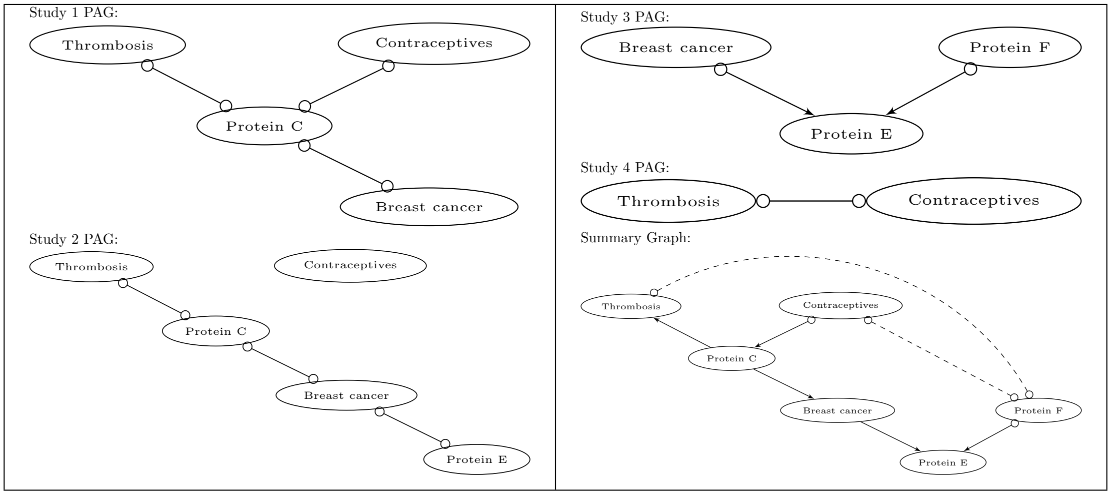


Figure 3.3: **COmbINE input - output for the motivating example**. Input PAGs and output summary graph for the motivating example, presented in Chapter 1.

As mentioned above, in the absence of statistical errors, all the constraints stemming from all PAGs $\mathcal{P}_i$ are simultaneously satisfiable. In practical settings however, it is possible that some of the PAGs have some erroneous features due to statistical errors, and these features can lead to conflicting constraints. To tackle this problem, Algorithm 4 uses the following technique: For every observed feature, instead of imposing the implied constraints on the formula $\Phi$, the algorithm adds a bi-conditional connecting the feature to the constraints. For example, if $X$ and $Y$ are found adjacent in $\mathcal{P}_i$, then instead of adding the constraints $\exists p_{XY} : inducing(X, Y, i)$ to $\Phi$, we add the bi-conditional $adjacent(X, Y, \mathcal{P}_i) \leftrightarrow \exists p_{XY} : inducing(X, Y, i)$. The antecedents of the conditionals are stored in a list of literals $\mathcal{F}$. The conflict resolution strategy is then imposed on this list of literals, selecting a subset $\mathcal{F}'$ that results in a satisfiable SAT formula $\Phi \wedge \mathcal{F}'$. The formula $\Phi \wedge \mathcal{F}'$ is expressed in Conjunctive Normal Form (CNF) so it can be input to standard SAT solvers.

Recall that the propositional variables of $\Phi$ correspond to the features of the actual underlying SMCM (its edges and endpoints). Some of these variables have the same value in all the possible truth-setting assignments of $\Phi \wedge \mathcal{F}'$, meaning the respective features are invariant in all possibly underlying SMCMs. Such variables are called **backbone** variables of $\Phi \wedge \mathcal{F}'$ (Hyttinen et al., 2013). The actual value of a backbone variable is called the polarity of the variable. For sake of brevity, we say an edge or endpoint has polarity 0/1 if the corresponding variable is a backbone variable in $\Phi \wedge \mathcal{F}'$ and has polarity 0/1. Based on the backbone of $\Phi \wedge \mathcal{F}'$, the final step of COmbINE is to construct the summary graph $\mathcal{S}$. $\mathcal{S}$ has the following types of edges and endpoints:

- **Solid Edges:** Edges in $\mathcal{H}$ that have polarity 1 in $\Phi \wedge \mathcal{F}'$, meaning that they are present in all possibly underlying SMCMs.

- **Absent Edges:** Edges that are not in $\mathcal{H}$ or edges in $\mathcal{H}$ that have polarity 0 in $\Phi \wedge \mathcal{F}'$, meaning that they are absent in all possibly underlying SMCMs.

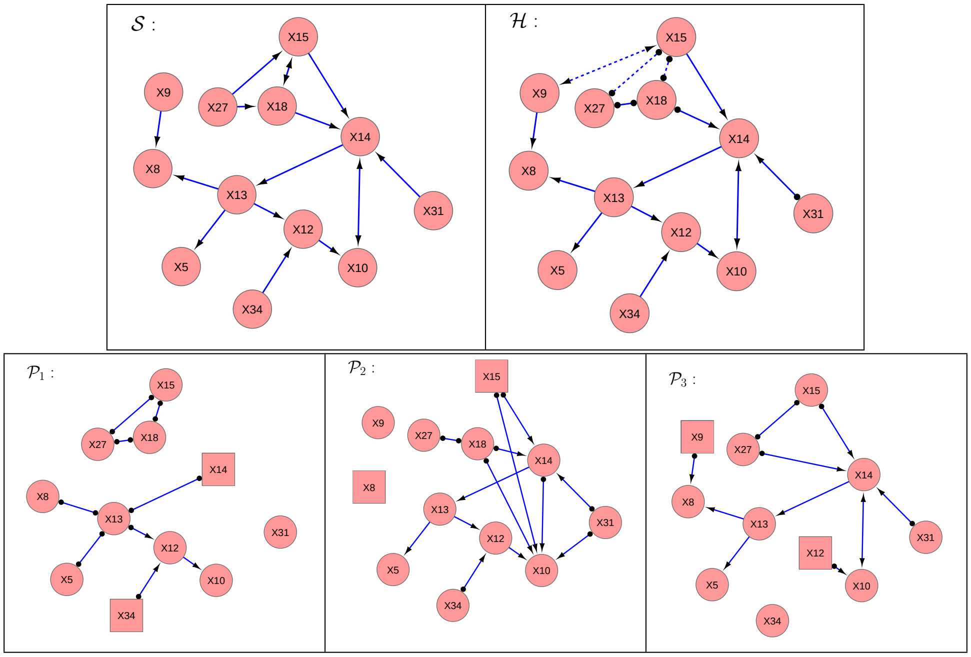


Figure 3.4: **An example of COmbINE input - output**. Graph $\mathcal{S}$ is the actual, data-generating, underlying SMCM over 12 variables. PAGs $\mathcal{P}_1, \mathcal{P}_2$ and $\mathcal{P}_3$ are the output of FCI ran with an oracle of conditional independence on three different marginals of $\mathcal{G}$. $\mathcal{H}$ is the output of COmbINE algorithm. The sets of latent variables (with respect to the union of observed variables) per data set are: $\mathbf{L}_1 = \{X9\}$, $\mathbf{L}_2 = \{\emptyset\}$, $\mathbf{L}_3 = \{X18\}$. The sets of manipulated variables (annotated as rectangle nodes instead of circles in the respective graphs) are: $\mathbf{I}_1 = \{X14, X34\}$, $\mathbf{I}_2 = \{X15, X8\}$, $\mathbf{I}_3 = \{X9, X12\}$. Notice that $X10$ and $X31$ are adjacent in $\mathcal{P}_2$, but not in $\mathcal{P}_1$ or $\mathcal{P}_3$. This happens because there exists an inducing path in the underlying SMCM ($X31 \longrightarrow X14 \longleftrightarrow X10$ in $\mathcal{S}$) that is "broken" by the manipulation of $X14$ and $X12$, respectively. Also notice a dashed edge between $X9$ and $X15$, which cannot be excluded since the variables have never been observed unmanipulated together. Even if the link existed, it would be destroyed in both $\mathcal{P}_2$ and $\mathcal{P}_3$, where both variables are observed. All graphs were visualized in Cytoscape (Smoot et al., 2011).

- **Dashed Edges:** Edges in $\mathcal{H}$ that are not backbone variables in $\Phi \wedge \mathcal{F}'$, meaning that there exists at least one possibly underlying SMCM where this edge is present and one where this edge is absent.
- **Solid Endpoints:** Endpoints in $\mathcal{H}$ that are backbone variables in $\Phi \wedge \mathcal{F}'$, meaning that this orientation is invariant in all possibly underlying SMCMs.
- **Dashed (circled) Endpoints:** Endpoints in $\mathcal{H}$ that are not backbone variables in $\Phi \wedge \mathcal{F}'$, meaning that there exists at least one possibly underlying SMCM where this orientation does not hold.

We use the term **solid features** of the summary graph to denote the set of solid edges, absent edges and solid endpoints of the summary graph.

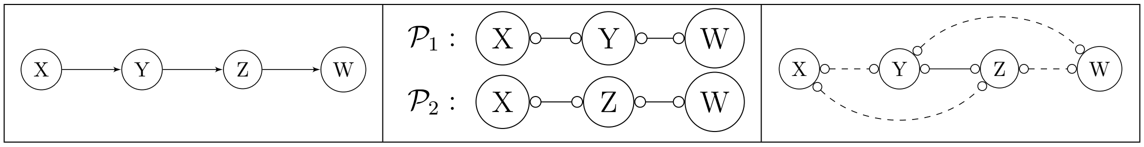


Figure 3.5: **A detailed example of a non-trivial inference**. From left to right: The true underlying SMCM over variables $X$, $Y$, $Z$, $W$; PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$ over $\{X, Y, W\}$ and $\{X, Z, W\}$, respectively; The output $\mathcal{H}$ of Algorithm 2 ran with an oracle of conditional independence. Notice that, the edges in $\mathcal{P}_1$ can not both simultaneously occur in a consistent SMCM $\mathcal{S}$: This would make $X{\circ\!\!-\!\!\circ}Y{\circ\!\!-\!\!\circ}W$ an inducing path for $X$ and $W$ with respect to $\mathbf{L}_2 = \{Y\}$ and contradict the features of $\mathcal{P}_2$, where $X$ and $W$ are not adjacent. Similarly, $X{\circ\!\!-\!\!\circ}Z{\circ\!\!-\!\!\circ}W$ cannot occur in any possibly underlying SMCM $\mathcal{S}$. The only possible edge structures that explain all the observed adjacencies and definite non colliders are $X{\circ\!\!-\!\!\circ}Y{\circ\!\!-\!\!\circ}Z{\circ\!\!-\!\!\circ}W$ or $X{\circ\!\!-\!\!\circ}Z{\circ\!\!-\!\!\circ}Y{\circ\!\!-\!\!\circ}W$. Either way, $Y$ and $Z$ share an edge in all consistent SMCMs, and the algorithm will predict a solid edge between $Y$ and $Z$, even if the two have not been measured in the same data set. This example is discussed in detail in (Tsamardinos et al., 2012).

Overall, *Algorithm 2 takes as input a set of data sets and a list of parameters and outputs a summary graph that has all invariant edges and orientations of the SMCMs that satisfy as many constraints as possible (according to some strategy).* The algorithm is capable of non-trivial inferences, like for example the presence of a solid edge among variables never measured together. Figures 3.3 3.4 and 3.5 illustrate the output of Algorithm 2, along with the corresponding input PAGs.

## 3.3 Soundness and Completeness

We claim that, given an oracle of conditional independence, the SAT-generating procedure described in Algorithm 4 results in a SAT instance $\Phi \wedge \mathcal{F}$ that is satisfied by all and only the possibly underlying SMCMs for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$ (i.e., every SMCM that entails the exact same conditional independencies as those obtained by the oracle for every experiment, after the removal of edges incoming to the manipulated variables). Lemma 3.3.3 proves that the every possibly underlying SMCM satisfies $\Phi \wedge \mathcal{F}$, while Lemma 3.3.5 proves that if $\mathcal{S}$ is a mixed graph satisfying $\Phi \wedge \mathcal{F}$, $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.

In all subsequent Lemmas, theorems and proofs we employ the assumptions and notation presented at the beginning of this chapter (Assumptions A1-A3 and notation presented beneath them). We also assume the algorithms are run with an oracle of conditional independence and infinite maximum conditioning set size and maximum path length.

Recall that, Theorem 3.0.1 proves that a $\mathcal{S}$ is possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^n$ and $\{\mathbf{I}_i\}_{i=1}^N$ if and only if the result of manipulating $\mathbf{I}_i$, adding necessary edges to create a Markov equivalent MAG and then marginalizing out variables in $\mathbf{L}_i$ produces a MAG $\mathcal{M}_i$ that belongs to the Markov equivalence class represented by $\mathcal{P}_i$ for all experiments.

The following Lemma proves that no inducing and ancestral paths present in the true underlying SMCM are ruled out during the construction of the initial search graph, and is necessary for subsequent proofs. We prove that $\mathcal{H}_{in}$ has a superset of edges and a subset of orientations compared to $\mathcal{S}$.

**Lemma 3.3.1** *If $\mathcal{H}_{in}$ is the initial search graph returned by Algorithm 3 for $\{\mathcal{P}_i\}_{i=1}^N$, and $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$, then the following hold: If p is an ancestral path in $\mathcal{S}$, then p is a possibly ancestral path in $\mathcal{H}_{in}$. Similarly, if p is an inducing path with respect to $\mathbf{L}$ in $\mathcal{S}$, then p is a possibly inducing path with respect to $\mathbf{L}$ in $\mathcal{H}_{in}$.*

**Proof** We will first prove that $\mathcal{H}_{in}$ has a superset of edges compared to $\mathcal{S}$, and therefore any path in $\mathcal{S}$ is a path also in $\mathcal{H}_{in}$. If $X$ and $Y$ are adjacent in $\mathcal{S}$, then one of the following holds:

1. $\exists i$ s.t. $X, Y \in \mathbf{O}_i \setminus \mathbf{I}_i$. Then the edge is present in $\mathcal{S}^{\mathbf{I}_i}$, and $X$ and $Y$ are adjacent in $\mathcal{P}_i$: the edge is added to $\mathcal{H}_{in}$ in Line 3 of Algorithm 3.
2. $\nexists i$ s.t. $X, Y \in \mathbf{O}_i \setminus \mathbf{I}_i$. Then the edge is added to $\mathcal{H}_{in}$ in Line 8 of Algorithm 3.

Therefore, every edge in $\mathcal{S}$ is present also in $\mathcal{H}_{in}$. We must also prove that no orientation in $\mathcal{H}$ is oriented differently in $\mathcal{S}$: $\mathcal{H}_{in}$ has only arrowhead orientations, so we must prove that, if $X \star\!\!\!\longrightarrow Y$ in $\mathcal{H}_{in}$ and $X$ and $Y$ are adjacent in both graphs, $X \star\!\!\!\longrightarrow Y$ in $\mathcal{S}$.

Arrowheads are added to $\mathcal{H}_{in}$ in Lines 5, 9 or 10 of the Algorithm. Arrowheads added in Line 5 occur in all $\mathcal{P}_i$. If X$\star\!\!\!\longrightarrow Y$ in any $\mathcal{P}_i$, this means that $Y$ is not an ancestor of $X$ in $\mathcal{S}^{\mathbf{I}_i}$. Assume that $X \longleftarrow Y$ in $\mathcal{S}$: If $X$ in $\mathbf{I}_i$, the edge would be absent in $\mathcal{S}^{\mathbf{I}_i}$ and $\mathcal{P}_i$. If $X \notin \mathbf{I}_i$, $X$ would be ancestor of $Y$ in $\mathcal{S}^{\mathbf{I}_i}$, which is a contradiction. Therefore, if $X$ and $Y$ are adjacent in $\mathcal{S}$, X$\star\!\!\!\longrightarrow Y$ in $\mathcal{S}$.

Arrows added to $\mathcal{H}_{in}$ in Lines 9 and 10 correspond to cases where an edge is not present in any $\mathcal{P}_i$, $\nexists i$ s.t. $X, Y \in \mathbf{O}_i \setminus \mathbf{I}_i$, but $\exists i$ s.t. $X, Y \in \mathbf{O}_i$, $X \in \mathbf{I}_i$ and $Y \notin \mathbf{I}_i$. Then an arrow is added towards $X$. Assume the opposite holds: $X \longrightarrow Y$ in $\mathcal{S}$, then $X \longrightarrow Y$ in $\mathcal{S}^{\mathbf{I}_i}$, and since both variables are observed in experiment $i$ the edge would be present in $\mathcal{P}_i$, which is a contradiction. Thus, if the edge is present in $\mathcal{S}$, the edge is oriented into $X$.

Thus, $\mathcal{H}_{in}$ has a superset of edges of $\mathcal{S}$, and for any edge present in both graphs, the orientations are the same. Thus, if $p$ is an ancestral path in $\mathcal{S}$, then $p$ is a possibly ancestral path in $\mathcal{H}_{in}$. Similarly, if $p$ is a possibly inducing path with respect to $\mathbf{L}$ in $\mathcal{S}$, then $p$ is a possibly inducing path with respect to $\mathbf{L}$ in $\mathcal{H}_{in}$.

We can now prove that if a SMCM $\mathcal{S}$ entails all and only the observed conditional independencies for all experiments (and is therefore a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$), then $\mathcal{S}$ satisfies $\Phi \wedge \mathcal{F}$. We say that $\mathcal{S}$ *satisfies a constraint* $\phi$ if the truth-values assigned to *edge*, *arrow* and *tail* variables by their corresponding configuration in $\mathcal{S}$ satisfies $\phi$. To simplify the proof, we first prove the following lemma:

**Lemma 3.3.2** *If $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$, and $X \longrightarrow Y$ is in $\mathcal{P}_i$, then $\mathcal{S}$ satisfies $ancestor(X, Y, i)$. Similarly, if $X \circ\!\!\!\longrightarrow Y$ is in $\mathcal{P}_i$, then $\mathcal{S}$ satisfies $\neg ancestor(Y, X, i)$.*

**Proof** By Theorem 3.0.1 SMCMtoMAG $(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} \in \mathcal{P}_i$. Thus, if $X \longrightarrow Y$ is in $\mathcal{P}_i$, then $X$ is an ancestor of $Y$ in $\mathcal{S}^{\mathbf{I}_i}$ (there exists an ancestral path from $X$ to $Y$ in $\mathcal{S}^{\mathbf{I}_i}$). Let $p_1, \ldots, p_M$ be the possibly ancestral paths (there exists at least one: if $X \longrightarrow Y$ in $\mathcal{P}_i$, then $X {\star}\!\!\longrightarrow\!\!{\star} Y$ is a possibly inducing path in $\mathcal{H}_{in}$) from $X$ to $Y$ in $\mathcal{H}_{in}$. The constraint $ancestor(X, Y, i)$ is realized in $\Phi \wedge \mathcal{F}$ as $ancestor(Y, X, i) \wedge [ancestor(Y, X, i) \leftrightarrow ancestral(p_1, i) \vee ancestral(p_2, i) \cdots \vee ancestral(p_M, i)]$. This is equivalent to $ancestral(p_1, i) \vee ancestral(p_2, i) \cdots \vee ancestral(p_M, i)$. If a path is ancestral in $\mathcal{S}^{\mathbf{I}_i}$, the path is also ancestral in $\mathcal{S}$. By Lemma 3.3.1, if a path is ancestral in $\mathcal{S}$, the path is possibly ancestral in $\mathcal{H}_{in}$. Hence, at least one of $p_1, \ldots, p_M$ is ancestral in $\mathcal{S}^{\mathbf{I}_i}$, and $\mathcal{S}$ satisfies $ancestor(X, Y, i)$.

If $X \circ\!\!\longrightarrow Y$ is in $\mathcal{P}_i$, then, since SMCMtoMAG $(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} \in \mathcal{P}_i$, there can be no ancestral path from $Y$ to $X$ in $\mathcal{S}^{\mathbf{I}_i}$). Let $p_1, \ldots, p_M$ be the possibly ancestral paths (if any) from $Y$ to $X$ in $\mathcal{H}_{in}$. The constraint $\neg ancestral(Y, X, i)$ is realized in $\Phi \wedge \mathcal{F}$ as $\neg ancestor(Y, X, i) \wedge [ancestor(Y, X, i) \leftrightarrow ancestral(p_1, i) \vee ancestral(p_2, i) \cdots \vee ancestral(p_M, i)]$. This is equivalent to $\neg ancestral(p_1, i) \wedge \neg ancestral(p_2, i) \cdots \wedge \neg ancestral(p_M, i)$. None of these paths are ancestral in $\mathcal{S}^{\mathbf{I}_i}$, therefore $\mathcal{S}$ satisfies $ancestor(X, Y, i)$.

We can now prove that any possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$ satisfies $\Phi \wedge \mathcal{F}$.

**Lemma 3.3.3** *For an oracle of conditional independence, if $\mathcal{S}$ is a possibly underlying model for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$, and $\Phi \wedge \mathcal{F}$ is the conjunction of the outputs of Algorithm 4, $\mathcal{S}$ satisfies $\Phi \wedge \mathcal{F}$.*

**Proof** By Theorem 3.0.1, since $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$, $\mathcal{M}_i =$ SMCMtoMAG$(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} \in \mathcal{P}_i \quad \forall i$.

1. **Constraints added in Lines 8, 9 of Algorithm 4**. These constraints are satisfied since $\mathcal{S}$ is an acyclic mixed graph.

2. **Adjacency constraints added in Lines 4, 5, 6 of Algorithm 4**. Assume that for a pair of variables $X$, $Y$ adjacent in $\mathcal{P}_i$, there exist $M$ possibly inducing paths in $\mathcal{H}_{in}$, namely $p_1, , \ldots, p_M$. For this adjacency, the following constraint is added in $\Phi \wedge \mathcal{F}$ in Lines 4 and 5 of Algorithm 4:

   $$adjacent(X, Y, \mathcal{P}_i) \wedge [adjacent(X, Y, \mathcal{P}_i) \leftrightarrow inducing(p_1, i) \vee \cdots \vee inducing(p_M, i)],$$

   which is equivalent to
   $$inducing(p_1, i) \vee \cdots \vee inducing(p_M, i).$$

   Since $\mathcal{M}_i \in \mathcal{P}_i$, $X$ and $Y$ are adjacent in $\mathcal{M}_i$. By Proposition 2.3.3 there exists an inducing path $p^*$ between $X$ and $Y$ with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$. By Lemma 3.3.1, this path is a possibly inducing path in $\mathcal{H}_{in}$, thus, $\exists i \in [1, \ldots, M]$ such that $p^* = p_i$. Thus, the constraint $inducing(p_1, i) \vee \cdots \vee inducing(p_M, i)$ is satisfied by $\mathcal{S}$.

   Similarly, if $X$ and $Y$ are not adjacent in $\mathcal{P}_i$, the constraint

   $$\neg adjacent(X, Y, \mathcal{P}_i) \wedge [adjacent(X, Y, \mathcal{P}_i) \leftrightarrow inducing(p_1, i) \vee \cdots \vee inducing(p_M, i)]$$

   is added to $\Phi \wedge \mathcal{F}$ in Lines 4 and 6 of Algorithm 4. The constraint is equivalent to

   $$\neg inducing(p_1, i) \wedge \cdots \wedge \neg inducing(p_M, i).$$

Since $X$ and $Y$ are not adjacent in $\mathcal{M}_i$, by Proposition 2.3.3 there exists no inducing path with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$. Thus, none of the paths (if any) $p_1, \ldots, p_M$ is inducing with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$, and the constraint $\neg inducing(p_1, i) \wedge \cdots \wedge \neg inducing(p_M, i)$ is satisfied by $\mathcal{S}$.

3. **Unshielded (non) collider constraints added in Lines 13,14, 15,16 of Algorithm 4.** For an unshielded collider $X{\star\!\!-\!\!\star}Y\ {\star\!\!-\!\!\star}Z$ in $\mathcal{P}_i$, the constraint

$$\begin{aligned}&col(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge\\ &\quad\big[col(\langle X,Y,Z\rangle,\mathcal{P}_i)\rightarrow unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge collider(\langle X,Y,Z\rangle,\mathcal{P}_i)\big],\end{aligned}$$

which is equivalent to

$$unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge collider(\langle X,Y,Z\rangle,\mathcal{P}_i)$$

is added in Lines 14 and 15. As shown in Figure 3.1,

$$unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)\leftrightarrow adjacent(X,Y,\mathcal{P}_i)\wedge adjacent(Y,Z,\mathcal{P}_i)\wedge\neg adjacent(X,Z,\mathcal{P}_i)$$

and

$$collider(\langle X,Y,Z\rangle,\mathcal{P}_i)\leftrightarrow\neg ancestor(Y,X,i)\wedge\neg ancestor(Y,Z,i)$$

. Since $\mathcal{M}_i \in \mathcal{P}_i$, $X{\star\!\!\rightarrow}Y{\leftarrow\!\!\star}Z$ is an unshielded triple in $\mathcal{M}_i$, $adjacent(X,Y,\mathcal{P}_i)\wedge adjacent(Y,Z,\mathcal{P}_i)\wedge$ $\neg adjacent(X,Z,\mathcal{P}_i)$ is satisfied (as described above for adjacency constraints). Since $X{\star\!\!\rightarrow}Y{\leftarrow\!\!\star}Z$ in $\mathcal{P}_i$, by Lemma 3.3.2 constraints $\neg ancestor(Y,X,i)\wedge\neg ancestor(Y,Z,i)$ are satisfied by $\mathcal{S}$.

For an unshielded definite non collider $X{\star\!\!-\!\!\star}Y\ {\star\!\!-\!\!\star}Z$ in $\mathcal{P}_i$, the constraint

$$\begin{aligned}&dnc(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge\\ &\quad\big[dnc(\langle X,Y,Z\rangle,\mathcal{P}_i)\rightarrow unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge\neg collider(\langle X,Y,Z\rangle,\mathcal{P}_i)\big],\end{aligned}$$

is added in Lines 13 and 16 of Algorithm 4, which is equivalent to

$$unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge\neg collider(\langle X,Y,Z\rangle,\mathcal{P}_i).$$

Since $\mathcal{M}_i \in \mathcal{P}_i$, $X{\star\!\!-\!\!\star}Y\ {\star\!\!-\!\!\star}Z$ is an unshielded triple in $\mathcal{M}_i$, so $unshielded(\langle X,Y,Z\rangle,\mathcal{P}_i)$ is satisfied by $\mathcal{S}$ as described above. Moreover, since either $Y{\rightarrow}X$ in $\mathcal{M}_i$, or $Y{\rightarrow}Z$ in $\mathcal{M}_i$, by Lemma 3.3.2 $ancestor(Y,X,i)\vee ancestor(Y,Z,i)$ is satisfied by $\mathcal{S}$.

4. **Discriminating (non) collider constraints added in Lines 19, 20,21, 22 of Algorithm 4**. If $\langle W,\ldots,X,Y,Z\rangle$ is a discriminating path for $Y$ in $\mathcal{P}_i$, and $Y$ is a collider on the path in $\mathcal{P}_i$, the following constraint is added in $\Phi\wedge\mathcal{F}$ and in Lines 19 and 21 of Algorithm 4:

$$\begin{aligned}&col(\langle X,Y,Z\rangle,\mathcal{P}_i)\wedge\\ &\quad\big[col(\langle X,Y,Z\rangle,\mathcal{P}_i)\rightarrow discriminating(p_{WZ},Y,\mathcal{P}_i)\wedge collider(\langle X,Y,Z\rangle,\mathcal{P}_i)\big],\end{aligned}$$

which is equivalent to

$$discriminating(p_{WZ},Y,\mathcal{P}_i)\wedge collider(\langle X,Y,Z\rangle,\mathcal{P}_i).$$

Since $\mathcal{M}_i \in \mathcal{P}_i$, the path is discriminating for $Y$ in $\mathcal{M}_i$ and the triple is a collider in $\mathcal{M}_i$. The constraint for the discriminating path is analyzed as a conjunction of the individual features ((non) adjacencies and endpoints) of the path as shown in Figure 3.1. Since the path is discriminating in $\mathcal{M}_i$, all these adjacency and ancestry constraints are satisfied by $\mathcal{S}$, by the proof for adjacency constraints and Lemma 3.3.2. In addition, the triple is a collider in $\mathcal{M}_i$, thus $\mathit{collider}(\langle X, Y, Z\rangle, \mathcal{P}_i)$ is satisfied by $\mathcal{S}$ as described for unshielded colliders.

Similarly, if $\langle W, \ldots, X, Y, Z\rangle$ is a discriminating path for $Y$ in $\mathcal{P}_i$, and $Y$ is a definite non collider on the path in $\mathcal{P}_i$, the following constraint is added in $\Phi \wedge \mathcal{F}$ and in Lines 20 and 22 of Algorithm 4:

$$\begin{aligned}&\mathit{dnc}(\langle X, Y, Z\rangle, \mathcal{P}_i)\wedge\\ &\quad\big[\mathit{dnc}(\langle X, Y, Z\rangle, \mathcal{P}_i) \rightarrow \mathit{discriminating}(p_{WZ}, Y, \mathcal{P}_i) \wedge \neg \mathit{collider}(\langle X, Y, Z\rangle, \mathcal{P}_i)\big],\end{aligned}$$

which is equivalent to

$$\mathit{discriminating}(p_{WZ}, Y, \mathcal{P}_i) \wedge \neg \mathit{collider}(\langle X, Y, Z\rangle, \mathcal{P}_i).$$

Since $\mathcal{M}_i \in \mathcal{P}_i$, the path is discriminating for $Y$ in $\mathcal{M}_i$ and the triple is a non-collider in $\mathcal{M}_i$. The constraint for the discriminating path satisfied by $\mathcal{S}$ as described above. In addition, the triple is a non-collider in $\mathcal{M}_i$, thus $\neg\mathit{collider}(\langle X, Y, Z\rangle, \mathcal{P}_i)$ is satisfied by $\mathcal{S}$ as described for unshielded definite non colliders.

Thus, $\mathcal{S}$ satisfies all constraints in $\Phi \wedge \mathcal{F}$.

To prove completeness for Algorithm 4, we must show that the opposite also holds: If $\mathcal{S}$ is a truth-setting assignment of $\Phi\wedge\mathcal{F}$, $\mathcal{S}$ entails all and only the conditional independencies observed in $\{\mathcal{J}_i\}_{i=1}^N$ for each experiment. According to Theorem 3.0.1, we need to show that any truth setting assignment of $\Phi\wedge\mathcal{F}$ results, in each experiment $i$ (after the respective procedures of manipulation, conversion to MAG and marginalization) in a MAG $\mathcal{M}_i$ that belongs to the Markov equivalence class represented by $\mathcal{P}_i$. Thus, we need to show that $\mathcal{M}_i$ has the same adjacencies and colliders with order as any MAG $\mathcal{M}' \in \mathcal{P}_i$. Proving that $\mathcal{M}_i$ and any $\mathcal{M}' \in \mathcal{P}_i$ have the same adjacencies is straight-forward. We then use induction to the order of the triple to show that the two MAGs also share the same colliders with order. The following lemma proves that discriminating paths with order are present in all members of the equivalence class, and therefore they are (definite) discriminating paths with order in $\mathcal{P}_i$ (Lemma 3.3.4.) Thus, all (non) colliders with order in $\mathcal{P}_i$ are identified and added to the SAT formula in Lines 19 and 20 of Algorithm 4.

**Lemma 3.3.4** *If $p = \langle W, V_1, \ldots, V_n, Y, Q\rangle$ is a discriminating path with order $r$ in $\mathcal{M}$, then the path is a discriminating path with order $r$ in $\mathcal{P} = [\mathcal{M}]$.*

**Proof** We will show that the path is a discriminating path with order $r$ in any $\mathcal{M}' \in \mathcal{P}$. Since $\mathcal{M}'$ and $\mathcal{M}$ are Markov equivalent, the two share the same colliders with order. Thus, every triple $\langle V_{i-1}, V_i, V_{i+1}\rangle$ is a collider with order in $\mathcal{M}$. Lemma 3.10 in Ali et al. (2009) states that if a path $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ is discriminating for $Y$ in a MAG $\mathcal{M}$, then in any Markov equivalent MAG $\mathcal{M}'$ in which $V_i$ are colliders on the same path, $V_i \rightarrow Q$ in $\mathcal{M}'$ for $i = 1, \ldots, N$, and therefore the path is discriminating with order $r$ in $\mathcal{M}'$. Thus, the path is discriminating with order $r$ in all members of $[\mathcal{M}]$. It is therefore a discriminating path with order $r$ in $\mathcal{P}$.

We can now prove that any truth-setting assignment for $\Phi \wedge \mathcal{F}$ corresponds to a SMCM $\mathcal{S}$ that is possibly underlying for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.

**Lemma 3.3.5** *For an oracle of conditional independence, if $\Phi \wedge \mathcal{F}$ is the conjunction of the outputs of Algorithm 4, and $\mathcal{S}$ a mixed graph that satisfies $\Phi \wedge \mathcal{F}$, then $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.*

**Proof** We need to prove that (a) $\mathcal{S}$ is an acyclic mixed graph and (b) $\mathcal{M}_i = \text{SMCMtoMAG}(\mathcal{S}^{\mathbf{I}_i})[_{\mathbf{L}_i} \in \mathcal{P}_i \quad \forall i$. To prove the latter, we need to prove that for each $i$, if $\mathcal{M}' \in \mathcal{P}_i$, $\mathcal{M}_i$ and $\mathcal{M}'$ are Markov equivalent. Thus, we must show that $\mathcal{M}_i$ and $\mathcal{M}'$ share the same edges and colliders with order.

- $\mathcal{S}$ **is a SMCM:** $\mathcal{S}$ satisfies the constraints added in Lines 8 and 9 respectively. Therefore, $\mathcal{S}$ has no tail-tail edges, every endpoint is an arrow or a tail (not exclusively) and $\mathcal{S}$ has no directed cycles.
- $\mathcal{M}_i$ **and** $\mathcal{M}'$ **share the same edges**: If $X$ and $Y$ are adjacent in $\mathcal{M}'$, then $X$ and $Y$ are adjacent in $\mathcal{P}_i$. $\mathcal{S}$ satisfies the constraints added in Line 4 of Algorithm 4, therefore there exists an inducing path with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$. Thus, $X$ and $Y$ are adjacent in $\mathcal{M}_i$. If $X$ and $Y$ are not adjacent in $\mathcal{M}'$, $X$ and $Y$ are not adjacent in $\mathcal{P}_i$ and by the same constraints there exists no inducing path with respect to $\mathbf{L}_i$ in $\mathcal{S}^{\mathbf{I}_i}$, therefore $X$ and $Y$ are not adjacent in $\mathcal{M}_i$.
- $\mathcal{M}_i$ **and** $\mathcal{M}'$ **share the same colliders with order:** We will prove this by induction to order $r$: For order $= 0$, if $\langle X, Y, Z\rangle$ is an unshielded collider in $\mathcal{M}'$, the triple is an unshielded collider in $\mathcal{P}_i$. Since $\mathcal{M}'$ and $\mathcal{M}_i$ share the same edges, $X \star\!\!-\!\!\star Y \star\!\!-\!\!\star Z$ is an unshielded triple in $\mathcal{M}_i$. $\mathcal{S}$ satisfies the constraints added in Line 13 of Algorithm 4, and therefore $Y$ is not an ancestor of $X$ nor $Z$ in $\mathcal{S}^{\mathbf{I}_i}$. Thus, $X \star\!\!\rightarrow Y \leftarrow\!\!\star Z$ in $\mathcal{M}_i$. If the triple is an unshielded collider in $\mathcal{M}_i$, then the triple is unshielded in $\mathcal{M}'$. If the triple is a non-collider in $\mathcal{M}'$, then $\mathcal{S}$ satisfies the constraints added in Line 14 of Algorithm 4, and $Y$ is an ancestor of either $X$ or $Z$ in $\mathcal{S}^{\mathbf{I}_i}$. But then the triple is a non-collider in $\mathcal{M}_i$, which is a contradiction. Thus, $\mathcal{M}_i$ and $\mathcal{M}'$ share the same colliders with order 0.

  For the induction step, we assume that $\mathcal{M}_i$ and $\mathcal{M}'$ share the same colliders with order $s < r$. We will show that the two MAGs also share the same colliders with order $r$. We will first show that a path $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ is discriminating for $\langle V_n, Y, Q\rangle$ with order $r$ in $\mathcal{M}_i$ iff the path is discriminating for $\langle V_n, Y, Q\rangle$ with order $r$ in $\mathcal{M}'$.

  If $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ is discriminating with order $r$ in $\mathcal{M}'$, by Lemma 3.3.4 the path is discriminating with order $r$ in $\mathcal{P}_i$. $\mathcal{S}$ satisfies the constraints added in Lines 20 and 19 and therefore the path is discriminating in $\mathcal{M}_i$. Moreover, every triple on the path is a collider with order $< r$ in $\mathcal{M}'$ and by the induction hypothesis $\mathcal{M}'$ and $\mathcal{M}_i$ share the same colliders with order $< r$, thus the path has order $r$ in $\mathcal{M}_i$.

  If $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ is discriminating with order $r$ in $\mathcal{M}_i$, then, by the induction hypothesis, every triple on the path is a collider with the same order $< r$ in $\mathcal{M}'$. We will show that $V_i \rightarrow Q \quad \forall i$, and therefore $\langle W, V_1, \ldots, V_n, Y, Q\rangle$ is a discriminating path with order $r$ in $\mathcal{M}'$.

  The proof is similar to that of Lemma 3.10 in Ali et al. (2009). We will use induction on $i$. First, consider the $(V_1, Q)$ edge in $\mathcal{M}'$. If $V_1 \leftarrow\!\!\star Q$, then $W \star\!\!\rightarrow V_1 \leftarrow\!\!\star Q$ forms a collider with order 0 in $\mathcal{M}'$, but an non-collider with order 0 in $\mathcal{M}_i$, which is a contradiction. Thus, $V_1 \longrightarrow Q$ in $\mathcal{M}'$.

  Suppose that $V_j \longrightarrow Q$ for $1 \leq j \leq i$ in $\mathcal{M}'$. Then, the path $\langle W, V_1, \ldots, V_i, Q\rangle$ forms a discriminating path for $V_i$ with the same order $< r$ in both graphs, and $\langle V_{i-1}, V_i, Q\rangle$ is a non-collider in $\mathcal{M}_i$. By Lemma 3.3.4, the path is a discriminating path with order in $\mathcal{P}_i$, and therefore

$\Phi \wedge \mathcal{F}$ includes discriminating path constraints for this path added in Lines 19 and 21 or 20 and 22 of Algorithm 4. Thus, the triple can only be a non-collider in $\mathcal{M}_i$ if it is a non-collider in $\mathcal{M}'$. Since $V_{i-1} \leftrightarrow V_i$ in $\mathcal{M}'$, $V_i \longrightarrow Q \quad \forall i$ and the path is discriminating in $\mathcal{M}'$ with order $r$.

We have shown that $\mathcal{M}_i$ and $\mathcal{M}'$ share the same discriminating paths with order $r$. It is now easy to show that a triple is a collider with order $r$ in $\mathcal{M}'$ iff it is a collider with order $r$ in $\mathcal{M}_i$. If $\langle V_n, Y, Z \rangle$ is a collider with order $r$ in $\mathcal{M}'$, then there exists a discriminating path with order $r$ in both graphs and in $\mathcal{P}_i$. Thus, $\mathcal{S}$ satisfies the constraints added in Lines 19 and 21 of Algorithm 4, by which $Y$ is not an ancestor of $V_n$ nor $Q$ in $\mathcal{S}^{\mathbf{I}_i}$, and therefore the triple is a collider in $\mathcal{M}_i$, and it has order at most $r$. But by the induction hypothesis, the $\mathcal{M}'$ and $\mathcal{M}_i$ share the same colliders with order $< r$, thus the triple has order $r$ in $\mathcal{M}_i$. Similarly, if the triple is a collider with order $r$ in $\mathcal{M}_i$, there exists a discriminating path with order $r$ in $\mathcal{M}$; and therefore in $\mathcal{P}_i$. Thus, $\mathcal{S}$ satisfies the constraints added in Lines 19 and 21 of Algorithm 4 or in Lines 20 and 22 of algorithm 4. Hence, the triple must be in $\mathcal{M}'$, otherwise the triple would be a non-collider in $\mathcal{M}_i$. In addition, the triple has order at most $r$ in $\mathcal{M}'$ and by the induction hypothesis the triple can not have order $< r$ in $\mathcal{M}'$, so the triple has order $r$ in $\mathcal{M}'$. Thus, $\mathcal{M}'$ and $\mathcal{M}_i$ share the same colliders with order.

Thus, if $\mathcal{S}$ a mixed graph that satisfies $\Phi \wedge \mathcal{F}$, then $\mathcal{S}$ is a SMCM and SMCMtoMAG($\mathcal{S}^{\mathbf{I}_i}$)$[_{\mathbf{L}_i} \in \mathcal{P}_i \quad \forall i$, so by Theorem 3.0.1, $\mathcal{S}$ is a possibly underlying SMCM for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.

We can now prove soundness and completeness of Algorithm 2:

**Theorem 3.3.6 (Soundness and completeness of Algorithm 2)** *If $\mathcal{H}$ is the output of Algorithm 2, then the following hold:*
***Soundness****: If a feature (edge, absent edge, endpoint) is* solid *in $\mathcal{H}$, then this feature is present in* all *SMCMs that are possibly underlying for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$.*
***Completeness****: If a feature is present in* all *SMCMs that are possibly underlying for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$, the feature is solid in $\mathcal{H}$.*

**Proof** *Soundness:* Solid features correspond to backbone variables. By Lemma 3.3.3 every possibly underlying SMCM $\mathcal{S}$ for $\{\mathcal{J}_i\}_{i=1}^N$ and $\{\mathbf{I}_i\}_{i=1}^N$ satisfies the final formula $\Phi \wedge \mathcal{F}$. Thus, if a core variable has the same value in all the possible truth-setting assignments of $\Phi \wedge \mathcal{F}$, this feature is present in all possibly underlying SMCMs. *Completeness:* By Lemma 3.3.5 the final formula $\Phi \wedge \mathcal{F}$ of Algorithm 2 is satisfied only by possibly underlying SMCMs. Thus, if a core variable is present in *all* consistent SMCMs, the corresponding core variable will be a backbone variable for $\Phi \wedge \mathcal{F}$.

## 3.4 Related Work

Methods for causal discovery have been, for the most part, limited to the analysis of a single data set. However, the great advancement of intervention and data collection technology has led to a vast increase of available data sets, both observational and experimental. Therefore, over the last few years, there have been a number of works that focus on causal discovery from multiple sources. Algorithms in that area may differ in the formalism they use to model causality or in the type of heterogeneity in the studies they co-analyze. In any case, the goal is always to discover the single underlying data-generating causal mechanism.

One group of algorithms focuses on combining observational data that measure overlapping variables. The idea that combining causal graphs can lead to additional inferences was introduced in Danks (2005). Sound and complete procedures for learning the common characteristics of MAGs from data sets measuring overlapping variables were then presented in Tillman et al. (2008) and Triantafillou et al. (2010). Tillman et al. (2008) handles conflicts by ignoring conflicting evidence, while the method presented in Triantafillou et al. (2010) only works with an oracle of conditional independence. Tillman and Spirtes (2011) present an algorithm for the same task that handles a limited type of conflicts (those concerning p-values for the same pair of variables stemming from different data sets) by combining the p-values for conditional independencies that are testable in more than one data sets. Claassen and Heskes (2010b) present a sound, but not complete, algorithm for causal structure learning from multiple independence models over overlapping variables by transforming independencies into a set of causal ancestry rules.

Another line of work deals with learning causal models from multiple experiments. Cooper and Yoo (1999) use a Bayesian score to combine experimental and observational data in the context of causal Bayesian networks. Hauser and Bühlmann (2012) extend the notion of Markov equivalence for DAGs to the case of interventional distributions arising from multiple experiments, and propose a learning algorithm. Tong and Koller (2001) and Murphy (2001) use Bayesian network theory to propose experiments that are most informative for causal structure discovery. Eberhardt and Scheines (2007) and Eaton and Murphy (2007b) discuss how some other types of interventions can be modeled and used to learn Bayesian networks. Hyttinen et al. (2012a) provides an algorithm for learning linear cyclic models from a series of experiments, along with sufficient and necessary conditions for identifiability. This method admits latent confounders but uses linear structural equations to model causal relations and is therefore inherently limited to linear relations. Meganck et al. (2006) propose learning SMCMs by learning the Markov equivalence classes of MAGs from observational data and then designing the experiments necessary to convert it to a SMCM.

Finally, there is a limited number of methods that attempt to co-analyze data sets measuring overlapping variables under different experimental conditions. In Hyttinen et al. (2012b) the authors extend the methods of Hyttinen et al. (2012a) to handle overlapping variables, again under the assumption of linearity. Hyttinen et al. (2013) propose a constraint-based algorithm for learning causal structure from different manipulations of overlapping variable sets. The method works by transforming the observed $m$-connection and $m$-separation constraints into a SAT instance. The method uses a path analysis heuristic to reduce the number of tests translated into path constraints. Causal insufficiency is allowed, as well as feedback cycles. However, this method cannot handle conflicts and therefore relies on an oracle of conditional independence. Moreover, the method can only scale up to about 12 variables. Claassen and Heskes (2010a) present an algorithm for learning causal models from multiple experiments; the experiments here are not hard manipulations, but general experimental conditions, modeled like variables that have no parents in the graph but can cause other variables in some of the conditions.

Hyttinen et al. (2013) presented an algorithm similar to COmbINE  named SAT-based causal structure discovery (SBCSD). SBCSD is also capable of learning causal structure from manipulated data-sets over overlapping variable sets. In addition, if linearity is assumed, it can admit feedback cycles. SBCSD also uses similar techniques for converting conditional (in)dependencies into a SAT

| # variables | # max parents | Running time<br>**Median** (5 %ile, 95 %ile) | | | Completed instances/<br>total instances | | |
|---|---|---|---|---|---|---|---|
| | | COmbINE | SBCSD | SBCSD′ | COmbINE | SBCSD | SBCSD′ |
| 10 | 3 | **17**(1, 113) | **149**(14, 470)* | **91**(30, 369)* | 50/50 | 30/50 | 48/50 |
| | 5 | **80**(4, 1192) | **365**(133, 500)* | **264**(68, 554)* | 50/50 | 16/50 | 32/50 |
| 14 | 3 | **28**(4, 6361)* | – | **451**(407, 492)* | 49/50 | 0/50 | 4/50 |
| | 5 | **272**(23, 16107)* | – | – | 43/50 | 0/50 | 0/50 |

Table 3.1: **Comparison of running times for COmbINE and SBCSD for networks of 10 and 14 variables**. The table reports the median running time along with the 5 and 95 percentiles, as well as the number of instances (problem inputs) in which each algorithm managed to complete; *numbers are computed only on the problems for which the algorithm completed.

instance. However, the algorithm requires all $m$-connections to constrain the search space (at least the ones that guarantee completeness), while COmbINE uses inducing paths to avoid that. For each adjacency $X{\ast\!\!-\!\!\!-\!\!\ast}Y$ in a data set, COmbINE creates a constraint specifying that at least one path between the variables is inducing with respect to $\mathbf{L_i}$. In contrast, SBCSD creates a constraint specifying that at least one path between the variables is $m$-connecting path given each possible conditioning set. So, both algorithms are forced to check every possible path, yet COmbINE examines each path once (with respect to $\mathbf{L_i}$), while SBCSD examines it for multiple possible conditioning sets. The latter choice may be necessary to deal with cyclic structures, but leads to significantly larger SAT problems when acyclicity is assumed.

## Comparison to SBCSD

SBCSD is not presented with a conflict resolution strategy and so it can only be tested by using an oracle of conditional independence. Equipping SBCSD with such a strategy is possible, but it may not be straightforward: SBCSD computes the SAT backbone incrementally for efficiency, which complicates pre-ranking constraints according to some criterion. Since SBCSD cannot handle conflicts, we compared it to the complete version of our algorithm (infinite maxK and maximum path length) using an oracle of conditional independence. Since no statistical errors are assumed, the initial search graph for COmbINE includes all observed arrows. Both algorithms are sound and complete, hence we only compare running time. SBCSD uses a path-analysis heuristic to limit the number of tests to perform. However, the authors suggest that in cases of acyclic structures, this heuristic could be substituted with the FCI test schedule. To better characterize the behavior of SBCSD on acyclic structures, we equipped the original implementation as suggested[2]. We denote this version of the algorithm as SBCSD′. Also note, that the available implementation of SBCSD by its authors has an option to restrict the search to acyclic structures, which was employed in the comparative evaluation. Finally, we note that SBCSD is implemented in C, while COmbINE is implemented in Matlab.

For the comparative evaluation, we simulated random acyclic networks with 10 and 14 variables. The default parameters were used to generate 50 problem instances for networks with 3 and 5 maximum parents per variable. Both algorithms were run on the same computer, with 4GB of available memory. SBCSD reached maximum memory and aborted without concluding in several

[2] However, we do not include the Possible d-Separating step of FCI; this step hardly influences the quality of the algorithm Colombo et al. (2012). Thus, the timing results of Table 3.1 are a lower bound on the execution time of the SBCSD algorithm.

cases for networks of 10 variables, and *in all cases for networks of 14 variables.* SBCSD′ slightly improves the running time over SBCSD. Median running time along with the 5 and 95 percentiles as well as number of cases completed are reported in Table 3.1. The metrics for each algorithm were calculated only on the cases where the algorithm completed.

The results in Table 3.1 indicate that COmbINE is more time-efficient than SBCSD and SBCSD′. While the running times do depend on implementation, the fact that SBCSD have much higher memory requirements indicates that the results must be at least in part due to the more compact representation of constraints by COmbINE . COmbINE managed to complete all cases for networks of 10 and most cases for 14 variables, while SBCSD completed less than 50% and 0%, respectively. SBCSD′ completed most cases for 10 variables but only 4% of cases for 14 variables. Interestingly, the percentiles for COmbINE are quite wide spanning two orders of magnitude for problems with maxParents equal to 5 (we cannot compute the actual 95 percentile for SBCSD since it did not complete for most problems). Thus, performance highly depends on the input structure. Such heavy-tailed distributions are well-noted in the constraint satisfaction literature (Gomes et al., 2000). We also note the fact that COmbINE seems to depend more on the sparsity and less on the number of variables, while SBCSD's time increases monotonically with the number of variables. Based on these results, we would suggest the use of COmbINE for problems where acyclicity is a reasonable assumption and the number of variables is relatively high.

Chapter 4

# Estimating posterior probabilities of pairwise features in causal graphs

*The statistical constraints imposed on the underlying SMCM by COmbINE are obtained by the available data sets using appropriate tests of independence. Statistical errors can result in conflicting constraints, making the resulting SAT formula unsatisfiable. In this chapter, we present a method for estimating posterior probabilities for some of these constraints: Adjacencies and non-adjacencies learnt using constraint-based algorithms. PROPeR, the proposed method, uses p-values calculated by the constraint-based algorithm and has no computational overhead. Even though the method is approximate, it produces calibrated probability estimates and performs on par with more expensive Bayesian methods. The estimates obtained by PROPeR can the be used to equip COmbINE with a conflict resolution strategy.*

Constraint-based algorithms such as FCI are a popular choice for learning causal models; they are fast, scalable, and usually guarantee soundness and completeness in the sample limit. However, for smaller sample sizes, identification of false constraints poses a challenge: An erroneous identification of a conditional independence can propagate through the network and lead to erroneous edge identifications or conflicting orientations even in seemingly unrelated parts of the network. Particularly for networks with many variables and small sample sizes, error propagation can result in unreliable networks. Thus, some of the literals included in list $\mathcal{F}$ in Algorithm 2 will be false, resulting in conflicting information.

Algorithm COmbINE as presented in the previous chapter, handles conflicts by admitting a conflict resolution strategy *str* which is able to select a list of non-conflicting literals. Many methods may be used for selecting such a subset, but for most of them one needs to assign a measure of confidence to each literal. Literals in $\mathcal{F}$ include the *pairwise features* of each observed PAG, i.e. adjacencies and non-adjacencies. In this chapter, we present a novel approach for obtaining posterior probabilities for these features.

Constraint-based algorithms query the data for conditional independencies and then use the results to constrain the search space of possible causal models. Failure to identify which parts of the output

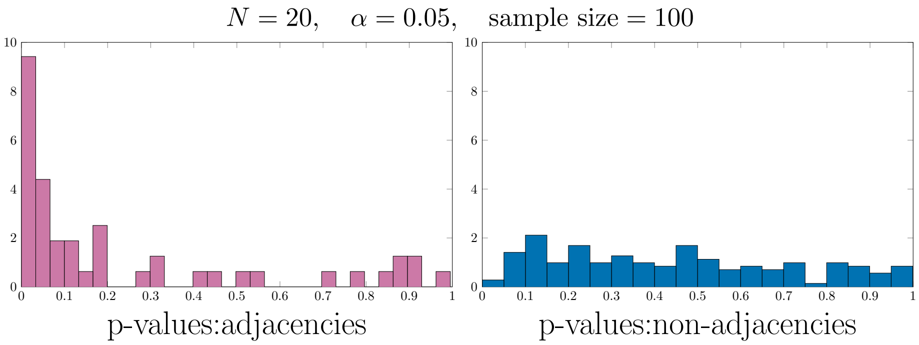


Figure 4.1: **Representative p-values for adjacencies and non-adjacencies**. Normalized histograms of 190 representative p-values identified by the PC skeleton algorithm for a random network of 20 variables. p-values corresponding to adjacencies in the data-generating network (left) follow a distribution with decreasing density. p-values corresponding to non-adjacencies in the data-generating network (right) follow a uniform distribution in the interval $[\alpha, 1]$. A smaller number of predictions fall in the $[0, \alpha]$ interval. This bias is introduced due to constraint-base search strategy: while the representative p-value is below the threshold, the algorithm performs more tests. Naturally, in real scenarios, we do not know which p-values come from which distribution

of a constraint-based algorithm are reliable is partly due to the nature of conditional independence tests: The test returns a p-value, which stands for the probability of getting a test statistic at least as extreme as the the one actually observed in the data, given that the null hypothesis (conditional independence) is true. If this probability is lower than a chosen significance threshold (typically 5-10%), the null hypothesis is rejected, and the alternative hypothesis is implicitly accepted. While lower p-values indicate higher confidence conditional dependencies, the p-value can not be interpreted as the probability of a conditional independence, and it therefore cannot be used to compare a conditional dependence to a conditional independence in terms of belief. Thus, the decisions made by the constraint-based algorithm (accept or reject a conditional independence) cannot be evaluated in terms of confidence.

We propose Posterior RatiO PRobability (PROPeR), a method for identifying posterior probabilities for all (non) adjacencies of a causal network learnt with a constraint-based algorithm. We use the term **pairwise relations** to denote adjacencies and non-adjacencies in a causal graph (ignoring orientations).

For each pair of variables, a constraint-based algorithm tries a number of conditional tests of independence. We use the maximum p-value obtained for every pair of variables as a representative of the corresponding pairwise relation. Posterior probabilities are then estimated as a function of these representative p-values. The method has no significant computational overhead, and can therefore scale up to the same number of variables as the algorithm of choice. Moreover, it does not depend on any additional assumptions (e.g. acyclicity, causal sufficiency, parametric assumptions) and can therefore be used with any constraint-based algorithm equipped with an appropriate test of conditional independence.

```
Algorithm 5: PROPeR
input  : causal network $\mathcal{G}$ over $\mathbf{V}$, representative p-values $\{p_{XY}\}$
output: Probability estimates $P(X—Y), P(\neg X—Y)$
Estimate $\hat{\pi_0}$ from $\{p_{XY}\}$ using the method described in Storey and Tibshirani (2003);
Find $\hat{\xi}$ that minimizes $-\sum_{(X,Y)\in\mathbf{V}^2} log(\hat{\pi_0} + (1-\hat{\pi_0})\xi p_{XY}^{\xi-1})$;
foreach $(X,Y) \in \mathbf{V}^2$ with representative p-value $p_{XY}$ do
    $E_0(p_{XY}) \leftarrow \frac{\hat{\pi_0}}{\hat{\xi} p_{XY}^{\hat{\xi}-1}(1-\hat{\pi_0})}$;
    $P(X—Y) \leftarrow \frac{1}{E_0(p_{XY})+1}$,   $P(\neg X—Y) \leftarrow \frac{E_0(p_{XY})}{E_0(p_{XY})+1}$;
end
```

## 4.1 Algorithm PROPeR

In this section, we present the PROPeR algorithm for estimating posterior probabilities of pairwise relations in causal networks. PROPeR takes as input the causal skeleton $\mathcal{G}$ returned by a constraint-based algorithm and a set of representative p-values and outputs a posterior probability estimate for every adjacency and non-adjacency in $\mathcal{G}$. We use $P(X—Y)$ and $P(\neg X—Y)$ to denote the posterior probability of the adjacency and non-adjacency of $X$ and $Y$ in $\mathcal{G}$, respectively.

According to the pairwise Markov condition, a non-adjacency in a causal graph $\mathcal{G}$ over variables $\mathbf{V}$ corresponds to a conditional independence given a subset of $\mathbf{V}$. In contrast, an adjacency in $\mathcal{G}$ corresponds to the lack of such a subset: If $X$ and $Y$ are adjacent in $\mathcal{G}$, there exists no subset $\mathbf{Z}$ of observed variables such that $X \perp\!\!\!\perp Y \mid \mathbf{Z}$. Thus, edge $X—Y$ will be present in $\mathcal{P}$ if the data support the null hypothesis

$$H_0 : \exists \mathbf{Z} \subset \mathbf{V} : X \perp\!\!\!\perp Y \mid \mathbf{Z} \textit{ less} \text{ than the alternative } H_1 : \forall \mathbf{Z} \subset \mathbf{V} : X \not\perp\!\!\!\perp Y \mid \mathbf{Z} \tag{4.1}$$

For a network with $N$ variables, this complex set of hypotheses involves $|2^{N-2}|$ conditional independencies. To simplify Equation 4.1, we use a surrogate conditioning set. For each pair of variables, during the skeleton search, a constraint-based algorithm performs a number of tests, each for a different conditioning set. To avoid performing all possible tests, most algorithms avoid conditioning sets that are theoretically not likely to be $d$-separating the variables, and also use a threshold on the cardinality of attempted conditioning sets. Let $p_{XY}$ be the maximum p-value of any attempted test of conditional independence between $X$ and $Y$, and let $\mathbf{Z}_{XY}$ be the corresponding conditioning set. $p_{XY}$ is used in constraint-based algorithms to determine whether $X$ and $Y$ are adjacent. If $p_{XY}$ is lower than the threshold $\alpha$, the edge is present in $\mathcal{G}$. Otherwise, the edge is absent in $\mathcal{G}$. We approximate Equation 4.1 with the following set of hypotheses:

$$H_0 : X \perp\!\!\!\perp Y \mid \mathbf{Z}_{XY} \text{ against the alternative } H_1 : X \not\perp\!\!\!\perp Y \mid \mathbf{Z}_{XY}, \tag{4.2}$$

Under $H_0$, the p-values follow a uniform distribution. Under $H_1$, the p-values follow a distribution with decreasing density. Sellke et al. Sellke et al. (2001) propose using Beta alternatives to model the distribution of the p-values under the null and the alternative hypotheses, respectively: $Beta(1,1)$ is the uniform distribution and describes the distribution of the p-values under the null hypothesis. $Beta(\xi,1), \quad 0 < \xi < 1$ is a distribution defined in $(0,1)$ with density decreasing in $p$. It is therefore suitable to model the distribution of p-values under the alternative hypothesis. Figure 4.1 shows an example of the distributions of representative p-values under $H_0$ and $H_1$, identified using the PC skeleton on data simulated from a known network. Equation 4.2 can be re-formulated on the basis of the representative p-value:

$$H_0 : p_{XY} \sim Beta(1,1) \text{ against } H_1 : p_{XY} \sim Beta(\xi,1) \text{ for some } \xi \in (0,1). \tag{4.3}$$

We can now estimate whether adjacency is more probable than non-adjacency for a given representative p-value $p$, *by estimating which of the Beta alternatives it is most likely to follow.* We use $\mathbf{V}^2 = \{(X,Y),\, X, Y \in \mathbf{V}, X \neq Y\}$ to denote the set of unordered pairs of $\mathbf{V}$, i.e. the set of pairwise relations in a causal skeleton $\mathcal{G}$. Let $\mathbf{p} = \{p_{XY} : (X,Y) \in \mathbf{V}^2\}$ be the set of the representative p-values for each pairwise relation. We assume that this population of p-values follows a mixture of $Beta(\xi,1)$ and $Beta(1,1)$ distributions. If $\pi_0$ is the proportion of p-values following $Beta(1,1)$, then the corresponding probability density function is:

$$f(p|\xi,\pi_0) = \pi_0 + (1-\pi_0)\xi p^{\xi-1}$$

For given estimates $\hat{\pi_0}$ and $\hat{\xi}$, the posterior odds of $H_0$ against $H_1$ for variables $X$, $Y$ is

$$\begin{aligned} E_0(p_{XY}) = \frac{P(p_{XY}|H_0)P(H_0)}{P(p_{XY}|H_1)P(H_1)} = \\ \frac{P(p_{XY}|p_{XY} \sim Beta(1,1))P(p_{XY} \sim Beta(1,1))}{P(p_{XY}|p_{XY} \sim Beta(\hat{\xi},1))P(p_{XY} \sim Beta(\hat{\xi},1))} = \frac{\hat{\pi_0}}{\hat{\xi} p_{XY}^{\hat{\xi}-1}(1-\hat{\pi_0})}. \end{aligned} \tag{4.4}$$

Obviously, if $E_0(p_{XY}) > 1$, non-adjacency is more probable than adjacency for the pair of variables $X$, $Y$. Notice that for some $\hat{\xi}$ and $\hat{\pi_0}$, it is possible that $E_0(p_{XY}) > 1$, while $X$ and $Y$ are adjacent in $\mathcal{G}$.

Based on the ratios in Equation 4.4, we can obtain the probability estimates:

$$P(X\text{—}Y) = \frac{1}{1+E_0(p_{XY})}, \quad P(\neg X\text{—}Y) = \frac{E_0(p_{XY})}{1+E_0(p_{XY})} \tag{4.5}$$

To estimate the probabilities in Equation 4.5, we need to obtain estimates for $\hat{\pi_0}$ and $\hat{\xi}$. To estimate $\pi_0$, we use the method described in Storey and Tibshirani (2003). The authors propose fitting a natural cubic spline to the distribution of the p-values to estimate the proportion of p-values that come from the null hypothesis.

The method requires that the p-values are i.i.d., an assumption that is clearly violated for the sample of p-values obtained during a skeleton identification algorithm: Typically, the tests of independence attempted by constraint-based network learning algorithms depend on the results of previously attempted tests. Moreover, each $p_{XY}$ is the maximum among many attempted tests. Finally, the p-values coming from the null hypothesis are not uniform, since independence is only accepted if $p > \alpha$. Thus, the obtained estimate $\hat{\pi_0}$ may be biased. Nevertheless, we believe that the estimates produced using this method are reasonable approximations. An example of the distribution of representative p-values coming from $H_0$ and $H_1$ is illustrated in Figure 4.1.

For a given $\hat{\pi_0}$, the likelihood for a set of representative p-values $\{p_{XY}\}$ is

$$L(\xi) = \prod_{(X,Y)\in\mathbf{V}^2} (\hat{\pi_0} + (1-\hat{\pi_0})\xi p_{XY}^{\xi-1}).$$

The respective negative log likelihood is

$$-LL(\xi) = -\sum_{(X,Y)\in\mathbf{V}^2} log(\hat{\pi_0} + (1-\hat{\pi_0})\xi p_{XY}^{\xi-1}). \tag{4.6}$$

Equation 4.6 can easily be optimized for $\xi$. Algorithm 5 describes how to obtain probability estimates for all pairwise relations given their representative p-values.

## 4.2 Related Work

Friedman et al. (1999) propose a method for estimating probabilities on features of Bayesian networks. They use bootstrap to resample the data and learn a Bayesian network from each sampled data set. The probability of a structural feature is then estimated as the proportion of appearances of the feature in the resulting networks. Friedman and Koller (2003) present a Bayesian method for estimating probabilities of features using MCMC samples over variable orderings. The methods are evaluated in terms of the classification performance (i.e. how accurately they accept or reject a feature), but not in terms of the calibration of predicted probability estimates.

Koivisto and Sood (2004) and Koivisto (2006) present algorithms for identifying exact posterior probabilities of edges in Bayesian networks. The methods use a dynamic programming strategy and constrain the search space of candidate causal models by bounding the number of possible parents per variable. The algorithms require a special type of non-uniform prior that does not respect Markov equivalence. Thus, resulting probabilities may be biased. Subsequent methods try to fix this problem by using MCMC simulations to compute network priors (Eaton and Murphy, 2007b) or exploiting special types of nodes (Tian and He, 2009). All methods in this category scale up to about 25 variables, since the minimum time and space requirement of these algorithms is $\mathcal{O}(n2^n)$.

Claassen and Heskes (2012b) propose a method for estimating Bayesian probabilities of a feature as a normalized sum of the posterior probabilities of all networks that entail this feature. The method requires exhaustive search of the space of possible networks, and is therefore not applicable for networks with more than 5-6 variables. The authors propose using this method as a standalone test of conditional independence, and also use it to decide on features inside a constraint-based algorithm. Pena et al. (2004) estimate the confidence of a feature as the fraction of models containing the feature out of the different locally optimal models.

## 4.3 Experimental Evaluation of PROPeR

We performed a series of experiments to characterize the behavior of the proposed algorithms.

### Calibration of Estimated Probabilities

We initially used simulated data to examine if the returned probability estimates are calibrated. We generated random DAGs with 10 and 20 variables, where each variable had 0 to 5 parents (randomly selected). The networks were then coupled with random parameters to create linear gaussian networks (continuous data) or discrete Bayesian networks (binary data). For continuous variables, a minimum correlation coefficient of 0.2 was imposed on the parameters to avoid weak interactions.

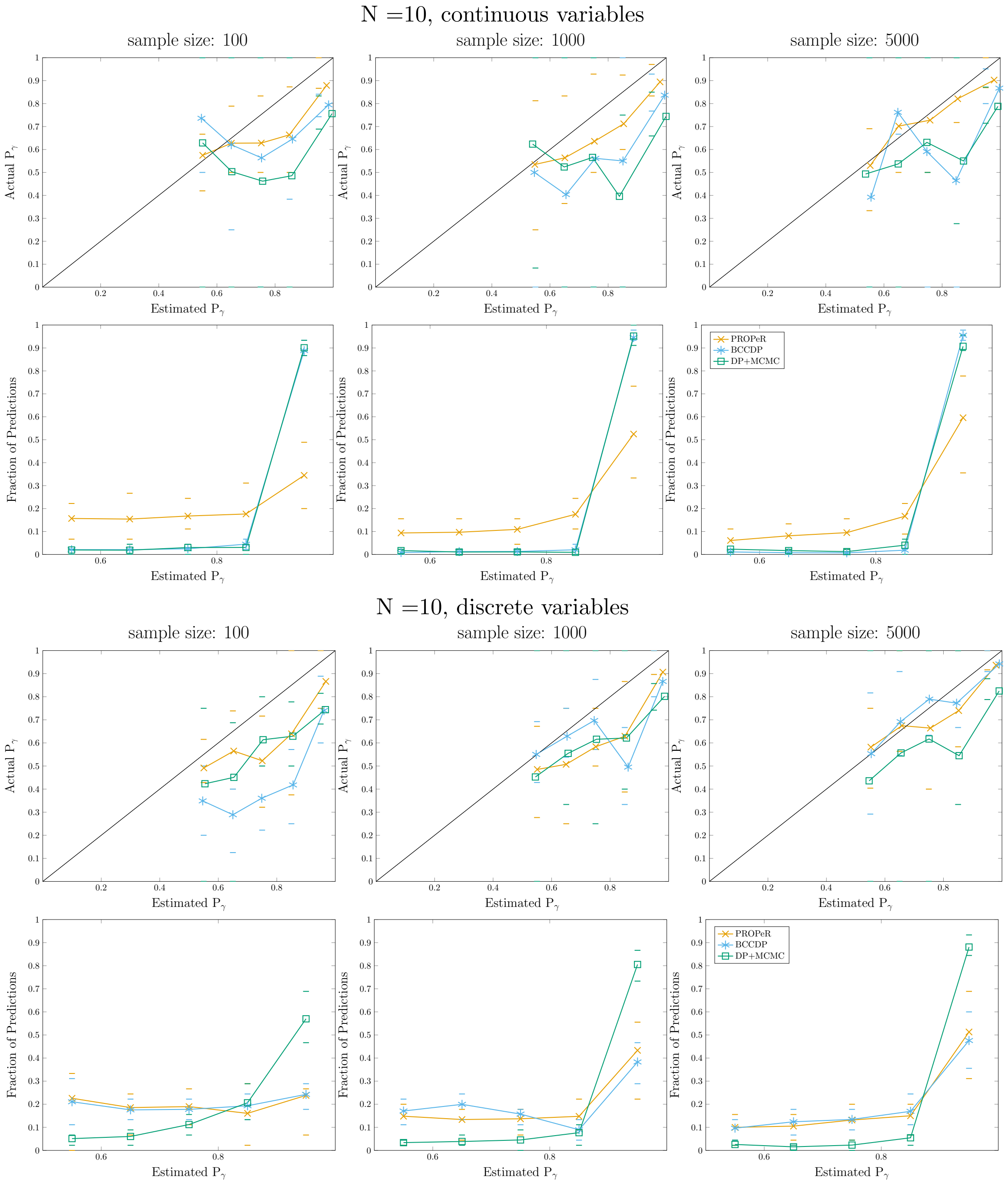


Figure 4.2: **Probability calibration plots for PROPeR, BCCD-P and MCMC+DP for networks of 10 variables.** Bars indicate the quartiles. All methods tend to overestimate probabilities. Bayesian scoring methods are often very confident: For continuous variables, most of the probability estimates predicted by BCCD-P or MCMC+DP lie in the interval [0.9, 1], while MCMC+DP exhibits similar behavior for discrete variables also.

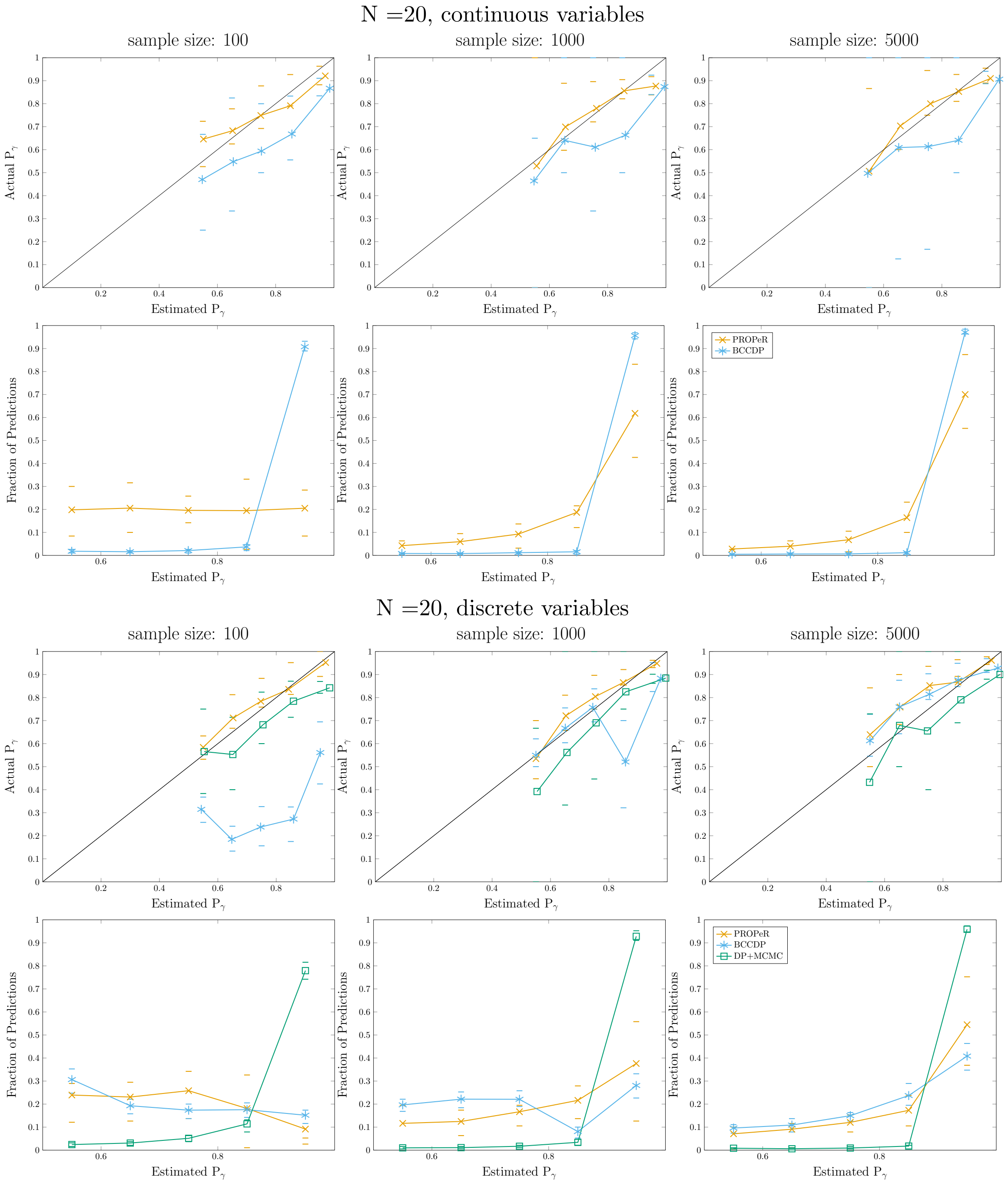


Figure 4.3: **Probability calibration plots for PROPeR, BCCD-P and MCMC+DP for networks of 20 variables.** Bars indicate the quartiles. Similar to the results in Figure 4.2, Bayesian scoring methods tend to overestimate probabilities. MCMC+DP produced memory errors and failed to complete in all iterations for the BGE score, and is therefore not inlcuded in the corresponding plot.

We then simulated networks of various sample sizes, to test the method's behavior in different settings.

We used the PC skeleton identification step Spirtes et al. (2000) with significance threshold $\alpha = 0.05$ and maximum conditioning set size 3 (explained below), modified to additionally return the maximum p-value encountered for each pair of variables. The set of maximum p-values was then used as input in Algorithm 5 to produce probability estimates for all pairwise relations. We compared our method against two alternative approaches:

1. **BCCD-P**: A method based on the BCCD algorithm presented in Claassen and Heskes (2012b). As mentioned above, the method estimates the posterior probability of a feature as a normalized sum of the posterior probabilities of DAGs that entail this feature. The algorithm scores all possible DAGs, and the authors use it to estimate probabilities for networks of at most 5 variables. To estimate the probabilities of pairwise relations, we scored the DAGs over variables $X$, $Y$ and $\mathbf{Z}_{XY}$, where $\mathbf{Z}_{XY}$ is the conditioning set maximizing the p-value of the tests $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ performed by PC. This means that the cardinality of $\mathbf{Z}_{XY}$ cannot exceed 3. For a fair comparison, we used 3 as the maximum conditioning set of PC in all experiments. The probability of an adjacency was estimated as: $P(X—Y) = \sum_{\mathcal{G} \vdash X—Y} P(\mathbf{D}|\mathcal{G})P(\mathcal{G})$. Consistent priors described in Claassen and Heskes (2012b) were pre-calculated and cached. To speed up the algorithm, we only scored one DAG per Markov equivalence class. For both approaches, we used the BDe metric for discrete data and the BGe metric for gaussian data. Both metrics are score-equivalent.

2. **DP+ MCMC**: The method presented in Eaton and Murphy (2007b) for identifying exact probabilities for edges in Bayesian networks. The method uses a combination of the DP algorithm Koivisto (2006) and MCMC sampling to correct the bias from the modular priors. We used the implementation provided by the authors in the BDAGL package. Maximum parents was set to 5, and the default parameters suggested by the authors in the package documentation were used. The method estimates probability estimates for directed edges, so we used $P(X—Y) = P(X \longrightarrow Y) + P(Y \longrightarrow X)$, $P(\neg X—Y) = 1 - P(X—Y)$.

To produce the probability calibration plots, the resulting predicted probabilities in [0.5, 1] were binned in 5 intervals. For every pair of variables, $P(X—Y)$ =1-$P(\neg X—Y)$. Thus, to consider each estimate once, we only need to consider half of the interval [0, 1]. If $N$ pairwise relations have probability estimates $\{\hat{P}_i\}_{i=1}^{N}$ that lie in interval $[\gamma, \gamma + 0.1]$, we expect that $\bar{\hat{P}}_i \times N$ of the corresponding relations will be true. The actual probability $P_\gamma$ for each interval is the fraction of relations with probability estimates in the given interval that are actually true in the data-generating graph. Figures 4.2 and 4.3 illustrate the mean estimated versus the mean actual probability for each bin, as well as the fraction of predictions in each bin for networks with 10 and 20 variables. Running times for all methods are shown in Figure 4.4.

Overall, results indicate that :

- PROPeR produces reasonable probability estimates, particularly in comparison to the more expensive BCCD-P and MCMC+DP approaches.

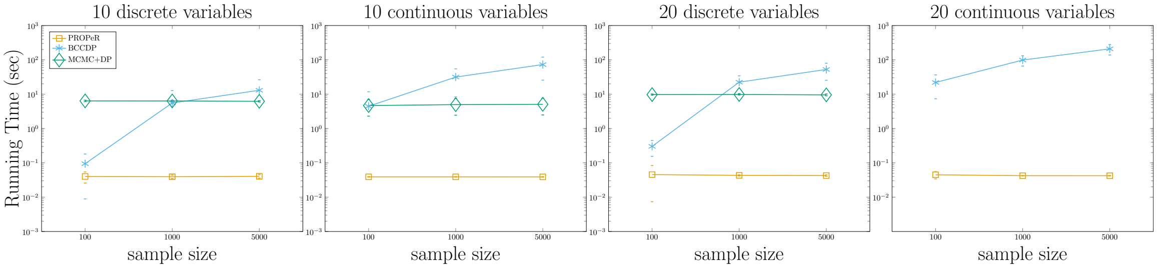


Figure 4.4: **Running times for PROPeR, BCCD-P and MCMC+DP**.

- MCMC+DP tends to identify very high (resp. very low) probabilities for the pairwise relations, even for small sample sizes for both metrics (BGE and BDE). BCCD has similar behavior for the BGE score, but not for the BDE score. This could explain the large deviations (and the seemingly unpredictable behavior) observed for these algorithms in the first four bins ([0.5 0.9]), since the means are computed over very few data points.

- As far as running times are concerned, both BCCD-P and MCMC+DP algorithms have (theoretically) exponential complexity with respect to the number of variables. BCCD-P also increases exponentially with sample size, but this is probably due to an increase in maximum conditioning set sizes reported by PC skeleton for larger sample sizes. i.e., BCCD-P iterates networks with many variables (4-5) for most pairwise relations. This also explains the poor performance of BCCD-P for the BDE metric and sample size 100: estimates are obtained by scoring smaller networks. The employed implementation of MCMC+DP failed to complete any iterations for N=20 and continuous variables.

We must point out that the calibration of the probability estimates *is not necessarily related to the predictive power of the respective approaches*, which depends more on the relative ranking of probabilities among pairwise relations, rather than the actual estimates. For example, MCMC+DP has been shown to produce rankings of edges with very high AUC Eaton and Murphy (2007b).

## 4.4 Using PROPeR to identify networks of high structural confidence

PROPeR is not used to improve the algorithm per se, but to produce confidence estimates for pairwise relations learnt from the algorithm and can be helpful in conflict resolution. However, even for a single data set, identifying which parts of the learnt network are reliable is of great importance for practitioners who use causal discovery methods, and are often interested in high-confidence pairwise connections among variables or in avoiding a specific type of error (e.g. false positive or false negative edges). It can also be useful for selecting subsequent experiments for a system under study, by pointing out relationships that are uncertain.

We use the estimates obtained by PROPeR, to identify neighborhoods of high structural confidence in causal networks. The proposed method, called BiND ($\beta$-NeighborhooDs), takes as input a causal graph $\mathcal{G}$ along with representative p-values for every pairwise relation in $\mathcal{G}$ and a desired threshold of confidence $\beta$. The algorithm outputs all neighborhoods in $\mathcal{G}$ for which all pairwise relations have confidence estimates above $\beta$. Internally, BiND uses PROPeR to obtain probability estimates for each pairwise relation, creates a graph $\mathcal{H}_\beta$ where edges correspond to pairwise relations with

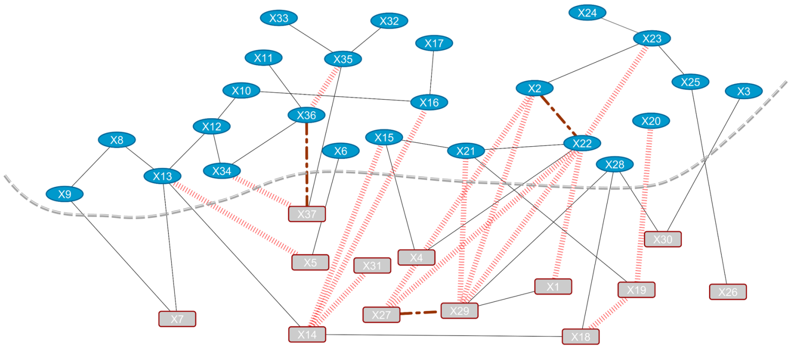


Figure 4.5: **An example maximum 0.8-neighborhood identified using Algorithm 6**. We used the DAG of the Alarm network coupled with random parameters to simulate 100 samples. PC-skeleton was used to obtain the network skeleton $\mathcal{G}$, consisting of 34 edges: 31 true positive edges (solid lines in the figure) and 3 false positive edges ($- \cdot -$ lines). 15 edges were not identified by the algorithm, even though they are present in the data-generating graph (false negative edges, depicted as |||| lines). Algorithm 6 was used to identify the maximum 0.8-neighborhoods of $\mathcal{G}$. One of the maximum 0.8-neighborhoods, consisting of 24 variables that share 17 adjacencies, is noted: elliptical blue nodes denote variables in the neighborhood, while the remaining variables are shown as rectangular grey nodes (the neighborhood is also separated from the rest of the network with a dashed grey line). The proportion of false inferences within the clique is far lower than the overall proportion of false inferences: The clique includes only two false negative edges and only one false positive. Most of the false inferences are pairwise relations between members and non-members of the neighborhood.

confidence above $\beta$, and then uses the Bron-Kerbosch algorithm to identify all maximal cliques in graph $\mathcal{H}_\beta$.

Algorithm 6 takes as input a causal skeleton $\mathcal{G}$, confidence estimates on $\mathcal{G}$'s pairwise relations and a confidence threshold $\beta$ and outputs the set of all $\beta$-neighborhoods in $\mathcal{G}$. In the previous section we presented a method for obtaining posterior probability estimates for all pairwise relations in a causal skeleton. In this section, we will use these estimates to identify neighborhoods of high structural confidence on the same skeleton. We define a neighborhood of structural confidence $\beta$ as follows:

**Definition 4.4.1 ($\beta$-neighborhood)** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a causal skeleton, and $\{P_{XY},\ (X,Y) \in \mathbf{V}^2\}$ the set of probability estimates:*

$$P_{XY} = \begin{cases} P(X \text{—} Y), \text{ if } (X,Y) \text{ adjacent in } \mathcal{G} \\ P(\neg X \text{—} Y), \text{ if } (X,Y) \text{ not adjacent in } \mathcal{G} \end{cases}$$

*A subgraph $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ of $G$ is a $\beta$-**neighborhood** iff: $\forall X, Y \in \mathbf{V}' : P_{XY} > \beta$ The size of a $\beta$-neighborhood $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ is $|\mathbf{V}'|$.*

Thus, a neighborhood of confidence $\beta$ is a subgraph of the causal network in which the posterior probability of every pairwise relation is above a given threshold $\beta$. For a causal skeleton and a set

```
Algorithm 6: BiND
input  : causal network $\mathcal{G}$ over $\mathbf{V}$, pairwise confidence estimates $P_{XY}$, confidence threshold $\beta$
output: $\beta$-neighborhoods $\{\mathcal{G}'\}$
$\mathcal{H}_\beta \leftarrow$ empty graph;
foreach $(X, Y), X, Y \in \mathbf{V}$ do
    if $P_{XY} \geq \beta$ then add $(X,\ Y)$ to $\mathcal{H}_\beta$;
end
$\{\mathbf{V}'\} \leftarrow$ Bron-Kerbosch($\mathcal{H}_\beta$);
$\{\mathcal{G}'\} \leftarrow$ subgraphs of $\mathcal{G}$ over $\{\mathbf{V}'\}$;
```

of confidence estimates on all pairwise relations, finding a $\beta$ - neighborhood can be reformulated as a graph theoretical problem: Let $\mathcal{H} = (\mathbf{V}, \mathcal{E}_\beta)$ be an undirected graph with edges defined as follows:

$$(X, Y) \in \mathcal{E}_\beta \text{ if } P_{XY} \geq \beta, \quad (X, Y) \notin \mathcal{E}_\beta \text{ if } P_{XY} < \beta \tag{4.7}$$

Variables $X$ and $Y$ are adjacent in $\mathcal{H}_\beta$ only if the probability of their respective pairwise relation in $\mathcal{G}$ is above the confidence threshold $\beta$. Finding $\beta$-neighborhoods in $\mathcal{G}$ is equivalent to identifying cliques in $\mathcal{H}_\beta$.

Naturally, a causal skeleton can have many $\beta$-neighborhoods. Moreover, if a subgraph $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ of $\mathcal{G}$ is a $\beta$-neighborhood, then every subgraph of $\mathcal{G}'$ is a $\beta$-neighborhood. More interesting inferences may be made by identifying all **maximal** $\beta$-neighborhoods on a graph:

**Definition 4.4.2** *Let $\mathcal{G}$=($\mathbf{V}$, $\mathcal{E}$) be a causal skeleton and $\mathcal{G}' = (\mathbf{V}', \mathbf{E}')$ be a $\beta$-neighborhood. $\mathcal{G}$ is a* ***maximal*** *$\beta$-neighborhood if $\nexists \mathbf{V}'' \supset \mathbf{V}'$ such that the subgraph $\mathcal{G}'' = (\mathbf{V}'', \mathcal{E}'')$ is a $\beta$-neighborhood.*

Thus, a maximal $\beta$-neighborhood is a $\beta$-neighborhood that is not part of a larger neighborhood. Identifying all maximal $\beta$-neighborhoods in $\mathcal{G}$ can be solved by finding all maximal cliques in the corresponding $\mathcal{H}_\beta$. Identifying maximal cliques is NP-hard Karp (1972), but algorithms that run in exponential time or identify approximate solutions are available. We use the Bron-Kerbosch algorithm Bron and Kerbosch (1973).

Maximal cliques can often be very small; for example, if no larger cliques exist, all adjacencies and all non-adjacencies with $P_{XY} > \beta$ are (trivial) maximal cliques of size 2. Another interesting problem that could be solved using Algorithm 6 is to identify the **maximum** $\beta$-neighborhoods of a causal skeleton, i.e. the maximal $\beta$-neighborhoods with the maximum possible number of variables. This is equivalent to identifying all maximum cliques in $\mathcal{H}_\beta$, and can be easily obtained from the output of Algorithm 6. Figure 4.5 shows an example maximum clique, identified using Algorithm 6 on simulated data. The neighborhood includes 24 out of 37 variables. While the neighborhood includes more than half of the total variables and edges of $\mathcal{G}$, the number of false positive and false negative edges within the neighborhood is much lower than the corresponding number in the entire skeleton.

To demonstrate the value of BiND, we simulated data of 100 and 1000 samples from random networks with 20 and 50 variables, as described above. For the causal skeletons identified with the PC

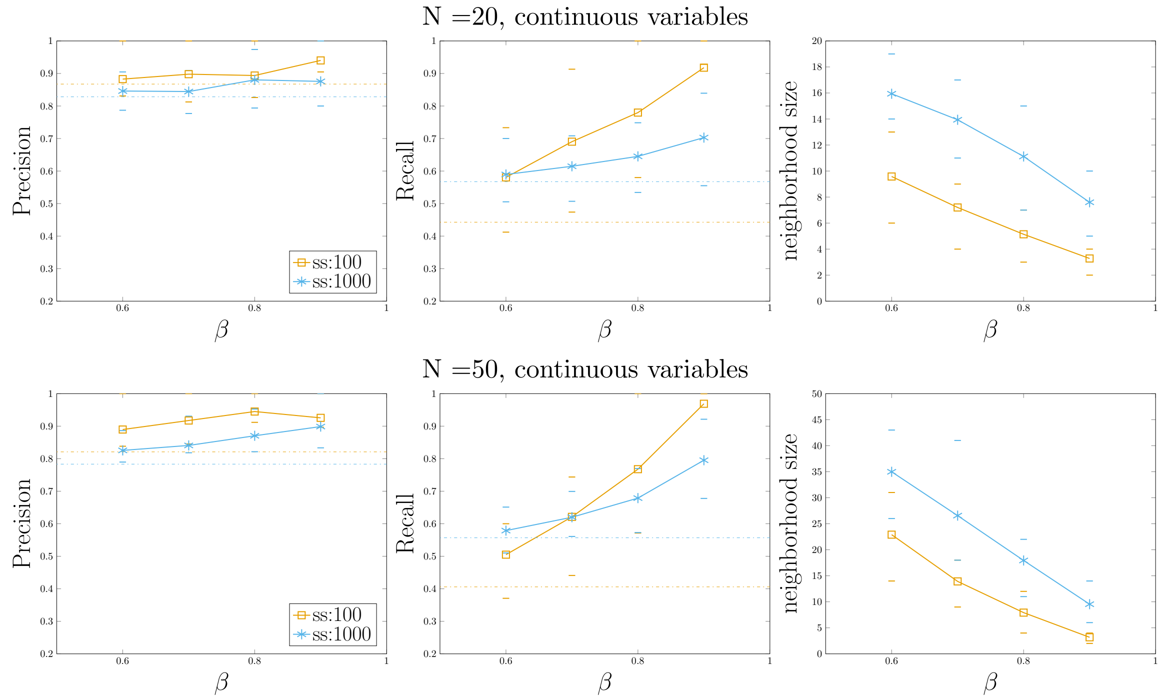


Figure 4.6: **Precision, recall, number and size of maximum cliques identified using BiND in networks of 20 continuous variables.** Bars indicate quartiles. Dashed horizontal lines show the mean baseline precision and recall (mean precision and recall of the output of PC skeleton. BiND identifies neighborhoods of higher structural precision and recall than the corresponding baseline, particularly for small sample sizes.

skeleton algorithm and the posterior probability estimates produced by PROPeR, all maximal $\beta$-neighborhoods for $\beta$=0.6, 0.7 and 0.9 were identified using Algorithm 6.

We examined the structural precision $(\frac{\#\text{ edges in }\mathcal{G}'\text{ and the ground truth}}{\#\text{ edges in }\mathcal{G}'})$ and recall $(\frac{\#\text{ edges in }\mathcal{G}'\text{ and the ground truth}}{\#\text{ edges in the ground truth}})$ of the resulting neighborhoods, compared to the baseline precision and recall for $\mathcal{G}$. As mentioned above,the maximal cliques can be very small and uninformative, particularly for high confidence thresholds. We are more interested in identifying large parts of the networks that we are confident about, and therefore focused in the maximum $\beta$-neighborhoods. Figure 4.6 illustrates the precision, recall and size of maximum $\beta$-neighborhoods for networks of 20 and 50 variables, for both discrete and continuous data. The algorithm took 85.56 seconds on average to identify the maximum 0.6-neighborhoods for 50 variables and 1000 samples (the most expensive case). Detailed time results are omitted due to space limitations.

Results indicate the following:

- The method identifies subgraphs with lower ratios of false inferences compared to the entire skeleton.

- For high confidence thresholds and small sample sizes, the algorithm cannot identify large neighborhoods.
- The algorithm is particularly useful in small sample sizes, where the overall recall is very low.

*Chapter 5*

# Experimental Evaluation of COmbINE

*PROPeR and similar algorithms presented in the previous chapter can be used to produce confidence estimates for the constraints employed by COmbINE . Based on these confidence estimates, we present four different strategies for selecting a list of non-conflicting constraints, and test their performance in simulated data sets. We also evaluate the performance of COmbINE against several factors that affect the robustness and the scalability of the algorithm. COmbINE is shown to perform reasonably against most factors, and scales up to 100 variables. To showcase the availability of real scenarios that can benefit from integrative causal analysis, we apply COmbINE on public mass cytometry data sets measuring intracellular proteins in the human immune system under different experiments.*

In chapter 3, we presented an algorithm for learning causal structure from data sets measuring overlapping sets of variables under different manipulations. The algorithm works by converting (non) adjacencies and (non) colliders with order into a constraint satisfaction instance. The instance is solved by all and only the SMCMs that could have generated all observed data sets. The algorithm was also proved to be sound and complete in the sample limit.The algorithm was evaluated against a similar algorithm that also admits cycles, and was proved to be computationally more efficient.

In this chapter, we also present a series of experiments to characterize how the behavior of COmbINE performs for varying input parameters and characteristics of the problem instance. In more detail, we explore (a) the learning accuracy of COmbINE for four different conflict resolution strategies (b) the learning accuracy of COmbINE as a function of the maximum path length considered by the algorithm, the density and size of the network to reconstruct, the number of input data sets, the sample size, and the number of latent variables, and (c) the computational time as a function of the above factors. Finally, we present a proof-of-concept application on real mass cytometry data on human T-cells.

## 5.1 Conflict Resolution

Recall that the literals in list $\mathcal{F}$ output by Algorithm 4 correspond to features of the PAGs learnt using FCI. In realistic settings, where an oracle of conditional independence is not present, some of these literals will be false due to imperfect statistical knowledge. Thus, output formula of Algorithm

2 is unsatisfiable. To avoid this problem, the Algorithm 2 as presented in chapter 3 admits a conflict resolution strategy *str*. The strategy takes as input the list of literals $\mathcal{F}$ and returns a modified list of non conflicting literals $\mathcal{F}'$. The goal any strategy *str* is to select a subset of these literals whose antecedents will not contradict each other.

Probably the most straight-forward would be to try to maximize the *number* of satisfied literals. This means satisfying as many adjacencies, non-adjacencies, colliders and non-colliders as possible. Recall that the SAT instance generated in Algorithm 2 consists of a set of hard-constraints (conditionals, no cycles, no tail-tail edges), which should always be satisfied (hard constraints), and a set of literals $\mathcal{F}$. To maximize the number of literals satisfied, while ensuring all hard-constraints are satisfied, the instance can be transformed a weighted max-SAT instance by assigning weight equal to one to each literal, and a weight larger than the number of all literals to each hard constraints.

Another approach would be to rank the list of literals in decreasing order of confidence, and then try to satisfy as many as possible *in the given order*. List $\mathcal{F}$ includes four types of literals, expressing different statistical information:

1. $adjacent(X, Y, \mathcal{P}_i)$: $X$ and $Y$ are not independent given any subset of $\mathbf{O}_i$.
2. $\neg adjacent(X, Y, \mathcal{P}_i)$: $X$ and $Y$ are independent given some $\mathbf{Z} \subset \mathbf{O}_i$
3. $col(\langle X, Y, Z\rangle, \mathcal{P}_i)$: $Y$ is in no subset of $\mathbf{O}_i$ that renders $X$ and $Z$ independent.
4. $dnc(\langle X, Y, Z\rangle, \mathcal{P}_i)$: $Y$ is in every subset of $\mathbf{O}_i$ that renders $X$ and $Z$ independent.

Assigning a measure of likelihood or posterior probability to every literal would enable their comparison. However, being able to assign a measure of confidence on the same scale for literals that represent so different statistical information is not an easy task. In chapter 4 we introduced a method for assigning posterior probabilities to the first two types of literals (adjacencies and non-adjacencies). Assigning posterior probabilities to (non) colliders is more complicated, since three (non) adjacencies *and* a list of conditional independencies are involved in the decision that a triple is a (non) collider. To simplify the problem, assign colliders and non-colliders with order to the same rank as the non-adjacency of the corresponding discriminating path's endpoints. We must point out, however, that this criterion is merely a heuristic, and measuring confidence in a (non) collider is a very interesting and complicated problem.

Recall that the posterior probabilities in PROPeR were computed based on the posterior odds

$$E_0(p_{XY}) = \frac{\hat{\pi_0}}{\hat{\xi} p_{XY}^{\hat{\xi}-1}(1-\hat{\pi_0})}.$$

$E_0(p_{XY}) > 1$ implies that for the test of independence represented by the p-value $p$, independence is more probable than dependence, while $E_0(p_{XY}) < 1$ implies the opposite. Moreover, the value of $E_0(p)$ *quantifies* this belief. Conversely, the corresponding posterior odds of $H_1$ against $H_0$ is

$$E_1(p) = \frac{\hat{\xi} p^{\hat{\xi}-1}(1-\hat{\pi_0})}{\hat{\pi_0}}.$$

We define the **maximum posterior ratio (MPR)** for a p-value $p$ to be the maximum between the two:

$$E(p) = max\{\frac{\hat{\pi_0}}{\hat{\xi}p^{\hat{\xi}-1}(1-\hat{\pi_0})}, \frac{\hat{\xi}p^{\hat{\xi}-1}(1-\hat{\pi_0})}{\hat{\pi_0}}\}. \tag{5.1}$$

MPR estimates can be used to heuristically quantify our confidence in the observed adjacencies and non-adjacencies. A strategy for selecting a list of non-conflicting literals based on MPR estimates is presented in Algorithm 7. Let $X$ and $Y$ be a pair of observed variables, and $p_{XY}$ be the maximum p-value reported during FCI for these variables. Then, if $E_0(p_{XY}) > E_1(p_{XY})$, the literal $\neg adjacent(X, Y, i)$ is added to $\mathcal{F}$ with confidence estimate $E(p_{XY})$. Otherwise, the literal $adjacent(X, Y, i)$ is added to $\mathcal{F}$ with a confidence estimate $E(p_{XY})$. The list can then be sorted in order of confidence, and the literals can be satisfied incrementally. Whenever a literal in the list is encountered that cannot be satisfied in conjunction with the ones already selected, it is ignored. Naturally, using MPR estimates is identical to using the maximum PROPeR estimate in terms of ranking.

Notice that, it is possible that for a p-value $E_0(p_{XY}) > E_1(p_{XY})$ (i.e., MPR determines independence is more probable), even though $p_{XY}$ is smaller than the FCI threshold used. In other words, given a fixed FCI threshold, dependence maybe accepted; but, when analyzing the set of p-values encountered to compute MPR, independence seems more probable. The reverse situation is also possible. The pseudo-code in Algorithm 7 (Lines 6—10) accepts the MPR decisions for dependencies and independencies; *this implies that some of the decisions made by FCI will be reversed.* Nevertheless, in anecdotal experiments we found that the literals for which this situation occurs are near the end of the sorted list; thus, whether one accepts the initial decisions of FCI based on a fixed threshold, or a dynamic threshold based on MPR usually does not have a large impact on the output of the algorithm.

Figure 5.1(left) shows how the MPR varies with the p-value for $\hat{\pi_0} = 0.6$ and several $\hat{\xi}$'s. The lowest possible value of the MPR is 1, and corresponds to the p-value $p$ for which $E_0(p) = E_1(p)$. Naturally, for the same $\xi$, this p-value (where the odds switch in favor of non-adjacency) is larger for a lower $\pi_0$. In Figure 5.1 for $\pi_0 = 0.6$ we can see an example of two p-values that correspond to the same $E$: An adjacency represented by a p-value of 0.0038 (0.0038 being the *maximum* p-value of any test performed by FCI for the pair of variables) is as likely as a non-adjacency represented by a p-value of 0.6373 (0.6373 being the p-value based on which FCI removed this edge).

The same potential pitfalls discussed in chapter 4 also apply here: The method used to obtain $\hat{\pi_0}$ assumes independent p-values, which is of course not the case since the test schedule of FCI depends on previous decisions. In addition, each p-value may be the maximum of several p-values; these maximum p-values may not follow a uniform distribution even when the non-adjacency (null hypothesis) is true. Finally, given that p-values stem from tests over different conditioning set sizes, p-values corresponding to adjacencies do not necessarily follow the same beta distribution. Thus, the approach presented here is at best an approximation.

In the algorithm as presented, a single beta is fit from the pooled p-values of FCI runs over all data-sets. This strategy is perhaps more appropriate when individual data-sets have a small number of p-values, so the pooled set provides a larger sample size for the fitting. Other strategies though, are also possible. One could instead fit a different beta for each data-set and its corresponding set of p-values. This approach could perhaps be more appropriate in case the PAG structures $\mathcal{P}_i$ vary

```
Algorithm 7: MPRstrategy
input  : SAT formula $\Phi$, list of literals $\mathcal{F}$, list of p-values $\{p_j\}$(each corresponding to an observed
         (non) adjacency)
output: List of non conflicting literals $\mathcal{F}'$
$\mathcal{F}' \leftarrow \emptyset$;
Estimate $\hat{\pi}_0$ from $\{p_j\}$ using the method described in Storey and Tibshirani (2003);
Find $\hat{\xi}$ that minimizes $-\sum_j log(\hat{\pi}_0 + (1-\hat{\pi}_0)\xi p_j^{\xi-1})$;
foreach literal $(\neg)adjacent(X, Y, \mathcal{P}_i) \in \mathcal{F}$ with p-value $p_j$ do
    $E_0(p_j) \leftarrow \frac{\hat{\pi}_0}{\hat{\xi} p_j^{\hat{\xi}-1}(1-\hat{\pi}_0)}$, $E_1(p_j) \leftarrow \frac{\hat{\xi} p_j^{\hat{\xi}-1}(1-\hat{\pi}_0)}{\hat{\pi}_0}$;
    if $E_1(p_j) < E_0(p_j)$ then
        add $\neg adjacent(X, Y, \mathcal{P}_i)$ to $\mathcal{F}$
    else
        add $adjacent(X, Y, \mathcal{P}_i)$ in $\mathcal{F}$
    end
    $Score(literal) \leftarrow max\{E_0(p_j), E_1(p_j)\}$;
end
foreach literal $collider(X, Y, Z, \mathcal{P}_i)$, $dnc(X, Y, Z, \mathcal{P}_i)$ do
    if $X$, $Y$, $Z$ is an unshielded triple in $\mathcal{P}_i$ then
        $Score(literal) \leftarrow Score(X, Z, \mathcal{P}_i)$;
    else if $\langle W \ldots X, Y, Z \rangle$ is discriminating for $Y$ in $\mathcal{P}_i$ then
        $Score(literal) \leftarrow Score(W, Z, \mathcal{P}_i)$;
    end
end
$\mathcal{F} \leftarrow$ sort $\mathcal{F}$ by descending score;
foreach $\phi \in \mathcal{F}$ do
    if $\Phi \wedge \phi$ is satisfiable then
        $\Phi \leftarrow \Phi \wedge \phi$;
        Add $\phi$ to $\mathcal{F}'$;
    end
end
```

greatly in terms of sparseness. In addition, one could also fit different beta distributions for each conditioning set size. Figure 5.2 shows the empirical distribution of p-values and the estimated $\hat{\pi}_0$ based on the p-values returned from FCI on 2, 5 and 10 input data sets, simulated from a network of 14 variables. Figure 5.1 (center) also show the calibration of PROPeR estimates when $\hat{\xi}$ is estimated from the pooled p-values, for data simulated from artificial networks.

The advantages of MPR-based strategy are also similar to the ones discussed for PROPeR: The method is based on p-values and thus, can be applied in different types of data (e.g., continuous and discrete) in conjunction with any appropriate test of independence. Moreover, since it is based on cached p-values, and fitting a beta distribution is efficient, it adds minimal computational complexity.

Naturally, other measures of confidence discussed in the previous chapter can also be used to rank the literals. Two methods for obtaining such measures were discussed in the previous chapter: A

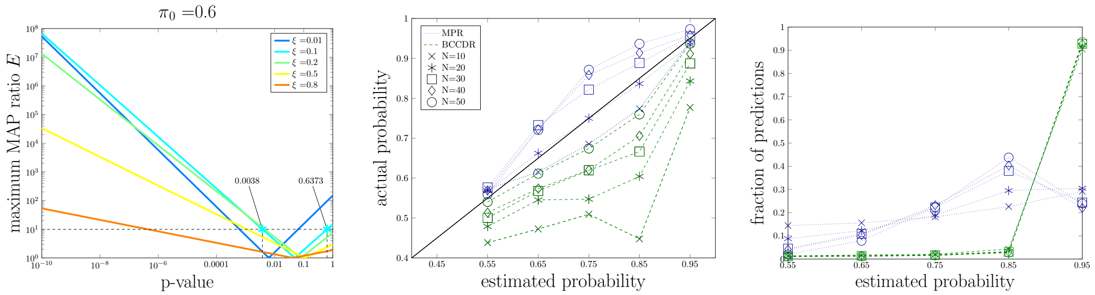


Figure 5.1: **Behaviour and calibration of MPR estimates.**(left) Log of the maximum posterior ratio $E(p)$ versus log of the p-value $p$ for $\hat{\pi_0} = 0.6$ and various $\hat{\xi}$. For $\hat{\xi} = 0.1$, an adjacency supported by a maximum p-value of 0.0038 corresponds to the same $E$ as a non-adjacency supported by a p-value of 0.6373. The intersection point of the line with the x axis is the p for which $E_0(p) = E_1(p) = 1$. (center) Probability calibration plots for confidence estimates obtained using MPR estimates $(1/(1 + E_0(p))$ for adjacencies, $E_0(p)/(1 + E_0(p))$ for non-adjacencies). For each interval of length 0.1 in [0.5, 1], the estimated confidences are plotted against the actual frequency of correctness of the corresponding constraints. The green lines correspond to estimates obtained using BCCDR (see section 5.2) The confidence estimates correspond to the experiments presented in Figure 5.3. (right). Number of confidences in each interval.

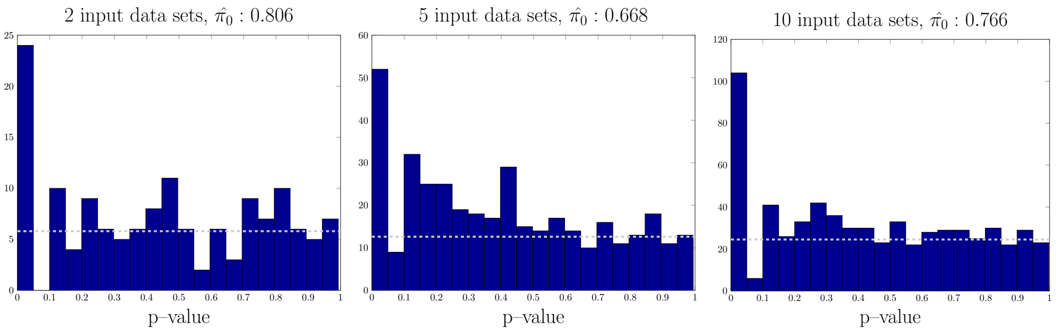


Figure 5.2: **Distribution of p-values and estimated** $\hat{\pi_0}$. We used the method of Storey and Tibshirani (2003) to estimate $\hat{\pi_0}$ for a sample of p-values corresponding to 2 (left), 5 (center) and 10 (right) input data sets. We generated networks by manipulating a marginal of the ALARM network (Beinlich et al., 1989) consisting of 14 variables. In each experiment, at most 3 variables were latent and at most 2 variables were manipulated. We simulated data sets of 100 samples each from the resulting manipulated graphs. We ran FCI on each data set with $\alpha = 0.1$ and $maxK = 5$ and cached the maximum p-value reported for each pair of variables. We used the p-values from all data sets to estimate $\hat{\pi_0}$. The dashed line corresponds to the proportion of p-values that come from the null distribution based on the estimated $\hat{\pi_0}$.

Bayesian method (Claassen and Heskes, 2012a) and an exact method (Eaton and Murphy, 2007b). The latter could not scale up to more than 20 variables, and we therefore did not use it in this context. The former is included in the comparative evaluation presented in Section 5.2.

| **Problem attribute** | **Default value used** |
|---|---|
| Number of variables in the generating DAG | 20 |
| Maximum number of parents per variable | 5 |
| Number of input data sets | 5 |
| Maximum number of latent variables per data set | 3 |
| Maximum number of manipulated variables per data set | 2 |
| Sample size per data set | 1000 |

Table 5.1: **Default values used in generating experiments in each iteration of COmbINE**. Unless otherwise stated, the input data sets of COmbINE were generated according to these values.

Instead of greedily trying to satisfy as many of the ranked constraints as possible, we could try to identify the subset of non-conflicting literals that maximize the sum of (satisfied) scores. Again, due to the hard constraints in the output of Algorithm 4, this has to be implemented as a weighted max SAT instance, where hard constraints have weights that are larger than the sum of literal scores.

## 5.2 Evaluation of COmbINE in Simulated Data

We now present a series of experiments evaluating COmbINE' s performance. All experiments were performed on data simulated from randomly generated networks as follows. The graph of each network is a DAG with a specified number of variables and maximum number of parents per variable. Variables are randomly sorted topologically and for each variable the number of parents is uniformly selected between 0 and the maximum allowed. The parents of each variable are selected with uniform probability from the set of preceding nodes. Each DAG is then coupled with random parameters to generate conditional linear gaussian networks. To avoid very weak interactions, minimum absolute conditional correlation was set to 0.2. Before generating a data set, the variables of the graph are partitioned to unmanipulated, manipulated, and latent. Mean value and standard deviation for the manipulated variables were set to 0 and 1, respectively. Subsequently, data instances are sampled from the network distribution, considering the manipulations and removing the latent variables. All experiments are performed on **conservative** families of targets; the term was introduced in Hauser and Bühlmann (2012) to denote families of intervention targets in which all variables have been observed unmanipulated at least once.

For each invocation of the algorithm, the problem instance (set of data sets) is generated using the parameters shown in Table 5.1. COmbINE default parameters were set as follows: maximum path length = 3, $\alpha = 0.1$ and maximum conditioning set $maxK = 5$, and the Fisher z-test of conditional independence. As far as orientations are concerned, in our experience, FCI is very prone to error propagation, we therefore used the rule in (Ramsey et al., 2006) for *conservative* colliders. Unless otherwise stated, Algorithm 7 is employed to resolve conflicts. SAT instances were solved using MINISAT2.0 (Eén and Sörensson, 2004) along with the modifications presented in Hyttinen et al. (2013) for iterative solving and computing the backbone with some minor modifications for sequentially performing literal queries. In the subsequent experiments, *one of the problem parameters in Table 5.1 is varied each time, while the others retain the values above.*

To measure learning performance, ideally one should know the correct output, i.e., the structure that the algorithm would learn if ran with an oracle of conditional independence, and unrestricted infinite maxK and maximum path length parameters. Notice that *the original generating DAG structure* *cannot serve as the correct output for comparison.* This is because the presence of manipulated and latent variables implies that not all structural features of the generating DAG can be recovered. For example, for the problem instance presented in Figure 3.5 (middle), the correct output, shown in Figure 3.5 (right), has one solid edge out of 5, no solid endpoint, one absent, and four dashed edges. Dashed edges and endpoints in the output of the algorithm can only be evaluated if one knows this correct output. Unfortunately, the correct output cannot be recovered in a timely fashion in most problems involving more than 15 variables, as shown in Section 3.4.

As a surrogate, we defined metrics that do not consider dashed edges or endpoints and can be directly computed by comparing the “solid” features of the output with the original data generating graph. Specifically, we used two types of precision and recall; one for edges (s-Precision/s-Recall) and one for orientations (o-Precision/o-Recall). Let $\mathcal{G}$ be the graph that generated the data (the SMCM stemming from the initial random DAG after marginalizing out variables latent in all data sets), and $\mathcal{H}$ be the summary graph returned by COmbINE. s-Precision and s-Recall were then calculated as follows:

$$\text{s-Precision} = \frac{\#\text{ solid edges in } \mathcal{H} \text{ that are also in } \mathcal{G}}{\#\text{ solid edges in } \mathcal{H}}$$

and

$$\text{s-Recall} = \frac{\#\text{ solid edges in } \mathcal{H} \text{ that are also in } \mathcal{G}}{\#\text{ edges in } \mathcal{G}}.$$

Similarly, orientation precision and recall are calculated as follows:

$$\text{o-Precision} = \frac{\#\text{ endpoints in } \mathcal{G} \text{ correctly oriented in } \mathcal{H}}{\#\text{ of orientations(arrows/tails) in } \mathcal{H}}$$

and

$$\text{o-Recall} = \frac{\#\text{ endpoints in } \mathcal{G} \text{ correctly oriented in } \mathcal{H}}{\#\text{ endpoints in } \mathcal{G}}.$$

Since dashed edges and endpoints do not contribute to these metrics, precision in particular could be favorable for conservative algorithms that tend to categorize all edges (endpoints) as dashed. To alleviate this problem, we accompany each precision / recall figure with the percentage of dashed edges out of all edges in the output graph to indicate how conservative is the algorithm. Similarly, we present the percentage of dashed (circled) endpoints out of all endpoints in the output graph. Finally, we note that in the experiments that follow, unless otherwise stated, we report the median, 5, and 95 percentile over 100 runs of the algorithm with the same settings.

### Evaluation of Conflict Resolution Strategies

In this section we evaluate our Maximum Map Ratio strategy (**MPR**) against three other alternatives: A ranking strategy where constraints are sorted based on Bayesian probabilities as proposed in Claassen and Heskes (2012a) (**BCCDR**), as well as a Max-SAT (**MaxSAT**) and a weighted max-SAT (**wMaxSAT**) approach.

**MPR**: This strategy sorts constraints according to the Maximum Map Ratio (Algorithm 7) and greedily satisfies constraints in order of confidence; whenever a new constraint is not satisfiable given the ones already selected, it is ignored (lines 21- 25 in Algorithm 7).

**BCCDR**: BCCDR sorts constraints according to Bayesian probability estimates of the literals in $\mathcal{F}$ as presented in Claassen and Heskes (2012a). The same greedy strategy for satisfying constraints

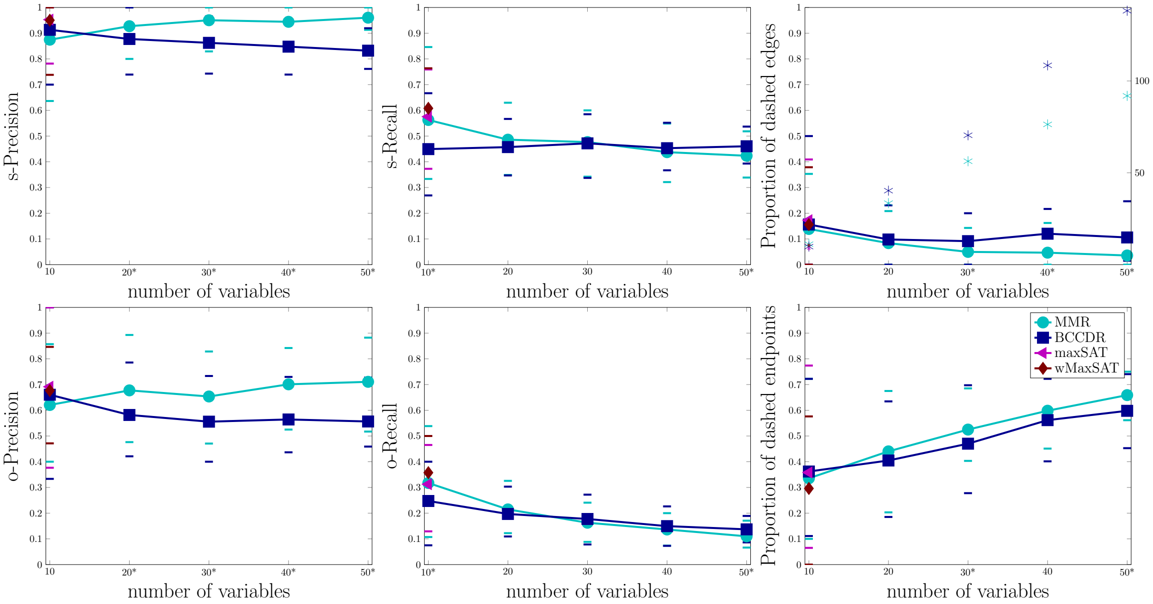


Figure 5.3: **Learning performance of COmbINE with various conflict resolution strategies**. From left to right: Median s-Precision, s-Recall, proportion of dashed edges (top) and o-Precision, o-Recall and proportion of dashed endpoints (bottom) for networks of several sizes for various conflict resolution strategies. Each data set consists of 100 samples. The numbers for wMaxSAT and maxSAT correspond to 22 and 23 cases, respectively, in which the algorithms managed to return a solution within 500 seconds. Coloured bars indicate 5 and 95 percentiles. Asterisks in the top right figure show the absolute number of literals rejected by each strategy (y axis on the right). Asterisks on x tick labels indicate cases where the behaviour of MPR and BCCDR are significantly different (paired t-test of equality of means with unknown but equal variances).

in order is employed. Briefly, the authors of (Claassen and Heskes, 2012a) propose a method for calculating Bayesian probabilities for any feature of a causal graph (e.g. adjacency, $m$-connection, causal ancestry). To estimate the probability of a feature, for a given data set $\mathbf{D}$, the authors calculate the score of all DAGs of $N$ variables. Let $\mathcal{G} \vdash f$ denote that a feature $f$ is present in DAG $\mathcal{G}$. The probability of the feature is then calculated as $P(f) = \sum_{\mathcal{G} \vdash \mathrm{f}} P(\mathbf{D}|\mathcal{G})P(\mathcal{G})$. Scoring all DAGs is practically infeasible for networks with more than 5 or 6 variables. Thus, for data sets with more variables, a subset of variables must be selected for the calculation of the probability of a feature. Following (Claassen and Heskes, 2012a), we use 5 as the maximum $N$ attempted.

The literals in $\mathcal{F}$ represent information on adjacencies: $(\neg)adjacent(X, Y, \mathcal{P}_i)$ and colliders: $(\neg)collider(X, Y, Z, \mathcal{P}_i)$. To apply the method above for a given feature, we have to select the variables used in the DAGs, a suitable scoring function, and suitable DAG priors. For (non) adjacencies $X {\star\!\!-\!\!\!-\!\!\star} Y$ in PAG $\mathcal{P}_i$, we scored the DAGs over variables $X$, $Y$ and $\mathbf{Z}$, for the conditioning set $\mathbf{Z}$ maximizing the p-value of the tests $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ performed by FCI. Since the total number of variables cannot exceed 5, the maximum conditioning set for FCI is limited to 3 in all experiments in this section for a fair comparison. (Non) colliders are assigned the same score as the non adjacency of their endpoints.

We use the BGE metric for gaussian distributions (Geiger and Heckerman, 1994) as implemented in the BDAGL package Eaton and Murphy (2007a) to calculate the likelihoods of the DAGs. This metric is score equivalent, so we pre-computed representatives of the Markov equivalent networks of up to 5 nodes, and scored only one network per equivalence class to speed up the method. Priors for the DAGs were also pre-computed to be consistent with respect to the maximum attempted number of nodes (i.e. 5) as suggested in Claassen and Heskes (2012a).

**MaxSAT**: This approach tries to satisfy as many literals in $\mathcal{F}$ as possible. To maximize the number of literals satisfied, while ensuring all hard-constraints are satisfied, each literal is assigned a weight of 1, and each hard-constraint is assigned a weight equal to the sum of all weights in $\mathcal{F}$ plus 10000. We use the akmaxsat (Kuegel, 2010) *weighted* max SAT solver to solve the final instance .The summary graph returned by Algorithm 2 is based on the backbone of the subset of literals selected by akmaxsat.

**wMaxSAT**: Finally, we augmented the above technique with a different weighted strategy that considers the importance of each literal. Specifically, each literal was weighted proportionally to the logarithm of the corresponding MPR. Again, each hard-constraint was assigned a weight equal to the sum of all weights in $\mathcal{F}$ plus 10000, to ensure that the solver will always satisfy these statements. The summary graph returned by Algorithm 2 is based on the backbone of the subset of literals selected by akmaxsat.

We ran all methods for networks of 10, 20, 30, 40 and 50 variables for data sets of 100 samples to test them on cases where statistical errors are common. For each network size we performed 50 iterations. **MaxSAT** and **wMaxSAT** often failed to complete in a timely fashion; to complete the experiments we aborted the solver after 500 seconds. We note that this amount of time corresponds to more than 10 times the maximum running time of the MPR method (calculating MPRs and solving the SAT instance), and more than twice times the maximum running time of the BCCDR-based method (for 50 variables). Cases where the solver did not complete were not included in the reported statistics. Unfortunately, *the methods using weighted max SAT solving failed to complete in most cases for 10 variables*, and all cases for more than 10 variables.

The results are shown in Figure 5.3, where we can see the median performance of both algorithms over 50 iterations. Overall, **MPR** exhibits better Precision and identifies more solid edges, while **BCCDR** exhibits slightly better Recall. **BCCDR** is better for variable size equal to 10, which could be explained from the fact that **MPR** is not provided with sufficient number of p-values to estimate $\hat{\pi}_0$ and $\hat{\xi}$. In terms of computational complexity, for networks of 50 variables, estimating the **BCCDR** ratios takes about 150 seconds on average, while estimating the **MPR** ratios takes less than a second. The more sophisticated search strategies **MaxSAT** and **wMaxSAT** do not seem to offer any significant quality benefits, at least for the single variable size for which we could evaluate them. Based on these results, we believe that **MPR** is a reasonable and relatively efficient conflict resolution strategy.

### COmbINE performance with increasing maximum path length

In this section, we examine the behavior of the algorithm when the length of the paths considered is limited, in which case the output is an approximation of the actual solution. The COmbINE pseudo-code in Algorithm 2 accepts the maximum path length as a parameter.

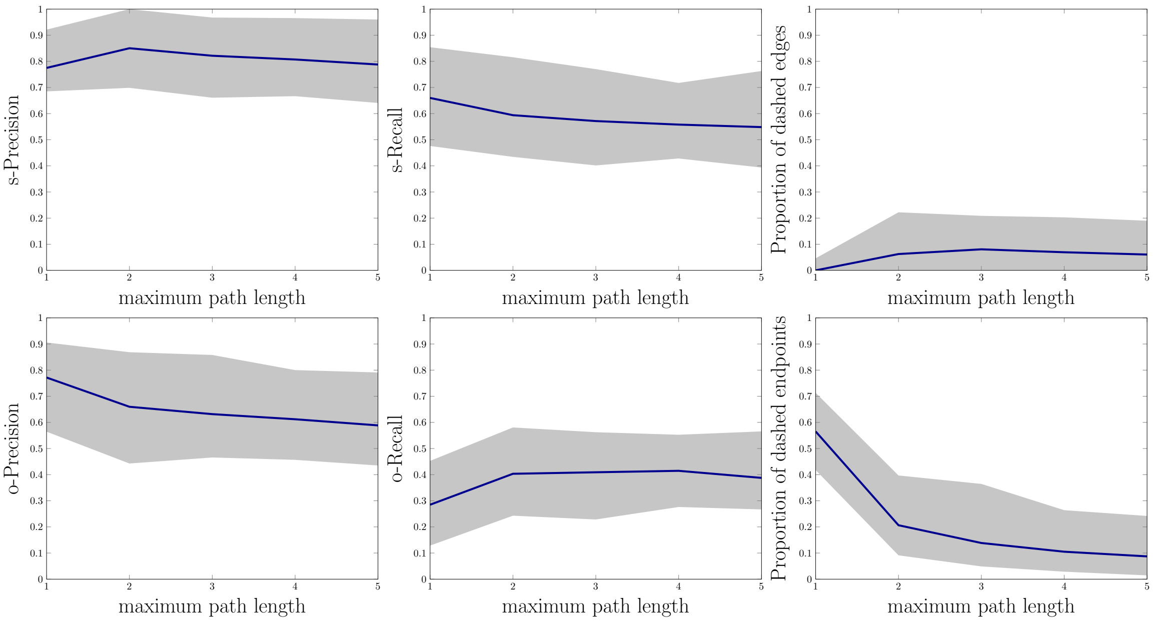


Figure 5.4: **Learning performance of COmbINE against maximum path length**. From left to right: s-Precision, s-Recall, percentage dashed edges and o-Precision, o-Recall and percentage of dashed endpoints (bottom) for varying maximum path length, averaged over all networks. Shaded area ranges from the 5 to the 95 percentile. Maximum path length 3 seems to be a be a reasonable trade-off between performance, percentage of dashed features, and efficiency.

| | | Actual $\mathcal{H}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | maxParents 3 | | | maxParents 5 | | |
| | **Edges** | solid | dashed | absent | solid | dashed | absent |
| $\hat{\mathcal{H}}$ | solid | **8.0** $(4.0, 12.0)$ | **0.0** $(0.0, 5.0)$ | **0.0** $(0.0, 4.0)$ | **9.0** $(3.0, 13.0)$ | **1.0** $(0.0, 10.0)$ | **1.0** $(0.0, 5.0)$ |
| | dashed | **0.0** $(0.0, 3.0)$ | **0.0** $(0.0, 4.0)$ | **0.0** $(0.0, 2.0)$ | **0.5** $(0.0, 4.0)$ | **0.5** $(0.0, 3.0)$ | **1.0** $(0.0, 2.0)$ |
| | absent | **1.0** $(0.0, 4.0)$ | **0.0** $(0.0, 3.0)$ | **31.0** $(24.0, 36.0)$ | **2.5** $(0.0, 8.0)$ | **1.5** $(0.0, 9.0)$ | **24.0** $(14.0, 34.0)$ |
| | **Endpoints** | arrow | circle | tail | arrow | circle | tail |
| | arrow | **8.0** $(4.0, 12.0)$ | **1.0** $(0.0, 5.0)$ | **0.0** $(0.0, 3.0)$ | **8.0** $(4.0, 13.0)$ | **3.0** $(0.0, 8.0)$ | **2** $(0.0, 5.0)$ |
| | circle | **1.0** $(0.0, 3.0)$ | **3.0** $(0.0, 14.0)$ | **0.0** $(0.0, 2.0)$ | **1.0** $(0.0, 5.0)$ | **3.0** $(0.0, 8.0)$ | **1.0** $(0.0, 4.0)$ |
| | tail | **0.0** $(0.0, 2.0)$ | **0.0** $(0.0, 5.0)$ | **4.0** $(0.0, 8.0)$ | **1.0** $(0.0, 5.0)$ | **1** $(0.0, 54.0)$ | **3.0** $(1.0, 6.0)$ |

Table 5.2: Confusion matrices reporting edge and endpoint counts of the output of COmbINE $\hat{\mathcal{H}}$ versus the actual summary graph $\mathcal{H}$, for 10 variables and 5 data sets of 1000 samples each. $\mathcal{H}$ was obtained using COmbINE with an oracle of conditional independence, and unconstrained maxK and maximum path length parameters. The table reports median values (bold) along with the 5 and 95 percentiles (in parenthesis). Results are in agreement with the metrics used for larger networks.

Learning performance as a function of the maximum path length is shown in Figure 5.4. Notice that when the path length is increased from 1 to 2 there is drop in the percentage of dashed endpoints, implying more orientations are possible. For length equal to 1, only unshielded and discriminating colliders are identified, while for length larger than 2 further orientations become possible thanks to reasoning with the inducing paths. When length is 1, notice that there are almost no dashed edges (except for the edges added in line 5 of Algorithm 3). When the maximum length increases,

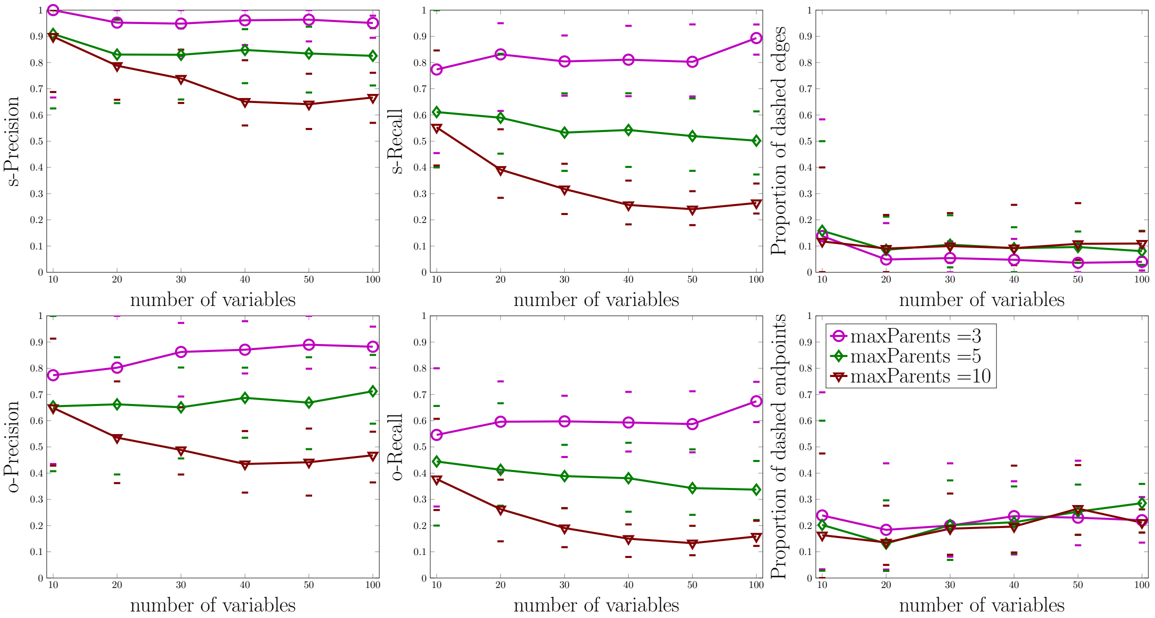


Figure 5.5: **Learning performance of COmbINE for various network sizes and densities**. From left to right: Median s-Precision, s-Recall, proportion of dashed edges (top) and o-Precision, o-Recall and proportion of dashed endpoints (bottom) for varying network size and density. Density is controlled by limiting the number of possible parents per variable. Coloured bars indicate 5 and 95 percentiles. As expected, the performance deteriorates as networks become denser.

adjacencies in one data set, can be explained with longer inducing paths in the underlying graph and more dashed edges appear. The learning performance of the algorithm is not monotonic with the maximum length. Explaining an association (adjacency) through the presence of a long inducing path may be necessary for asymptotic correctness. However, in the presence of statistical errors, allowing such long paths could lead to complicated solutions or the propagation of errors.

Overall, it seems any increase of the maximum path length above 3 does not significantly affect performance. It seems that a maximum path length of 3 is a reasonable trade-off among learning performance (precision and recall), percentage of uncertainties, and computational efficiency. These experiments justify our choice of maximum length 3 as the default parameter value of the algorithm.

### COmbINE performance as a function of network density and size

In Figure 5.5 the learning performance of the algorithm is presented as a function of network density and size. Density was controlled by the maximum parents allowed per variable, set by parameter maxParents during the generation of the random networks. For all network sizes, learning performance monotonically decreases with increased density, while the percentage of dashed features does not significantly vary. The size of the network has a smaller impact on the performance, particularly for the sparser networks. For dense networks, performance is relatively poor and becomes worse with larger sizes.

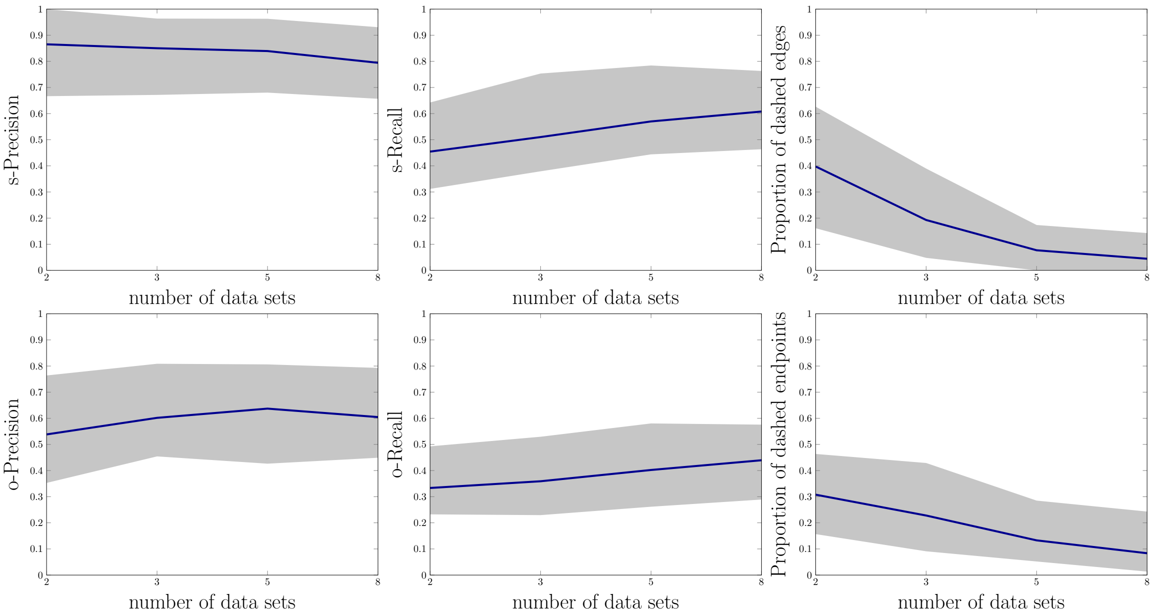


Figure 5.6: **Learning performance of COmbINE for varying number of input data sets**. From left to right: Median s-Precision, s-Recall, Proportion of dashed edges (top) and o-Precision, o-Recall and proportion of dashed endpoints of (bottom) for varying number of input data sets. Shaded area ranges from the 5 to the 95 percentile. Increasing the number of input data sets improves the performance of the algorithm.

We also calculated confusion matrices for edges and endpoints inferred by COmbINE against the *correct output* structure $\mathcal{H}$ for networks of 10 variables, where $\mathcal{H}$ can be obtained by running COmbINE with an oracle of conditional independence and unrestricted path length and conditioning set size. Table 5.2 shows the resulting confusion matrices for maxParents 3 and 5 and 5 data sets of sample size 1000. Overall, the results are in concordance with the metrics used for larger networks, and confirm that the method works best for sparser networks. Notice that for dense networks (for $N$=10 and maxParents =5, the networks have about 40% of all possible edges), there are cases where the actual correct output includes a large proportion of dashed edges, while constricting the maximum path length forces the algorithm to accept more solid features (hence the wide percentiles).

### COmbINE performance over sample size and number of input data sets

Figure 5.6 shows the performance of the algorithm with increasing the number of input data sets. As expected, the percentage of uncertainties (dashed features) is steadily decreasing with increased number of input data sets. Recall also steadily improves, while Precision is relatively unaffected. Figure 5.7 holds the number of input data set constant to the default value 5, while increasing the sample size per data set. Recall in particular improves with larger sample sizes, while the percentage of dashed endpoints drops.

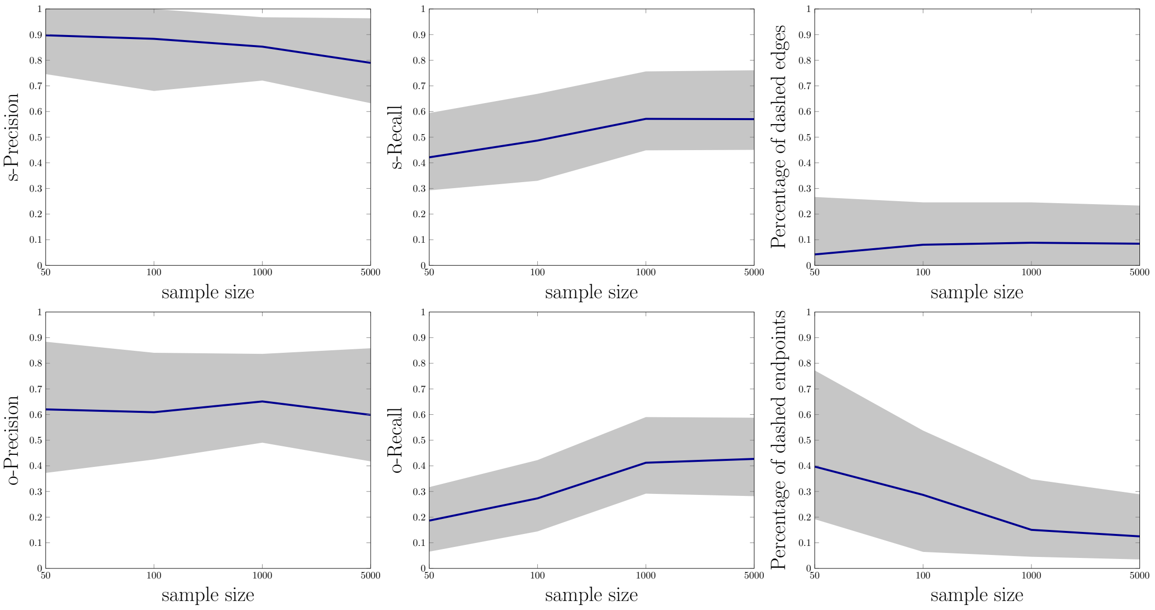


Figure 5.7: **Learning performance of COmbINE for varying sample size per data set**. From left to right: s-Precision, s-Recall, Proportion of dashed edges (top) and o-Precision, o-Recall and proportion of dashed endpoints of (bottom) for varying sample size per data set. Shaded area ranges from the 5 to the 95 percentile. Increasing the sample size improves the performance of the algorithm.

### COmbINE performance for increasing number of latent variables

We also examine the effect of confounding to the performance of COmbINE . To do so, we generated semi-Markov causal models instead of DAGs in the generation of the experiments: We generated random DAG networks of 30 variables and then marginalized out a percentage of the variables. Figure 5.8 depicts COmbINE's performance against 3, 6, and 9 of latent variables, corresponding to 10%, 20% and 30% of the total number of variables in the graph, respectively. Overall, confounding does not seem to greatly affect the performance of COmbINE. We must point out however, that s-Recall is lower than the s-Recall with no confounded variables for the same network size (see Figure 5.5).

### Running Time for COmbINE

The running time of COmbINE depends on several factors, including the ones examined in the previous experiments: Maximum path length, number of input data sets and sample size, and, naturally, the number of variables. Figure 5.9 illustrates the running time of COmbINE against these factors. As we can see in Figure 5.9, the restriction on the maximum path length is the most critical factor for the scalability of the algorithm.

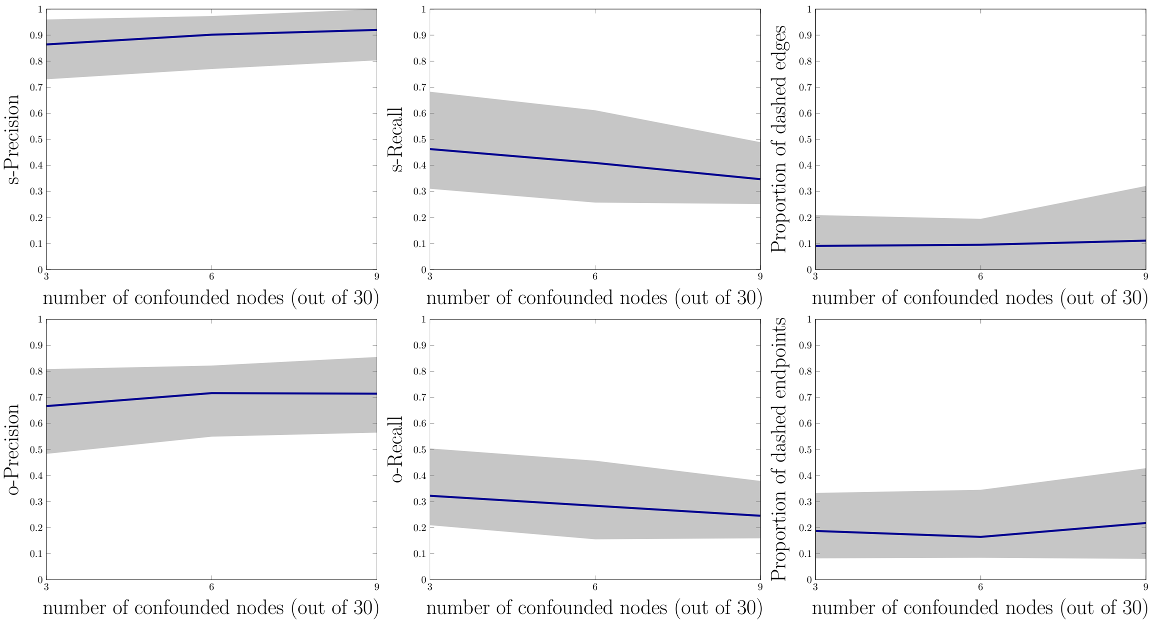
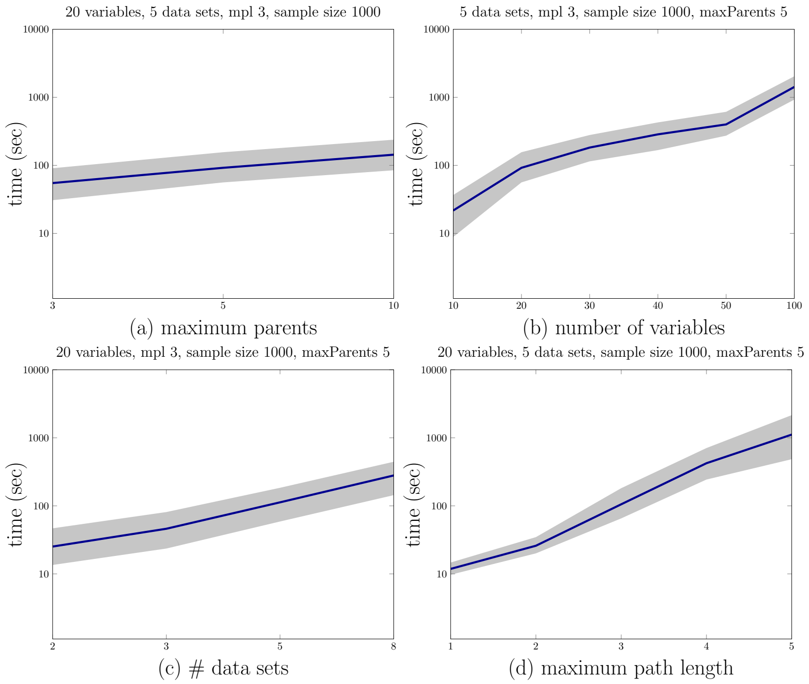


Figure 5.8: **Learning performance of COmbINE for varying percentage of confounded variables**. From left to right: s-Precision, s-Recall, percentage of dashed edges (top) and o-Precision, o-Recall and percentage of dashed endpoints (bottom) for varying number of confounded nodes for networks of 30 variables. Shaded area ranges from the 5 to the 95 percentile. Overall, the number of confounding variables does not seem to greatly affect the algorithm' s performance.

## 5.3 A case study: Mass Cytometry data

Mass cytometry (Bendall et al., 2011) is a recently introduced technique that enables measuring protein activity in cells, and its main use is to classify hematopoietic cells and identify signaling profiles in the immune system. Therefore, the proteins are usually measured in a sample of cells and then in a different sample of the same (type of) cells after they have been stimulated with a compound that triggers some kind of signaling behavior. Identifying the causal succession of events during cell signaling is crucial to designing drugs that can trigger or suppress immune reaction. Therefore in several studies both stimulated and un-stimulated cells are treated with several perturbing compounds to monitor the potential effect on the signaling pathway.

Mass cytometry data seem to be a suitable test-bed for causal discovery methods: The proteins are measured in single cells instead of representing tissue averages, the latter being known to be problematic for causal discovery (Chu et al., 2003), and the samples range in thousands. However, the mass cytometer can measure only up to 34 variables, which may be too low a number to measure all the variables involved in a signaling pathway. Moreover, about half of these variables are surface proteins that are necessary to distinguish (gate) the cells into sub-populations, but are not functional proteins involved in the signaling pathway. It is therefore reasonable for scientists to perform experiments measuring overlapping variable sets.



(a) maximum parents

(b) number of variables

(c) # data sets

(d) maximum path length

Figure 5.9: **Running time of COmbINE** . From left to right: Running time (in seconds) is plotted in logarithmic scale against maximum parents per variable and number of variables (top row); number of data sets and maximum path length (bottom row). Shaded area ranges from the 5 to the 95 percentile. The number of variables and the maximum path length seem to be the most critical factors of computational performance. Notice that, COmbINE scales up to problems with 100 total variables for limited path length and relatively sparse networks.

Bendall et al. (2011) and Bodenmiller et al. (2012) both use mass cytometry to measure protein abundance in cells of the immune system. In both studies, the samples were treated with several different signaling stimuli. Some of the stimuli were common in both studies. After stimulation with each activating compound, Bodenmiller et al. (2012) also test the cell's response to 27 inhibitors. One of these inhibitors is also used in Bendall et al. (2011). For this inhibitor, Bendall et al. (2011) measured bone marrow cell samples of a single donor. In Bodenmiller et al. (2012), measurements were taken from peripheral blood mononuclear cell (PBMC) samples of a (different) single donor. Despite differences in the experimental setup, the signaling pathway of every stimulus and every sub-population of cells is considered universal across (healthy) donors, so the data should reflect the same underlying causal structure.

We focused on two sup-populations of the cells, CD4+ and CD8+ T-cells, which are known to play a central role in immune signaling. The data were manually gated by the researchers in the original studies. We also focused on one of the stimuli present in both studies, PMA-Ionomycin, which is known to have prominent effects on T-cells. Proteins pBtk, pStat3, pStat5, pNfkb, pS6, pp38, pErk, pZap70, pSHP2 and pPlcg2 are measured in both data sets (initial p denotes that the concentration of the phosphorylated protein is measured). Four additional variables were included in the analysis, pAkt, pLat and pStat1 measured only in Bodenmiller et al. (2012) and pMAPK measured only in

| Data set | Source | latent ($\mathbf{L_i}$): | manipulated($\mathbf{I_i}$) | Donor |
|---|---|---|---|---|
| $\mathbf{D_1}$ | Bodenmiller et al. (2012) | pMAPK | pAkt | 1 |
| $\mathbf{D_2}$ | Bodenmiller et al. (2012) | pMAPK | pBtk | 1 |
| $\mathbf{D_3}$ | Bodenmiller et al. (2012) | pMAPK | pErk | 1 |
| $\mathbf{D_4}$ | Bendall et al. (2011) | pAkt, pLat, pStat1 | pErk | 2 |

Table 5.3: **Summary of the mass cytometry data sets co-analyzed with COmbINE**. The procedure was repeated for two sub-populations of cells, CD4+ cells and CD8+ cells.

Bendall et al. (2011). To be able to detect signaling behavior, we formed data sets that contain both stimulated and unstimulated samples.

As mentioned above, the cells were treated with several inhibitors. Some of these inhibitors target a specific protein, and some of them perturb the system in a more general or unidentified way. Specific inhibitors can be abundance inhibitors, which affect the level of measured protein, and activity inhibitors, which affect the function of measured proteins. The former are closer to ideal hard interventions. Activity inhibitors have been modelled in several ways in the literature. Sachs et al. (2005) model them as ideal interventions by manually setting the values to the lowest discretization level. Itani et al. (2010) propose splitting the target variable in two nodes, one used to represent the inhbition and the other used to represent the abundance. Mooij and Heskes (2013) propose modelling activity inhibitions by removing outgoing edges of the target variable. Notice that this type of modelling can be easily encoded in a SAT representation.

We used abundance inhibitors that we believe can be modeled as hard interventions (i.e. the compounds used to target these proteins are known to be specific and to have an effect in the phosphorylation levels of the target). The maximum dosage of each inhibitor was used. For all three interventions, the distribution of the target variable under zero dosage is differs significantly (according to a Kolmogorov-Smirnov test with significance threshold 0.05) from the distribution of the target variable for the maximum dosage, indicating that the inhibitor has an effect on the abundance of the target protein. Nevertheless, we must point out that the interventions may not be entirely ideal. More information on the specific compounds can be found in the respective publications.

We ended up with four data sets for each sub-population. Details can be found in Table 5.3. Protein interactions are typically non-linear, so we discretized the data into 4 bins. We ran Algorithm 2 with maximum path length 3. We used the $G^2$ test of independence for FCI with $\alpha = 0.05$ and maxK=5. We used Cytoscape (Smoot et al., 2011) to visualize the summary graphs produced by COmbINE, illustrated in Figure 5.10.

Unfortunately, the ground truth for this problem is not known for a full quantitative evaluation of the results. Nevertheless, this set of experiments demonstrates the availability of real and important data sets and problems that are suited integrative causal analysis. Second, these experiments provide a proof-of-concept for the specific algorithm. One type of interesting type of inference possible with COmbINE and similar algorithms is the prediction of a direct relation of pAkt and pMAPK in CD4+ cells, *even though the variables are not jointly measured in any of the input data sets.* Thus, methods for learning causal structure from multiple manipulations over overlapping variables potentially constitute a powerful tool in the field of mass cytometry.

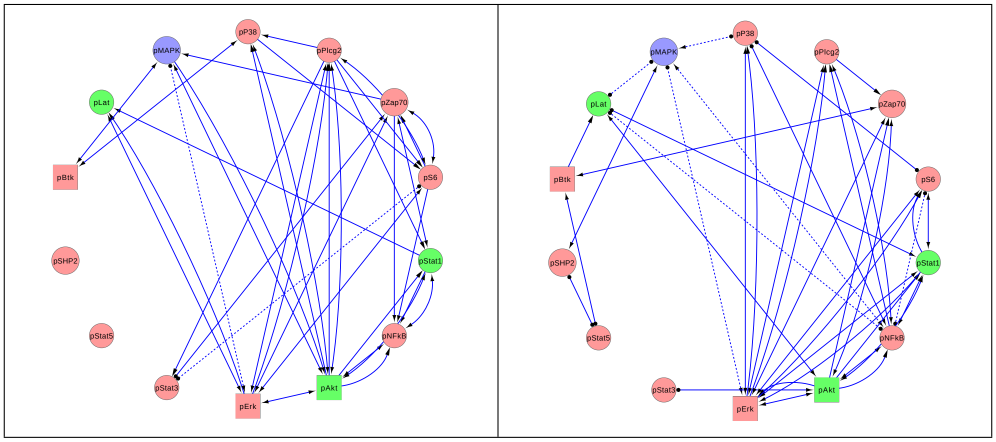


Figure 5.10: **A case study for COmbINE: Mass cytometry data**. COmbINE was run on 4 different mass cytometry data for two different cell populations: CD4+ T-cells (left) and CD8+ T-cells (right). In each data set, one variable was manipulated (pAkt, pBTk, pErk, pErk respectively). Variables pAkt, pLat and pStat1 are only measured in data sets 1-3, while pMAPK is only measured in data set 4.

We do not make any claims for the validity of the output graphs and they are presented only as a proof-of-concept, as there are several potential pitfalls. In addition to the potential imperfect manipulations described above, COmbINE also assumes lack of feedback cycles, which is not guaranteed in this system. We note however, that acyclic networks have been successfully used for reverse engineering protein pathways in the past (Sachs et al., 2005).

*Chapter 6*

# A non-trivial INCA prediction

*Unfortunately, for most real data sets, the causal ground truth is not known and cannot be established without experiments. Thus, validating the results of causal discovery algorithms is very difficult. In this chapter, we focus on a minimal scenario where COmbINE can predict significant association between variables never measured together. Furthermore, making additional parametric assumptions and employing the rules of path analysis allows the prediction of the strength of this association (i.e. the correlation coefficient of the two variables that are not jointly measured). We identify such cases in a 20 real-world data sets from a variety of scientific domains, and evaluate our predictions in held-out test data. We also compare against statistical matching, a family of methods able to produce similar estimates. Results indicate that causal assumptions produce predictions that are largely validated in real-world data.*

In the previous chapters, we presented the scope and motivation of integrative causal analysis of heterogeneous data sets. We also presented an algorithm that, given a set of overlapping data sets, can infer the possible underlying causal structures. The invariant characteristics of these causal structures are summarized using a *summary graph.* The validity of the algorithm given the causal Markov and Faithfulness assumptions is tested in simulated scenarios. But how often do these assumptions hold in *actual* data sets?

Validating causal discovery algorithms in real-world data sets is hardly possible, since the *true* causal structure is rarely established, particularly in multivariate systems. In this chapter, we identify a small scenario where INCA provides a testable prediction: A significant correlation between variables that have not been measured in the same data set. We then test this scenario in a variety of real-world data sets.

## 6.1 A Testable Scenario

We assume two i.i.d. data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ are provided on variables $\mathbf{O}_1 = \{X, Y, W\}$ and $\mathbf{O}_2 = \{X, Z, W\}$ respectively. We assume that the independence models of the data sets are $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$, in other words the one and only independence in $\mathcal{D}_1$ is $X \perp\!\!\!\perp W \mid Y$, and in $\mathcal{D}_2$ is $X \perp\!\!\!\perp W \mid Z$. Based on the input data it is possible to induce with existing

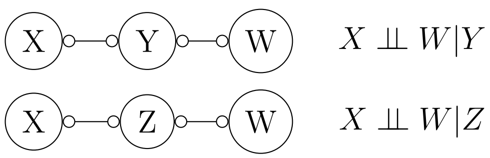


Figure 6.1: Definition of the co-analysis problem presented in Section 6.1: two observational i.i.d. data sets defined on variables $\mathbf{O}_1 = \{X, Y, W\}$ and $\mathbf{O}_2 = \{X, Z, W\}$ are used to identify the independence models $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$. These models are represented by PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$ shown in the figure. The problem is to identify one or all MAGs defined on $\mathbf{O} = \{X, Y, Z, W\}$ consistent with both $\mathcal{P}_1$ and $\mathcal{P}_2$.

causal analysis algorithms, such as FCI the following PAGs from each data set respectively:

$$\mathcal{P}_1 : X \circ - \circ Y \circ - \circ W$$

and

$$\mathcal{P}_2 : X \circ - \circ Z \circ - \circ W.$$

We will henceforth call this scenario **Example 6.1**. The PAGs corresponding to the input independence models are also shown graphically in Figure 6.1. The problem is to identify one or all MAGs defined on $\mathbf{O} = \{X, Y, Z, W\}$ consistent with the independence models $\mathcal{J}_1$ and $\mathcal{J}_2$, or equivalently, both PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$.

These two PAGs represent all the sound inferences possible about the structure of the data, when analyzing the data sets in isolation and independently of each other. But what are the possible causal structures over the union of variables $\mathbf{O} = \{X, Y, Z, W\}$ that are compatible with both independence models?

Since Example 6.1 does not include any manipulations, it can be solved by algorithms prior to SBCSD and COmbINE. The first algorithm to solve the problem is ION (Tillman et al., 2008), which identifies the set of PAGs (defined over $\mathbf{O}$) of all consistent MAGs. Subsequently, in Triantafillou et al. (2010), we proposed the algorithm Find Consistent MAG (FCM) that converts the problem to a satisfiability problem for improved computational efficiency. FCM returns one consistent MAG with all input PAGs. Similar ideas have been developed to learn joint structure from marginal structures in decomposable graphs such as undirected graphs (Kim and Lee, 2008) and Bayesian Networks (Kim, 2010). Going back to Example 6.1, Figure 6.2 shows all 14 consistent MAGs with the input PAGs in the scenario. The FCM algorithm arbitrarily returns one of them as the solution to the problem (of course, the algorithm can be easily modified to return all solutions). Figure 6.4 (right) shows the output of ION on the same problem.

COmbINE and SBCSD use SMCMs to represent causal relations. Since SMCMs have a many-to-one correspondence to MAGs, the set of possibly underlying SMCMs, shown in Figure 6.3, is larger than the set of possibly underlying MAGs. The summary graph returned by COmbINE is identical shown in Figure 6.4 (left).

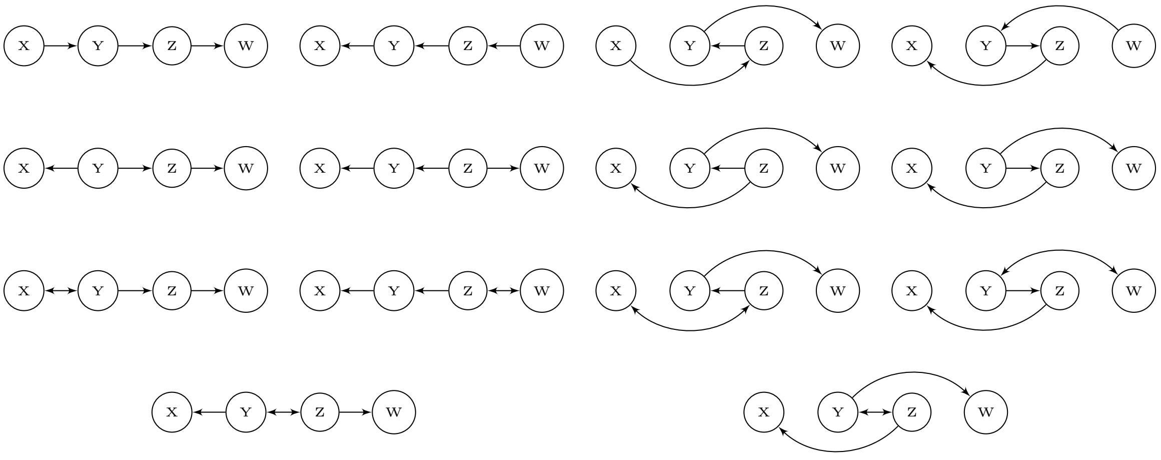


Figure 6.2: **Solution MAGS for the co-analysis problem of Example 6.1**: The 14 depicted MAGs are all and only the consistent MAGs with the PAGs shown in Figure 6.1. In all these MAGs the independencies $X \perp\!\!\!\perp W \mid Y$ and $X \perp\!\!\!\perp W \mid Z$ hold (and only them). Notice that, even though the edge $X - Y$ exists in $\mathcal{P}_1$ (Example 6.1), some of the consistent MAGs (the ones on the right of the figure) do not contain this edge: *adjacencies in the input PAGs do not simply transfer to the solution MAGs.* The FCM algorithm would arbitrarily output one of these MAGs as the solution of the problem of Example 6.1.

## 6.2 Predicting the Presence of Unconditional Dependencies

We now discuss how to implement the identification of the scenario presented in Section 6.1 to predict the presence of dependencies.

### Predictions of Dependencies

Recall that, in Example 6.1 we assume we are given two data sets on variables $\mathbf{O}_1 = \{X, Y, W\}$ and $\mathbf{O}_2 = \{X, Z, W\}$. We then determine, if possible, whether their independence models are respectively $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$ by a series of unconditional and conditional tests of independence. If this is the case, we predict an association between $Y$ and $Z$. The details of determining the independence model are important. Let us denote the $p$-value of an independence test with null hypothesis $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ as $p_{X \perp\!\!\!\perp Y \mid \mathbf{Z}}$. In the algorithms that follow, we make statistical decisions with the following rules:

- If $p_{X \perp\!\!\!\perp Y \mid \mathbf{Z}} \leq \alpha$ conclude $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$ (reject the null hypothesis).
- If $p_{X \perp\!\!\!\perp Y \mid \mathbf{Z}} \geq \beta$ conclude $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ (accept the null hypothesis).
- Otherwise, forgo making a decision.

The details are shown in Algorithm 8 named Full-Testing Rule, or FTR for short. We note a couple of observations. First, the algorithm is opportunistic. It does not produce a prediction whenever possible, but only for the case presented in Example 6.1. In addition, it makes a prediction only

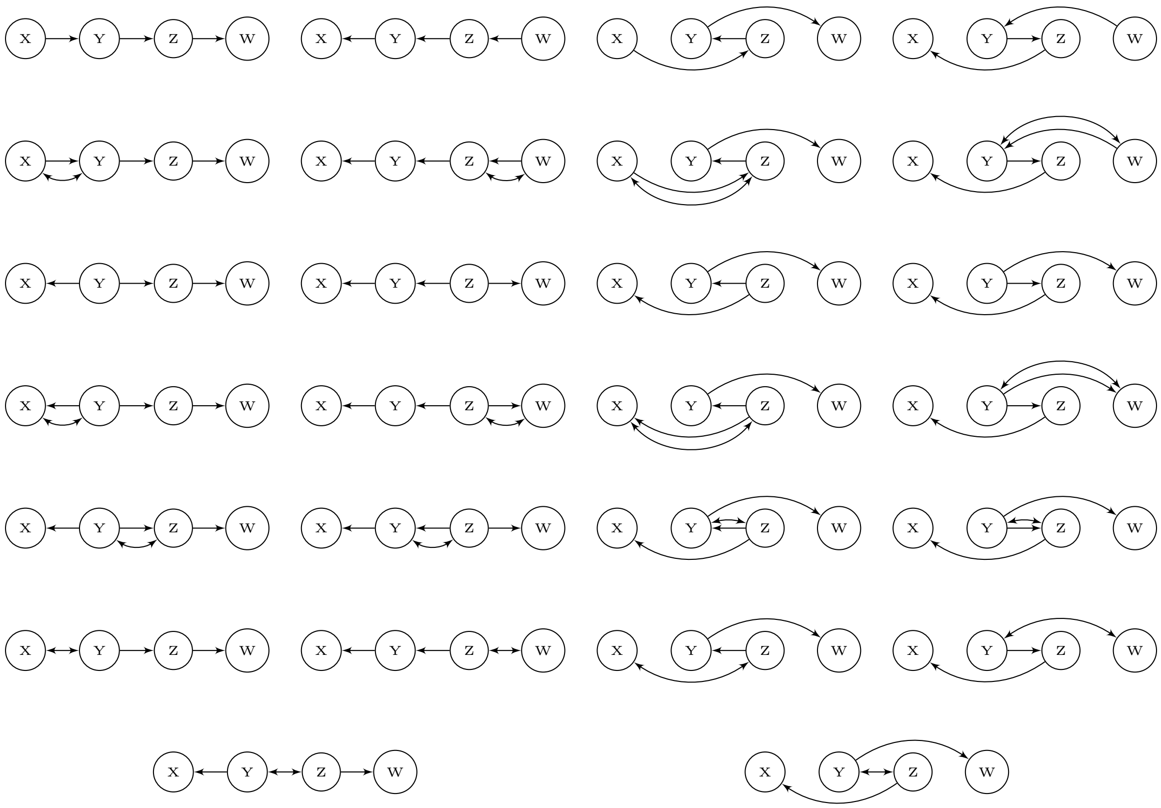


Figure 6.3: **Solution SMCMs for the co-analysis problem of Example 6.1**: The 26 depicted SMCMs are all and only the consistent SMCMs with the PAGs shown in Figure 6.1. In all these MAGs the independencies $X \perp\!\!\!\perp W \mid Y$ and $X \perp\!\!\!\perp W \mid Z$ hold (and only them).

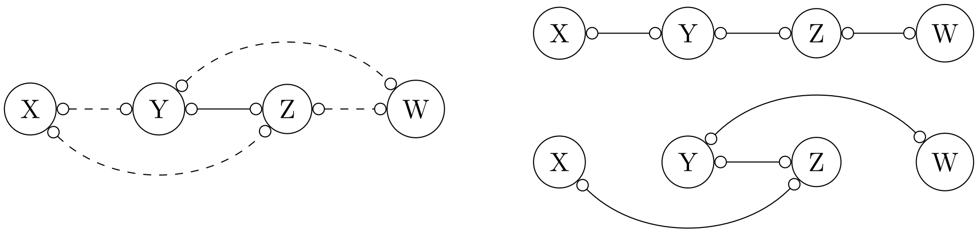


Figure 6.4: **Different summaries of the consistent SMCMs for Example 6.1**.(left) Summary graph representing the set of consistent SMCMs of Example 6.1. This graph is the output of the COmbINE algorithm on the problem of Example 6.1. Alternatively, the set of consistent MAGs can be represented with two PAGs (right). This is the output of the ION algorithm on the same problem.

when the $p$-values of the tests are either too high or too low to relatively safely accept dependencies and independencies. Second, to accept an independence model, for example, that $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$ all possible conditional and unconditional tests among the variables are performed. If any of these

**Algorithm 8:** Predict Dependency: Full-Testing Rule (**FTR**)

```
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X, Y, W\}$ and $\{X, Z, W\}$, respectively
if in $\mathcal{D}_1$ we conclude
        // determine whether $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$
        $X \perp\!\!\!\perp W \mid Y$ , $X \not\perp\!\!\!\perp Y \mid \emptyset$ , $Y \not\perp\!\!\!\perp W \mid \emptyset$ , $X \not\perp\!\!\!\perp W \mid \emptyset$ , $X \not\perp\!\!\!\perp Y \mid W$ , $Y \not\perp\!\!\!\perp W \mid X$
    and in $\mathcal{D}_2$ we conclude
        // determine whether $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$
        $X \perp\!\!\!\perp W \mid Z$ , $X \not\perp\!\!\!\perp Z \mid \emptyset$ , $Z \not\perp\!\!\!\perp W \mid \emptyset$ , $X \not\perp\!\!\!\perp W \mid \emptyset$ , $X \not\perp\!\!\!\perp Z \mid W$ , $Z \not\perp\!\!\!\perp W \mid X$
then
    Predict $Y \not\perp\!\!\!\perp Z \mid \emptyset$
    Predict either $(X \circ - \circ Y \circ - \circ Z \circ - \circ W)$ or $(X \circ - \circ Z \circ - \circ Y \circ - \circ W)$ holds
else
    Do not make a prediction
end
```

tests is inconclusive or contradictory to $\mathcal{J}_1$, the latter is not accepted and no prediction is made. In the terminology of Spirtes et al. (2000), we test for a *detectable failure of faithfulness.* Similar ideas have also been devised in Ramsey et al. (2006) and Spanos (2006). This rule characteristic is important in case one would like to generalize these ideas to larger graphs and sets of variables: performing all possible tests becomes quickly prohibitive, and the probability of statistical errors increases.

If however, one assumes the Faithfulness Condition holds among variables $\{X, Y, Z, W\}$, then it is not necessary to perform all such tests to determine the independence models. Algorithms for inducing graphical models from data, such as FCI and PC (Spirtes et al., 2000) are based on this observation to gain computational efficiency. The Minimal-Testing Rule, MTR for short, performs only a minimal number of tests that together with Faithfulness may entail that $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$ and lead to a prediction. The details are shown in Algorithm 9.

**Algorithm 9:** Predict Dependency Minimal-Testing Rule (**MTR**)

```
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X, Y, W\}$ and $\{X, Z, W\}$, respectively
if in $\mathcal{D}_1$ we conclude
        // determine whether $\mathcal{J}_1 = \{\langle X, W|Y\rangle\}$
        $X \perp\!\!\!\perp W \mid Y$ , $X \not\perp\!\!\!\perp Y \mid \emptyset$ , $Y \not\perp\!\!\!\perp W \mid \emptyset$
    and in $\mathcal{D}_2$ we conclude
        // determine whether $\mathcal{J}_2 = \{\langle X, W|Z\rangle\}$
        $X \perp\!\!\!\perp W \mid Z$ , $X \not\perp\!\!\!\perp Z| \mid$ emptyset , $Z \not\perp\!\!\!\perp W| \mid$ emptyset
then
    Predict $Y \not\perp\!\!\!\perp Z \mid \emptyset$
    Predict either $(X \circ - \circ Y \circ - \circ Z \circ - \circ W)$ or $(X \circ - \circ Z \circ - \circ Y \circ - \circ W)$ holds
else
    Do not make a prediction
end
```

### Heuristic Predictions of Dependencies Based on Transitivity

Is it really necessary to develop and employ the theory presented to make such predictions? Could there be other simpler and intuitive rules that are as predictive, or more predictive? For example, a common heuristic inference people are sometimes willing to make is the transitivity rule: if $Y$ is correlated with $X$ and $X$ is correlated with $Z$, then predict that $Y$ is also correlated with $Z$. The FTR and MTR rules defined also check these dependencies: $X \not\perp\!\!\!\perp Y \mid \emptyset$ in $\mathcal{D}_1$ and $X \not\perp\!\!\!\perp Z \mid \emptyset$ in $\mathcal{D}_1$, so one could object that any success of the rules could be attributed to the transitivity property often holding in Nature. We implement the Transitivity Rule (TR), shown in Algorithm 10 to compare against the INCA-based FTR and MTR rules. Obviously, the Transitivity Rule is not sound in general,[1] but on the other hand, FTR and MTR are also based on the assumption of faithfulness, which may as well be unrealistic. The verdict will be determined by experimentation.

**Algorithm 10:** Predict Dependency Transitivity Rule (**TR**)

```
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{Y, X\}$ and $\{X, Z\}$, respectively
if in $\mathcal{D}_1$: $Y \not\perp\!\!\!\perp X \mid \emptyset$ and in $\mathcal{D}_2$: $X \not\perp\!\!\!\perp Z \mid \emptyset$ then
    Predict $Y \not\perp\!\!\!\perp Z \mid \emptyset$
else
    Do not make a prediction
end
```

### Empirical Evaluation of Predicting Unconditional Dependencies

We have applied and evaluated the three rules against each-other as well as random predictions (prior probability of a pair being dependent) on real data, in a way that becomes testable. Specifically, given a data set $\mathcal{D}$ we randomly partition its samples to three data sets of equal size, $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_t$. The latter is hold out for testing purposes. In the first two data sets, we identify quadruples of variables $\{X, Y, Z, W\}$ for which the Full-Testing and the Minimal-Testing Rules apply. Notice that, the two rules perform tests among variables $\{X, Y, W\}$ in $\mathcal{D}_1$ and among variables $\{X, Z, W\}$ in $\mathcal{D}_2$; *the rules do not access the joint distribution of* $Y, Z$. Similarly, for the Transitivity Rule we identify triplets $\{X, Y, Z\}$ where the rule applies. Subsequently, we measure the predictive performance of the rules. In more detail:

- *Data Sets*: We selected data sets in an attempt to cover a wide range of sample-sizes, dimensionality (number of variables), types of variables, domains, and tasks. The decision for inclusion depended on availability of the data, ease of parsing and importing them. *No data set was a posteriori removed out of the study, once selected.* Table 6.1 assembles the list of data sets and their characteristics before preprocessing. Some minimal preprocessing steps were applied to several data sets. Details can be found in Tsamardinos et al. (2012).
- *Tests of Independence*: For discrete variables we have used the $G^2$-test (a type of likelihood ratio test) with an adjustment for the degrees-of-freedom used in Tsamardinos et al. (2006) and presented in detail in Tsamardinos and Borboudakis (2010). For continuous variables we have used a test based on the Fisher z-transform of the partial correlation as described in Spirtes

[1] The Transitivity Rule should be sound when the marginal of the three variables is faithful to a Markov Random Field.

| Name | Reference | # istances | # vars | Group Size | Vars type | Scient. domain |
|---|---|---|---|---|---|---|
| Covtype | Blackard and Dean (1999) | 581012 | 55 | 55 | N/O | Agricultural |
| Read | Guvenir and Uysal (2000) | 681 | 26 | 26 | N/C/O | Business |
| Infant-mortality | Mani and Cooper (2004) | 5337 | 83 | 83 | N | Clinical study |
| Compactiv | Alcalá-Fdez et al. (2009) | 8192 | 22 | 22 | C | Computer science |
| Gisette | Guyon et al. (2006a) | 7000 | 5000 | 50 | C | Digit recognition |
| Hiva | Guyon et al. (2006b) | 4229 | 1617 | 50 | N | Drug discovering |
| Breast-Cancer | Wang (2005) | 286 | 17816 | 50 | C | Gene expression |
| Lymphoma | Rosenwald et al. (2002) | 237 | 7399 | 50 | C | Gene expression |
| Wine | Cortez et al. (2009) | 4898 | 12 | 12 | C | Industrial |
| Insurance-C | Elkan (2001) | 9000 | 84 | 84 | N/O | Insurance |
| Insurance-N | Elkan (2001) | 9000 | 86 | 86 | N/O | Insurance |
| p53 | Danziger et al. (2009) | 16772 | 5408 | 50 | C | Protein activity |
| Ovarian | Conrads (2004) | 216 | 2190 | 50 | C | Proteomics |
| C&C | Frank and Asuncion (2010) | 1994 | 128 | 128 | C | Social science |
| ACPJ | Aphinyanaphongs et al. (2006) | 15779 | 28228 | 50 | C | Text mining |
| Bibtex | Tsoumakas et al. (2010) | 7395 | 1995 | 50 | N | Text mining |
| Delicious | Tsoumakas et al. (2010) | 16105 | 1483 | 50 | N | Text mining |
| Dexter | Guyon et al. (2006a) | 600 | 11035 | 50 | N | Text mining |
| Nova | Guyon et al. (2006b) | 1929 | 12709 | 50 | N | Text mining |
| Ohsumed | Joachims (2002) | 5000 | 14373 | 50 | C | Text mining |

Table 6.1: Data Sets included in empirical evaluation of Section 5. N- Nominal, O - Ordinal, C - Continuous.

et al. (2000). The two tests employed are typical in the graphical learning literature. In some cases ordinal variables were treated as continuous, while in others the continuous variables were discretized (Tsamardinos et al., 2012, see) so that every possible quadruple $\{X, Y, Z, W\}$ was either treated as all continuous variables or all discrete and one of the two tests above could be applied.

- *Significance Thresholds*: There are two threshold parameters: level $\alpha$ below which we accept dependence and level $\beta$ above which we accept independence; the TR rule only employs the $\alpha$ parameter. For FTR these thresholds were always set to $\alpha_{FTR} = 0.05$ and $\beta_{FTR} = 0.3$ without an effort to optimize them. Some minimal anecdotal experimentation with FTR showed that the performance of the algorithm is relative insensitive to the values of $\alpha_{FTR}$ and $\beta_{FTR}$ and the algorithm works without fine-tuning. Notice that FTR requires 10 dependencies and 2 independencies to be identified, while MTR requires 4 dependencies and 2 independencies, and TR requires 2 dependencies to be found. Thus, FTR is more conservative than MTR and TR for the same values of $\alpha$ and $\beta$. The Bonferroni correction for MTR dictates that $\alpha_{MTR} = \alpha_{FTR} \times \frac{4}{10} = 0.02$, while for TR we get $\alpha_{TR} = \alpha_{FTR} \times \frac{2}{10} = 0.01$ (TR however, does not require any independencies present so this adjustment may not be conservative enough). We run MTR with threshold values $\alpha_{MTR} \in \{0.05, 0.02, 0.002, 0.0002\}$, that is equal to the threshold of FTR, with the Bonferroni adjustment, and stricter than Bonferroni by one and two orders of magnitude. The $\beta_{MTR}$ parameter is always set to 0.3. In a similar fashion for TR, we set $\alpha_{TR} \in \{0.05, 0.01, 0.001, 0.0001\}$.

- *Identifying Quadruples*: In low-dimensional data sets (number of variables less than 150), we check the rules on all quadruples of variables. This is time-prohibitive however, for the larger data sets. In such cases, we randomly permute the order of variables and partition them into groups of 50 and consider quadruples only within these groups. The column named "Group Size" in Table 6.1 notes the actual sizes of the variable groups used.

- *Measuring Performance*: The ground truth for the presence of a predicted correlation is not known. We thus seek to statistically evaluate the predictions. Specifically, for each predicted pair of variables $X$ and $Y$, we perform a test of independence in the corresponding hold-out test set $\mathcal{D}_t$ and store its $p$-value $p_{X \perp\!\!\!\perp Y \mid \emptyset}$. The lower the $p$-value the higher the probability the pair is truly correlated. We consider as "accurate" a prediction whose $p$-value is less than a threshold $t$ and we report the accuracy of each rule.

**Definition 6.2.1 (Prediction Accuracy)** *We denote with $M_i^R$ and $U_i^R$ the multiset and set respectively of p-values of the predictions of rule $R$ applied on data set $i$. The p-values are computed on the hold-out test set. The accuracy of the rule on data set $i$ at threshold $t$ is defined as:*

$$Acc_i^R(t) = \#\{p <= t, p \in M_i^R\} / |M_i^R|.$$

*We also define the* average accuracy *over all data sets (each data set is weighted the same)*

$$\overline{Acc}^R(t) = \frac{1}{20} \sum_{i=1}^{20} Acc_i^R(t)$$

*and the* pooled accuracy *over the union of predictions (each prediction is weighted the same)*

$$\underline{Acc}^R(t) = \#\{p <= t, i = 1 \ldots 20, p \in M_i^R\} / \sum_i |M_i^R|.$$

The reason $M_i^R$ is defined as a multiset stems from the fact that a dependency $Y \not\perp\!\!\!\perp Z \mid \emptyset$ may be predicted multiple times if a rule applies to several quadruples $\{X_i, Y, Z, W_i\}$ or triplets $\{X_i, Y, Z\}$ (for the Transitivity Rule). The number of predictions of each rule $R$ (i.e., $|M_i^R|$) and the number of unique pairs $X - Y$ predicted correlated (i.e., $|U_i^R|$ ) are shown in Table 6.2. In some cases (e.g., data sets Read and ACPJ) the Full-Testing Rule does not make any predictions. Overall however, the rules typically make hundreds or even thousands of predictions.

*Overall Performance*: The accuracies at $t = 0.05$, $Acc_i(t)$, $\overline{Acc}(t)$, and $\underline{Acc}(t)$ for the three rules as well as the one achieved by guessing at random are shown in Figure 6.5. The Bonferroni adjusted thresholds for MTR and TR were used: $\alpha_{FTR} = 0.05, \alpha_{MTR} = 0.02, \alpha_{TR} = 0.01$. Over all predictions, the Full-Testing Rule achieves accuracy 96%, consistently higher than guessing at random, the MTR and the TR. The same results are also depicted in tabular form in Table 6.3, where additionally, the statistical significance is noted. The null hypothesis is that $Acc_i^{FTR}(0.05) \leq Acc_i^R(0.05)$, for $R$ being MTR or TR. The one-tail Fisher's exact test (Fisher, 1922) is employed when computationally feasible, otherwise the Pearson $\chi^2$ test (Pearson, 1900) is used instead. FTR is typically performing statistically significantly better than all other rules.

*Sensitivity to the $\alpha$ parameter*: The results are not particularly sensitive to the significance thresholds used for $\alpha$ for MTR and TR. Figures 6.8 (a-b) show the average accuracy $\overline{Acc}$ and the pooled accuracy $\underline{Acc}$ as a function of the *alpha* parameter used: no correction, Bonferroni correction, and stricter than Bonferroni by one and two orders of magnitude. The accuracy of MTR and TR improves as they become more conservative but never reaches the one by FTR even for the stricter thresholds of $\alpha_{MTR} = 0.0002$ and $\alpha_{TR} = 0.0001$.

| | # predictions $\lvert M_i^R \rvert$ | | | # unique predictions $\lvert U_i^R \rvert$ | | |
|---|---|---|---|---|---|---|
| Data Set | FTR$_{0.05}$ | MTR$_{0.02}$ | TR$_{0.01}$ | FTR$_{0.05}$ | MTR$_{0.02}$ | TR$_{0.01}$ |
| Covtype | 222 | 33277 | 54392 | 59 | 810 | 1431 |
| Read | 0 | 9 | 4713 | 0 | 9 | 260 |
| Infant Mortality | 22 | 2038 | 3736 | 10 | 427 | 1170 |
| Compactiv | 135 | 679 | 3950 | 69 | 193 | 231 |
| Gisette | 423 | 35824 | 134213 | 330 | 12340 | 31648 |
| hiva | 554 | 65967 | 151582 | 366 | 16174 | 34977 |
| Breast-Cancer | 1833 | 141643 | 470212 | 1371 | 68077 | 228610 |
| Lymphoma | 7712 | 188216 | 394572 | 4473 | 51794 | 122857 |
| Wine | 4 | 73 | 431 | 3 | 44 | 66 |
| Insurance-C | 1839 | 30569 | 40173 | 394 | 2212 | 3264 |
| Insurance-N | 226 | 18270 | 47115 | 95 | 1002 | 2527 |
| p53 | 46647 | 1645476 | 1995354 | 15181 | 95195 | 129372 |
| Ovarian | 539165 | 1604131 | 2015133 | 41600 | 48376 | 52646 |
| C&C | 99241 | 416934 | 301218 | 4168 | 5048 | 5050 |
| ACPJ | 0 | 219 | 16574 | 0 | 190 | 15994 |
| Bibtex | 1 | 3982 | 25948 | 1 | 1858 | 16087 |
| Delicious | 856 | 32803 | 105776 | 524 | 6042 | 21351 |
| Dexter | 0 | 2 | 117 | 0 | 2 | 116 |
| Nova | 0 | 124 | 3473 | 0 | 115 | 3280 |
| Ohsumed | 0 | 64 | 5358 | 0 | 60 | 5227 |

Table 6.2: Number of predictions $|M_i^R|$ and number of unique predictions $|U_i^R|$ with "Bonferroni" correction for rules FTR, MTR and TR. The rules typically make hundreds or even thousands of predictions.

*Sensitivity to $t$*: The results are also not sensitive to the particular significance level $t$ used to define accuracy. Figure 6.6 graphs $Acc_i^R(t)$ over $t = [0, 0.05]$ for two typical data sets as well as $\underline{Acc}(t)$ and $\overline{Acc}(t)$. The situation is similar and consistent across all data sets considered, which are shown in Figure 6.7. The lines of the Full Testing Rule rise sharply, which indicates that the $p$-values of its predictions are concentrated close to zero.

*Explaining the difference of FTR and MTR*: Asymptotically and when the data distribution is faithful to a MAG, the FTR and the MTR rules are both sound (100% accurate). However, when the distribution is not faithful, the performance difference could become large because FTR tests for faithfulness violations as much as possible in an effort to avoid false predictions. This may explain the large differences in accuracies observed in the Infant Mortality, Gisette, Hiva, Breast-Cancer, and Lymphoma data sets. When the distribution is faithful, but the sample is finite, we expect some but small differences. For example when MTR falsely determines that $X \not\perp\!\!\!\perp Y \mid \emptyset$ due to a false positive test, the FTR rule still has a chance to avoid an incorrect prediction by additionally testing $X \not\perp\!\!\!\perp Y \mid W$. To support this theoretical analysis we perform experiments with simulated data where the network structure is known. Specifically, we employ the structure of the ALARM (Beinlich et al., 1989), INSURANCE (Binder et al., 1997) and HAILFINDER (Abramson et al., 1996) Bayesian networks. We sample 20 continuous and 20 discrete pairs of data sets $D_1$ and $D_2$ from distributions faithful to the network structure using different randomly chosen parameterizations for

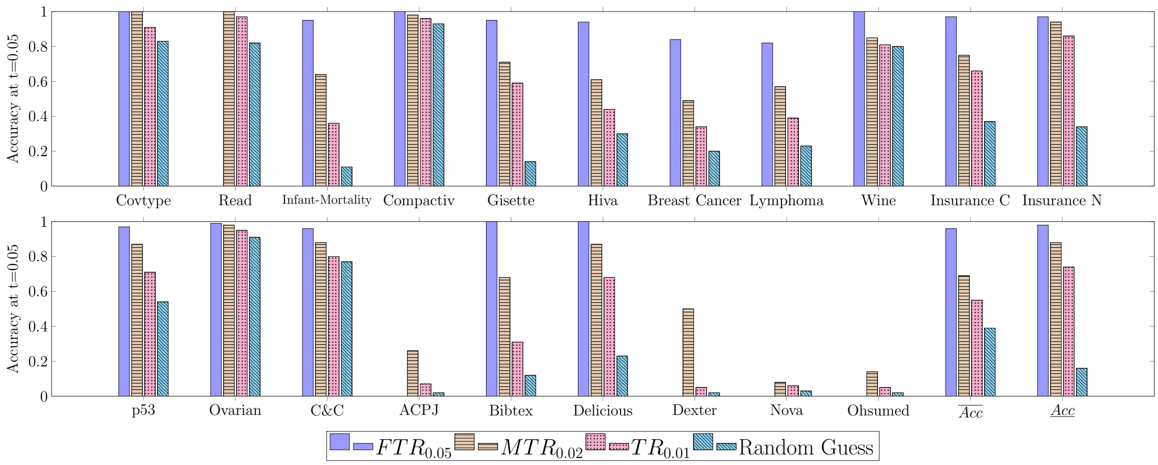


Figure 6.5: **Rules accuracy per data set.** Accuracies $Acc_i$ for each data set, as well as the average accuracy $\overline{Acc}$ (each data set weighs the same) and the pooled accuracy $\underline{Acc}$ (each prediction weighs the same). All accuracies are computed as threshold $t = 0.05$. FTR's accuracy is always above 80% and always higher than MTR, TR, and random guess.

the continuous case, and the original network parameters for the discrete case. We do the same for sample sizes 100, 500, 1000. Subsequently, we apply the FTR and MTR rules with $\alpha_{FTR} = 0.05$ and $\alpha_{MTR} = 0.02$ (Bonferroni adjusted) on each pair of $D_1$ and $D_2$ and all possible quadruples of variables. The true accuracy is not computed on a test data set $D_t$ but on the known graph instead by checking whether $Y$ and $Z$ are $d$-connected given $X$ and $W$. The mean true accuracies over all samplings are reported in Figure 6.9. The difference in performance on the faithful, simulated data is usually below 5%. In contrast, the largest difference in performance on the real data sets is over 35% (Breast-Cancer), while the difference of the pooled accuracies is 10%. Thus, violations of faithfulness seem to be the most probable explanation for the large difference in accuracy on the real data.

## Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:

- Notice that even if all predicted pairs are truly correlated, the accuracy may not reach 100% due to the presence of Type II errors (false negatives) *in the test set.*

- The FTR rule performs the test for the X-W association independently in both data sets. Given that the data in our experiments come from exactly the same distribution, they could be pooled together to perform a single test; alternatively, if this is not appropriate, the p-values of the tests could be combined to produce a single p-value (Tillman, 2009; Tsamardinos and Borboudakis, 2010).

- The results show that *the Full-Testing Rule accurately predicts the presence of dependencies*, statistically significantly better than random predictions, across all data sets, regardless of the type of data or the idiosyncracies of a domain. The rule is successful in gene-expression

| Data Set | $\text{FTR}_{0.05}$ | $\text{MTR}_{0.02}$ | $\text{TR}_{0.01}$ | Random Guess |
|---|---|---|---|---|
| Covtype | 1.00 | 1.00 | $0.91^{**}$ | $0.83^{**}$ |
| Read | - | 1.00 | 0.97 | 0.82 |
| Infant Mortality | 0.95 | $0.64^{**}$ | $0.36^{**}$ | $0.11^{\spadesuit}$ |
| Compactiv | 1.00 | 0.98 | $0.96^{*}$ | $0.93^{**}$ |
| Gisette | 0.95 | $0.71^{\spadesuit}$ | $0.59^{\spadesuit}$ | $0.14^{\spadesuit}$ |
| hiva | 0.94 | $0.61^{\spadesuit}$ | $0.44^{\spadesuit}$ | $0.30^{\spadesuit}$ |
| Breast-Cancer | 0.84 | $0.49^{\spadesuit}$ | $0.34^{\spadesuit}$ | $0.20^{\spadesuit}$ |
| Lymphoma | 0.82 | $0.57^{\spadesuit}$ | $0.39^{\spadesuit}$ | $0.23^{\spadesuit}$ |
| Wine | 1.00 | 0.85 | 0.81 | 0.80 |
| Insurance-C | 0.97 | $0.75^{\spadesuit}$ | $0.66^{\spadesuit}$ | $0.37^{\spadesuit}$ |
| Insurance-N | 0.97 | $0.94^{*}$ | $0.86^{**}$ | $0.34^{\spadesuit}$ |
| p53 | 0.97 | $0.87^{\spadesuit}$ | $0.71^{\spadesuit}$ | $0.54^{\spadesuit}$ |
| Ovarian | 0.99 | $0.98^{\spadesuit}$ | $0.95^{\spadesuit}$ | $0.91^{\spadesuit}$ |
| C&C | 0.96 | $0.88^{\spadesuit}$ | $0.80^{\spadesuit}$ | $0.77^{\spadesuit}$ |
| ACPJ | - | 0.26 | 0.07 | 0.02 |
| Bibtex | 1.00 | 0.68 | 0.31 | $0.12^{**}$ |
| Delicious | 1.00 | $0.87^{\spadesuit}$ | $0.68^{\spadesuit}$ | $0.23^{\spadesuit}$ |
| Dexter | - | 0.50 | 0.05 | 0.02 |
| Nova | - | 0.08 | 0.06 | 0.03 |
| Ohsumed | - | 0.14 | 0.05 | 0.02 |
| $\overline{ACC^R}$ | 0.96 | $0.69^{**}$ | $0.55^{**}$ | $0.39^{**}$ |
| $\underline{ACC^R}$ | 0.98 | $0.88^{\spadesuit}$ | $0.74^{\spadesuit}$ | $0.16^{\spadesuit}$ |

Table 6.3: **Rules accuracy after "Bonferroni" correction.** $ACC_i^R(t)$ at $t = 0.05$ with "Bonferroni" correction for rules FTR, MTR, TR and Random Guess. Marks *, **, and $\spadesuit$ denote a statistically significant difference from FTR at the levels of 0.05, 0.01, and machine-epsilon respectively.

data, mass-spectra data measuring proteins, clinical data, images and others. The accuracy of predictions is robustly always above 0.80 and over all predictions it is 0.96; the difference with random predictions is of course more striking in data sets where the percentage of correlations (prior probability) is relatively small, as there is more room for improvement.

- *The Full-Testing Rule is noticeably more accurate than the Minimal-Testing Rule*, due to testing whether the Faithfulness Condition holds in the induced PAGs. The result is important considering that most constraint-based algorithms assume the Faithfulness Condition to induce models, *but do not check whether the induced model is Faithful.* These results indicate that when the latter is not the case, the model (and its predictions) may not be reliable. On the other hand, the FTR rule is also noticeably more conservative: the number of predictions it makes is significantly lower than the one made by MTR. In some data sets (e.g., Compactiv, Insurance-N, and Ovarian) by using the MTR vs. the FTR one sacrifices a small percentage of accuracy (less than 3% in these cases) to gain one order of magnitude more predictions. However, caution should be exercised because in certain data sets MTR is over 35% less accurate than FTR.

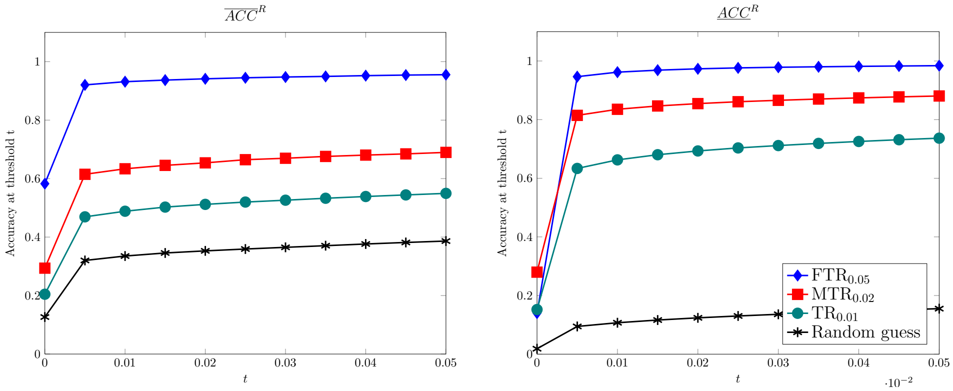


Figure 6.6: **Averaged and pooled accuracy per rule, for varying dependence threshold.** Average accuracy $\overline{ACC}^R(t)$ and pooled $\underline{ACC}^R(t)$. Predicted dependencies have $p$-values concentrated close to zero. The performance differences are insensitive to the threshold $t$ in the performance definition.

- *The Full-Testing Rule is more accurate than the Transitivity Rule.* Thus, the performance of the Full-Testing Rule cannot be attributed to simply performing a super-set of the tests performed by the Transitivity Rule.

- *Predictions are the norm case and not occur in contrived or rare cases only.* Even though there were few or no predictions for a couple of data sets, there are typically hundreds or thousands of predictions for each data set. This is the case despite the fact that we are only looking for a special-case structure and the search for these structures is limited within groups of 50 variables for the larger data sets. The results are consistent with the ones in Triantafillou et al. (2010), where larger structures were induced from simulated data.

- *FTR makes almost no predictions in the text data*:[2] this actually makes sense and is probably evidence for the validity of the method: it is semantically hard to interpret the presence of a word "causing" another word to be present.[3]

- FTR is an opportunistic algorithm that sacrifices completeness to increase accuracy, as well as improve computational efficiency and scalability. General algorithms for co-analyzing data over overlapping variable sets, such as ION (Tillman et al., 2008), IOD (Tillman and Spirtes, 2011) and cSAT (Triantafillou et al., 2010) could presumably make more predictions, and more general types of predictions (e.g., also predict independencies). However, their computational and learning performance on a wide range of domains and high-dimensional data sets is still an open question and an interesting future direction to pursue.

[2] The only predictions in text data are in Bibtex (1 prediction) and in Delicious (856), which are the only text data sets that are actually not purely bag-of-words data sets but include variables corresponding to tags. 66% of the predictions made in Delicious involves tag variables, as well as the single prediction in Bibtex.

[3] However, causality between words is still conceivable in our opinion: deciding to include a word in a document may change a latent variable corresponding to a mental state of the author, which in turn causes her to include some other word.

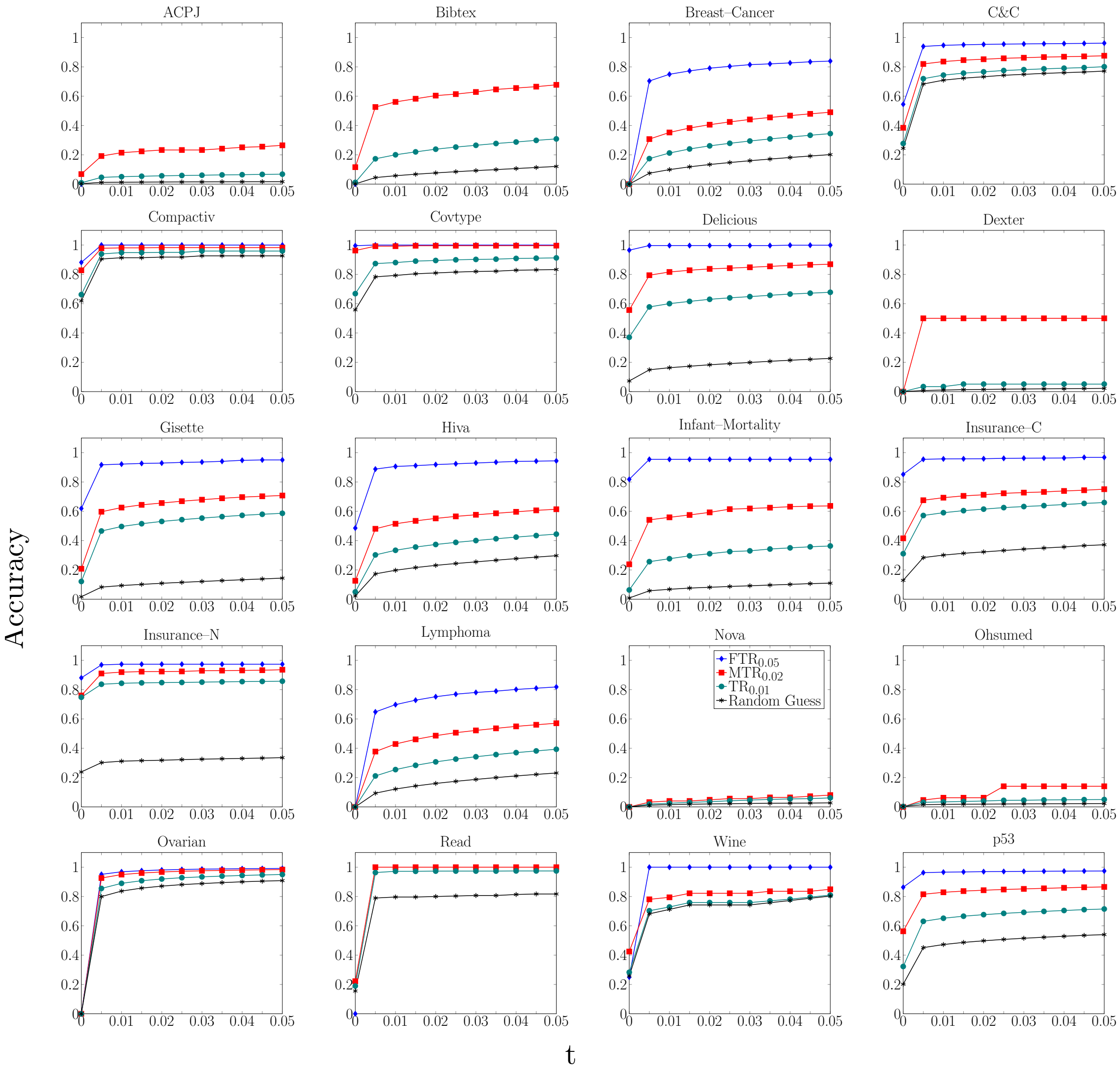


Figure 6.7: **Accuracy per data set and rule, for varying dependence threshold.** Accuracies $Acc_i^R(t)$ as a function of threshold $t$ for all data sets along. Overall, the performance differences are insensitive to the threshold $t$ in the performance definition.

## 6.3 Predicting the Presence of Conditional Dependencies

The FTR and the MTR not only predict the presence of the dependency $Y \not\perp\!\!\!\perp Z \mid \emptyset$ given two data sets on $\mathbf{O}_1 = \{X, Y, W\}$ and $\mathbf{O}_2 = \{X, Z, W\}$; the rules also predict that either $X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$ or $X \circ - \circ Z \circ - \circ Y \circ - \circ W$ is the model that generated both data sets (see Algorithms 8 and 9). Both of these models also imply the following dependencies:

$$Y \not\perp\!\!\!\perp Z \mid X,$$

$$Y \not\perp\!\!\!\perp Z \mid W,$$

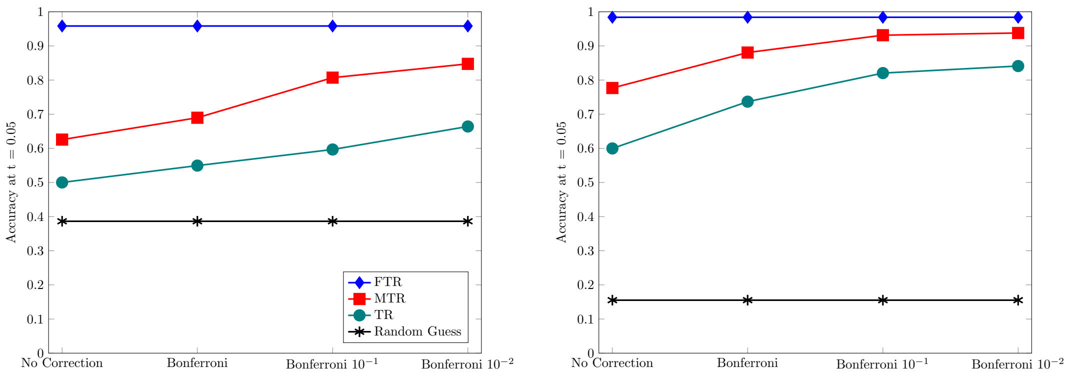


Figure 6.8: **Rules accuracies for varying independence threshold.** Average accuracy $\overline{Acc}(0.05)$ (left) and pooled accuracy $\underline{Acc}(0.05)$ (right) for each rule as a function of $\alpha$ thresholds used: $\alpha_{MTR} \in \{0.05, 0.02, 0.002, 0.0002\}$ and $\alpha_{TR} \in \{0.05, 0.01, 0.001, 0.0001\}$ corresponding to no correction, Bonferroni correction, and stricter than Bonferroni by one and two orders of magnitude respectively. FTR's performance is higher even when MTR and TR become quite conservative.

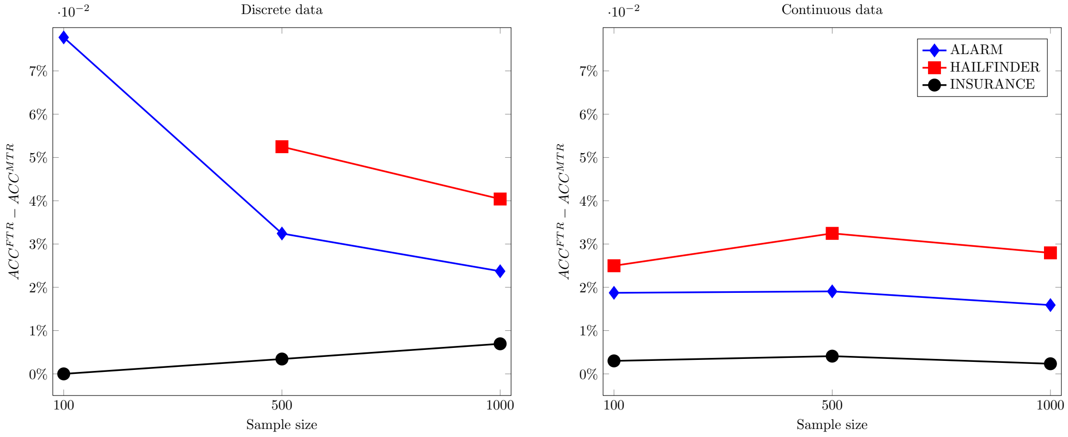


Figure 6.9: **Conservative testing increases accuracy.** Difference between $ACC^{FTR}$ and $ACC^{MTR}$ for discrete (left) and continuous (right) simulated data sets. Results calculated using the "Bonferroni" correction (i.e., $\text{FTR}_{0.05}$ and $\text{MTR}_{0.02}$). The difference between FTR and MTR is larger than 5% only in two cases with low sample size (ALARM and HAILFINDER networks); however, the difference steeply decreases as the sample size increases. No prediction was made for HAILFINDER with discrete data and 100 samples. The difference between FTR and MTR on faithful data is relatively small.

$$Y \not\perp\!\!\!\perp Z \mid \{X, W\}.$$

In other words, the rules predict that the dependency between $Y$ and $Z$ is not mediated by either $X$ or $W$ inclusively. To test whether all these predictions hold simultaneously at threshold $t$ we compute:

$$p^* = \max_{\mathbf{S} \subseteq \{X,W\}} p_{Y \perp\!\!\!\perp Z \mid \mathbf{S}}$$

and test whether $p^* \leq t$. The above dependencies are all the dependencies that are implied by the model but not tested by the FTR given that it has no access to the joint distribution of $Y$ and $Z$. Note that we forgo providing a value for $p^*$ when any of the conditional dependencies can not be calculated, that is, when there are not enough samples to achieve large enough power, see Tsamardinos and Borboudakis (2010). The accuracy of the predictions for all dependencies in the model, named Structural Accuracy because it scores all the dependencies implied by the structure of model, is defined in a similar fashion to *Acc* (Definition 6.2.1) but based on $p^*$ instead of $p$:

$$SAcc_i^R(t) = \#\{p^* <= t, p \in M_i^R\}/|M_i^R|.$$

The averaged and pooled *SAcc* for each FTR, MTR (with "Bonferroni" correction) and randomly selected quadruples is shown in Figure 6.10; *SAcc* for each data set is shown in Appendix 6.11. There is no line for the TR as it concerns triplets of variables and makes no predictions about conditional dependencies. Both FTR and MTR have maximum $p$-values $p^*$ concentrated around zero. The curves do not rise as sharp as those in Figure 6.6 since the $p^*$ values are always larger than the corresponding $p_{Y \perp\!\!\!\perp Z \mid \emptyset}$.

### Summary, Interpretation, and Conclusions

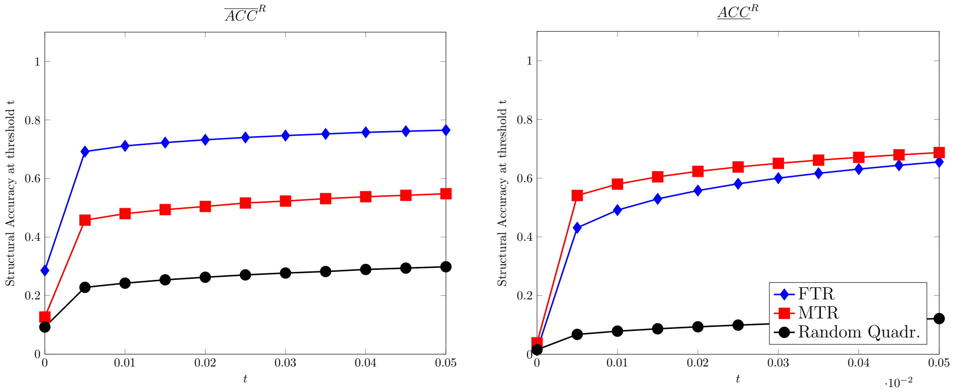


Figure 6.10: **Averaged and pooled structural accuracy per rule, for varying dependence threshold.** Averaged structural accuracy $\overline{SACC}^R(t)$ and pooled structural accuracy $\underline{SACC}^R(t)$. FTR outperforms MTR on most of the data sets, and thus $\overline{SACC}^{FTR}(t) > \overline{SACC}^{MTR}(t)$. However, since MTR ouperforms FTR on few data sets with a large number of predictions and so $\underline{SACC}^{MTR}(t)$ is slightly better than $\underline{SACC}^{FTR}(t)$ for $t <= 0.05$.

The results show that both the FTR and MTR rules correctly predict all the dependencies (conditional and unconditional) implied by the models involving the two variables never measured together. These results provide evidence that these rules often correctly identify the data generating structure.

## 6.4 Predicting the Strength of Dependencies

In this section, we present and evaluate ideas that turn the qualitative predictions of FTR to quantitative predictions. Specifically, for Example 6.1 we show *how to predict the strength of dependence* in addition to its existence. In addition to the Faithfulness Condition, we assume that when the FTR applies on quadruple $\{X, Y, Z, W\}$, all dependencies are linear with independent and normally

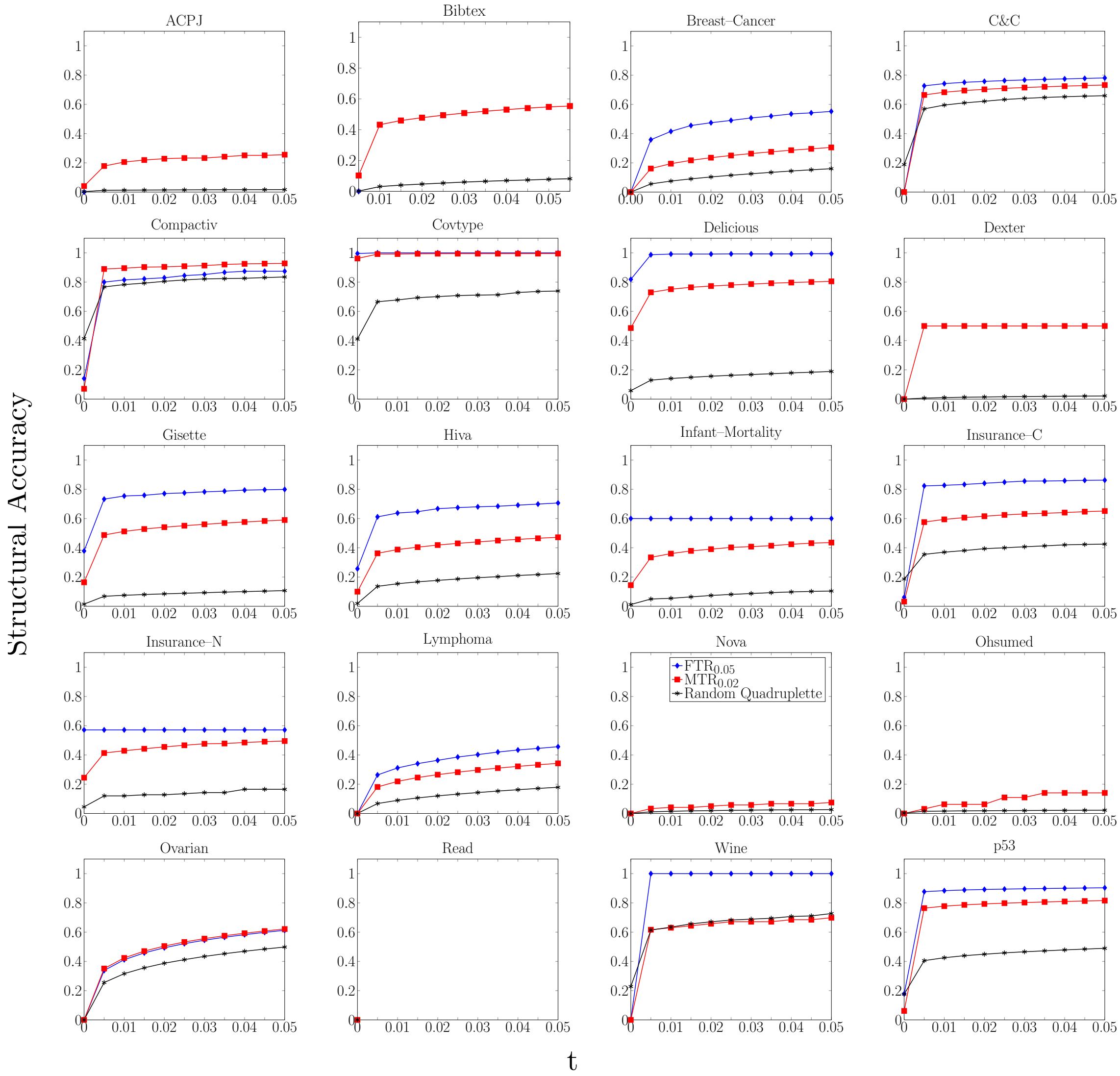


Figure 6.11: **Structural accuracy per data set and rule, for varying dependence threshold.** Structural Accuracies $SAcc_i^R(t)$ as a function of threshold $t$ for all data sets.

distributed error terms. However, the results of these section could possibly be generalized to more relaxed settings, for example, when some of the error terms are non-Gaussian (Shimizu et al., 2006, 2011). When the Full-Testing Rule applies, we can safely assume the true structure is one of the MAGs shown in Figure 6.2. Given linear relationships among the variables, we can treat these MAGs as linear Path Diagrams (Richardson and Spirtes, 2002). We also consider normalized versions of the variables with zero mean and standard deviation of one. Let us consider one of the possible MAGs:

$$M_1 : X \xleftarrow{\rho_{XY}} Y \xrightarrow{\rho_{YZ}} Z \xrightarrow{\rho_{ZW}} W$$

where $\rho_{XY}$ is the *regression coefficient* of regressing $X$ on $Y$, that is,

$$X = \rho_{XY} Y + \epsilon$$

and $\epsilon$ is the error term. Since we have standardized the variables, and since the above equation is simple linear regression, $\rho_{XY}$ coincides with the Pearson linear *correlation* between variables $X$ and $Y$. Thus, there is no need to distinguish the two.[4] Now notice that in all MAGs in Figure 6.2 there are no colliders. Thus, as in $M_1$ above, all regressions are simple regressions and all standardized regression coefficients coincide with their respective correlation coefficients, and so, for the rest of the section we will not differentiate between the two.

The rules of path analysis (Wright, 1934) dictate that the correlation between two variables, for example, $\rho_{XY}$ equals the sum of the contribution of every $d$-connecting path (conditioned on the empty set); the contribution of each path is the product of the correlations on its edges. For $M_1$ the above rule implies (among others):

$$\rho_{XZ} = \rho_{XY} \times \rho_{YZ}$$

because from $X$ to $Z$ there is a single path going through $Y$. Recall that the 14 consistent MAGs are represented by the following PAGs:

$$P_1 : X \circ - \circ Y \circ - \circ Z \circ - \circ W$$

and

$$P_2 : X \circ - \circ Z \circ - \circ Y \circ - \circ W.$$

All MAGs consistent with $P_1$ entail the same constraints on the coefficients using path analysis; similarly all MAGs consistent with $P_2$.[5] Specifically, if $P_1$ is the true structure we get the constraints

$$\rho_{XZ} = \rho_{XY} \times \rho_{YZ}, \tag{6.1}$$

$$\rho_{YW} = \rho_{YZ} \times \rho_{ZW}. \tag{6.2}$$

On the other hand, if $P_2$ is the true structure we obtain:

$$\rho_{XY} = \rho_{XZ} \times \rho_{YZ}, \tag{6.3}$$

$$\rho_{ZW} = \rho_{YZ} \times \rho_{YW}. \tag{6.4}$$

*We use $\rho$, $\hat{r}$, and $r$ to denote actual, predicted, and sample correlations, respectively.* The quantities that we observe are the *sample correlation coefficients*, denoted by $r$, for the pairs of variables measured together. Thus, we can compute the quantities $r_{XY}, r_{XZ}, r_{YW}, r_{ZW}$ from the data and we would like to predict $\rho_{YZ}$ without available data. From Equations 6.1, 6.2, 6.3, 6.4 above we obtain four possible estimators:

$$\text{If } P_1 \text{ is true } : \hat{r}^1_{YZ} \approx \frac{r_{XZ}}{r_{XY}} \text{ from Equation 6.1 and } \hat{r}^2_{YZ} \approx \frac{r_{YW}}{r_{ZW}} \text{ from Equation 6.2,} \tag{6.5}$$

[4] If $Y$ was a collider then it would have been regressed on multiple variables; in this case $\rho_{XY}$ should be the partial regression coefficient which in general does not coincide with the partial correlation coefficient, even for standardized variables.

[5] In general, the consistent MAGs may disagree on the unknown correlations. In this case, these parameters may not identifiable. However, one could analyze all possible MAGs to provide bounds on the unidentifiable quantities in a similar fashion to Balke and Pearl (1997) and Maathuis et al. (2009).

$$\text{if } P_2 \text{ is true } : \hat{r}^3_{YZ} \approx \frac{r_{XY}}{r_{XZ}} \text{ from Equation 6.3 and } \hat{r}^4_{YZ} \approx \frac{r_{ZW}}{r_{YW}} \text{ from Equation 6.4} \tag{6.6}$$

where the superscripts correspond to the equation used to produce the estimate. Notice that, each possible PAG provides two equations to predict $\rho_{YZ}$, that is, the parameter is overidentified. Also, the following important relation holds between the estimators:

$$\hat{r}^1_{YZ} = \frac{1}{\hat{r}^3_{YZ}} \text{ and } \hat{r}^2_{YZ} = \frac{1}{\hat{r}^4_{YZ}}.$$

This observation allows us to distinguish between PAGs $P_1$ and $P_2$: if $\hat{r}^1_{YZ}, \hat{r}^2_{YZ} \in [-1, +1]$, then their reciprocals $\hat{r}^3_{YZ}, \hat{r}^4_{YZ} \notin [-1, +1]$ and so, they are not valid estimates for a correlation. Thus, we can infer that $P_1$ is the true structure and employ only $\hat{r}^1_{YZ}, \hat{r}^2_{YZ}$ for estimation. Otherwise, the reverse holds $\hat{r}^3_{YZ}, \hat{r}^4_{YZ} \in [-1, +1]$, $P_2$ is the true structure and only $\hat{r}^3_{YZ}, \hat{r}^4_{YZ}$ should be used for estimation. Due to sampling errors it is plausible that we obtain conflicting information: $\hat{r}^1_{YZ} \in [-1, +1]$ but $\hat{r}^2_{YZ} \notin [-1, +1]$ (and so $\hat{r}^3_{YZ} \notin [-1, +1]$ and $\hat{r}^2_{YZ} \in [-1, +1]$). In that case, we forgo making any predictions.

The ramifications of the above analysis are important. In the case where all variables are jointly measured, the distribution is faithful, the relations are linear and the error terms follow Gaussian distributions, the set of statistically indistinguishable causal graphs is determined completely by the independence model and not by the parameterization of the distribution. However, in the case of incomplete data, where some variable sets are not jointly observed, the set of indistinguishable models also depends on the parameters of the distribution, even for linear relations and Gaussian error terms. In our scenario, by analyzing the estimable parameters we can further narrow down the set of equivalent consistent MAGs.

At this point in our analysis, we are left with two valid estimators, either $\hat{r}^1, \hat{r}^2$ or $\hat{r}^3, \hat{r}^4$. All estimators are computed as ratios. We report the mean of the two valid estimators as the predicted $\hat{r}_{YZ}$ for a more robust estimation. The above procedure is formalized in Algorithm 11, named FTR-S.

### Empirical Evaluation of the Predictions of Correlation Strength

As in Section 6.2, we partition each data set with continuous variables to three data sets $\mathcal{D}_1$, $\mathcal{D}_2$, and a test set $\mathcal{D}_t$. We then apply Algorithm 11 and predict the strength of correlation $\hat{r}_{YZ}$ for various pairs of variables; we compare the predictions with the sample correlation $r_{YZ}$ as estimated in $\mathcal{D}_t$. The results for one representative data set (Lymphoma) are shown in Figure 6.12(a): there is an apparent trend to overestimate the absolute value of the sample correlation.

There are several possible explanations for the bias of the method, including violations of normality, linearity, faithfulness, and even the known bias in the estimation of sample correlation coefficients (Zimmerman et al., 2003) that are used for making the predictions in Algorithm 11. In order to pinpoint the culprit, we generated data where all assumptions hold from the model $M_1$ shown in the beginning of this section, where we set the correlations $\rho_{XY}, \rho_{YZ}, \rho_{ZW}$ and the noise terms are independently and normally distributed. We used the entire spectrum of positive correlation coefficients for all three correlations to examine how the bias varies as a function of these correlations. We generated 1000 data sets of different sample sizes of 50, 70 and 100 samples. We then used Equation 6.1 to estimate $r_{YZ}$ in each experiment. *This set of experiments revealed no significant bias for any of the experimental settings* (results are not shown for brevity).

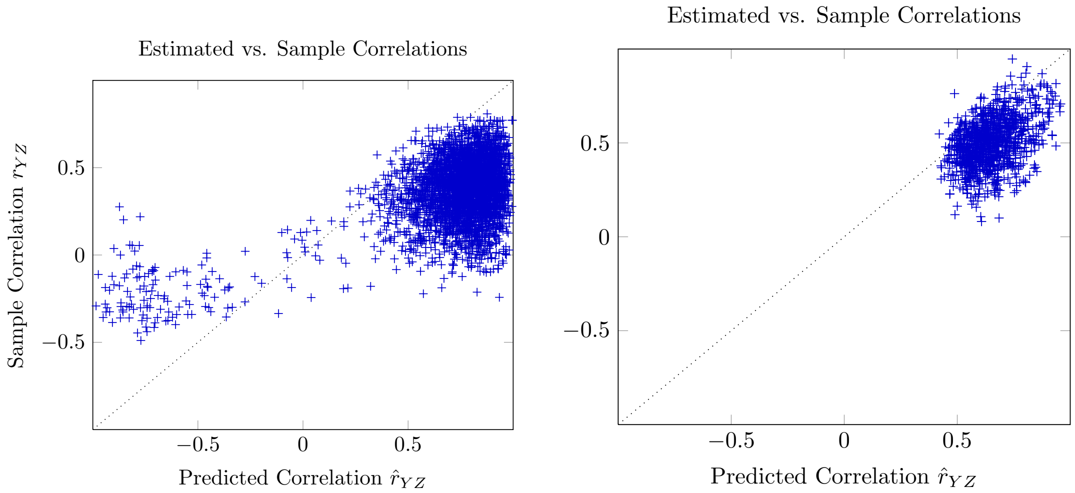
```
Algorithm 11: Predict Dependency Strength(FTR-S)
   Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X, Y, W\}$ and $\{X, Z, W\}$, respectively
 1 if Full-Testing Rule($\mathcal{D}_1$, $\mathcal{D}_2$) does not apply then return;
 2 ;
 3 In $\mathcal{D}_1$ compute $r_{XY}, r_{YW}$;
 4 In $\mathcal{D}_2$ compute $r_{XZ}, r_{ZW}$;
 5 $\hat{r}^1 \leftarrow \frac{r_{XZ}}{r_{XY}}$;
 6 $\hat{r}^2 \leftarrow \frac{r_{YW}}{r_{ZW}}$;
 7 $\hat{r}^3 \leftarrow \frac{r_{XY}}{r_{XZ}}$;
 8 $\hat{r}^4 \leftarrow \frac{r_{ZW}}{r_{YW}}$;
 9 if $\hat{r}^1, \hat{r}^2 \in [-1, 1]$ then
10     Predict $X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$;
11     Predict correlation $\hat{r}_{YZ} = \frac{1}{2}(\hat{r}^1 + \hat{r}^2)$;
12 end
13 else if $\hat{r}^3, \hat{r}^4 \in [-1, 1]$ then
14     Predict $X \circ\!-\!\circ Z \circ\!-\!\circ Y \circ\!-\!\circ W$;
15     Predict correlation $\hat{r}_{YZ} = \frac{1}{2}(\hat{r}^3 + \hat{r}^4)$;
16 end
17 else
18     Make no prediction
19 end
```

We next tested whether the bias is an artifact of the filtering by the FTR at Line 1 of the FTR-S algorithm. We re-run this procedure, but this time we kept only the predicted correlations that passed the FTR. By comparing Figure 6.12(a) produced on real data, and 6.12(b) on simulated data, we observe a similar behavior, indicating that FTR filtering seems a reasonable explanation for the bias.

An explanation of this phenomenon now follows. Suppose $M_1 : X \xleftrightarrow{\rho_{XY}} Y \xrightarrow{\rho_{YZ}} Z \xrightarrow{\rho_{ZW}} W$ is the data generating MAG. We expect that $\hat{r}_{YZ} = \frac{r_{XZ}}{r_{XY}}$ (the equality $\hat{r}_{YZ} = \frac{r_{YW}}{r_{ZW}}$ also holds but we ignore it to simplify the discussion). When sample correlations among $\{X, Y, Z, W\}$ pass the FTR, this means that both $r_{XZ}$ and $r_{XY}$ are above a cut-off threshold, as given by the Fisher test. For example, for a data set with 70 samples, two variables are considered dependent ($\rho \neq 0$) if their sample correlation is more that 0.2391 (in absolute value), whereas for a data set with 50 samples, this threshold is 0.2852.

Filtering with the Fisher test introduces a bias in the estimation of $r$. The bias of the estimation without filtering, $r_u$ is $B_{r_u} = E[r_u - \rho] = \overline{r_u} - \rho$, while the bias of the estimation with filtering $r_f$ is $B_{r_f} = E[r_f - \rho] = \overline{r_f} - \rho$, where $|r_f| \geq t$. The threshold $t$, as mentioned above, is the threshold determined by the Fisher test and depends on sample size. *The lower the sample size, the higher the threshold $t$, and so the higher the introduced bias $B_{r_f}$. In addition, the lower the $|\rho|$ the higher the bias $B_{r_f}$.*

Figure 6.13 illustrates these points pictorially. In this example, the distribution of the sample correlation $r$ of two variables for sample size 70 when the true correlation is $\rho \in \{0.2, 0.4\}$. For



(a) Lymphoma Data Set

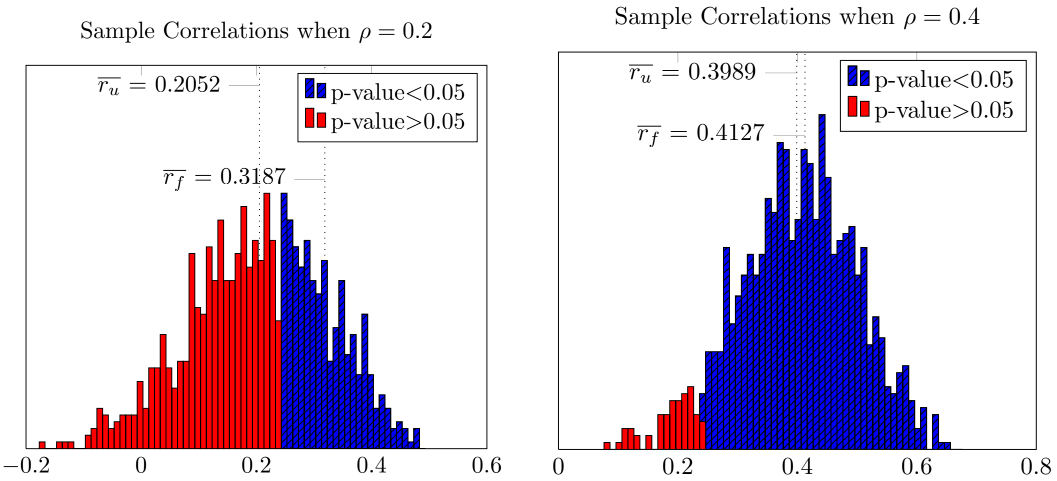
(b) Simulated Data where FTR Rule Applies

Figure 6.12: **Bias in estimation of the correlation coefficient.** (a) Predicted ($\hat{r}_{YZ}$) vs sample ($r_{YZ}$) correlation for the Lymphoma data set. There is an obvious trend to over-estimate the correlation in absolute value. (b) Simulated results from model $\mathcal{M}_1$ when $\rho_{XZ}$ and $\rho_{YW}$ are lower than 0.4 and observed correlations *are found significant* (FTR applies). The FTR constraint that the observed correlations are significant reproduces a similar behavior in the simulated data, explaining the bias.

unfiltered estimations, the bias is $B_{r_u}$ is 0.0052 and -0.0011 for $\rho$ equal to 0.2 and 0.4 respectively, whereas for filtered estimations the corresponding values $B_{r_f}$ are 0.1187 and 0.0127.

Going back to the prediction $\hat{r}_{YZ} = \frac{r_{XZ}}{r_{XY}}$ notice that the numerator is always lower (in absolute value) than the denominator. Therefore, when filtered, it is, on average more overestimated than the denominator. This implies that, on average, the fraction leads to overestimating the absolute value of $\rho_{YZ}$. The lower the values of $|r_{XZ}|$ and $|r_{XY}|$, the larger we expect this bias to be. The situation is similar for all fractions involved in Equations 6.5 and 6.6. This hypothesis is confirmed in the data as illustrated in Figure 6.14 where the predictions are grouped by the mean absolute values of the denominators used in their computation.

The bias should be a function of sample size, the absolute value of the correlations employed for its computation, and the significance thresholds of the FTR rule. However, a full theoretical treatment of the bias is out of the scope of the work. In the experiments that follow we remove the linear trend to over-estimate (*calibrate*) by regressing the sample correlations $r_{YZ}$ on the predicted $\hat{r}_{YZ}$: the final calibrated prediction is $s \times \hat{r}_{YZ} + i$. For each data set the intercept $i$ and slope $s$ of the regression are estimated by training on the remaining data sets (leave-one-data-set-out validation). The effect of this calibration is shown in Figure 6.15. To avoid repetition, the detailed set of results is presented in the comparative evaluation to statistical matching in Section 6.6.



(a)

(b)

Figure 6.13: **Overestimation of the sample correlation coefficient $r$ is larger for lower $\rho$.** Histograms of the sample correlations for (a) $\rho = 0.2$ and (b) $\rho = 0.4$ for sample size 70. Red bars correspond to cases where the Fisher test returns a p-value $> 0.05$, whereas blue bars correspond to p-values $< 0.05$. The dashed lines indicate the mean sample correlation for filtered and unfiltered correlations. The lower the $\rho$ , the more overestimated the sample correlations that pass the Fisher test, therefore the difference between the two means is larger.

## Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:

- FTR coupled with parametric assumptions can be used to predict the strength of dependency (correlation), providing quantitative predictions. This is equivalent to constructing a prediction model for variables not jointly observed.

- In the case of incomplete data, where some variable sets are not jointly observed, the set of indistinguishable models also depends on the parameters of the distribution, even for linear relations and Gaussian error terms. In contrast, in the case where all variables are jointly measured and the distribution is faithful the set of statistically indistinguishable causal graphs is completely determined by the independence model (again, also assuming linearity and Gaussian error terms).

- In this simple scenario, *given the correct structure*, path analysis of the induced MAGs provides easy solutions for predicting the strength of dependence. However, *searching for the correct MAG models* by applying the FTR incurs bias on the predictions that should be taken into account.

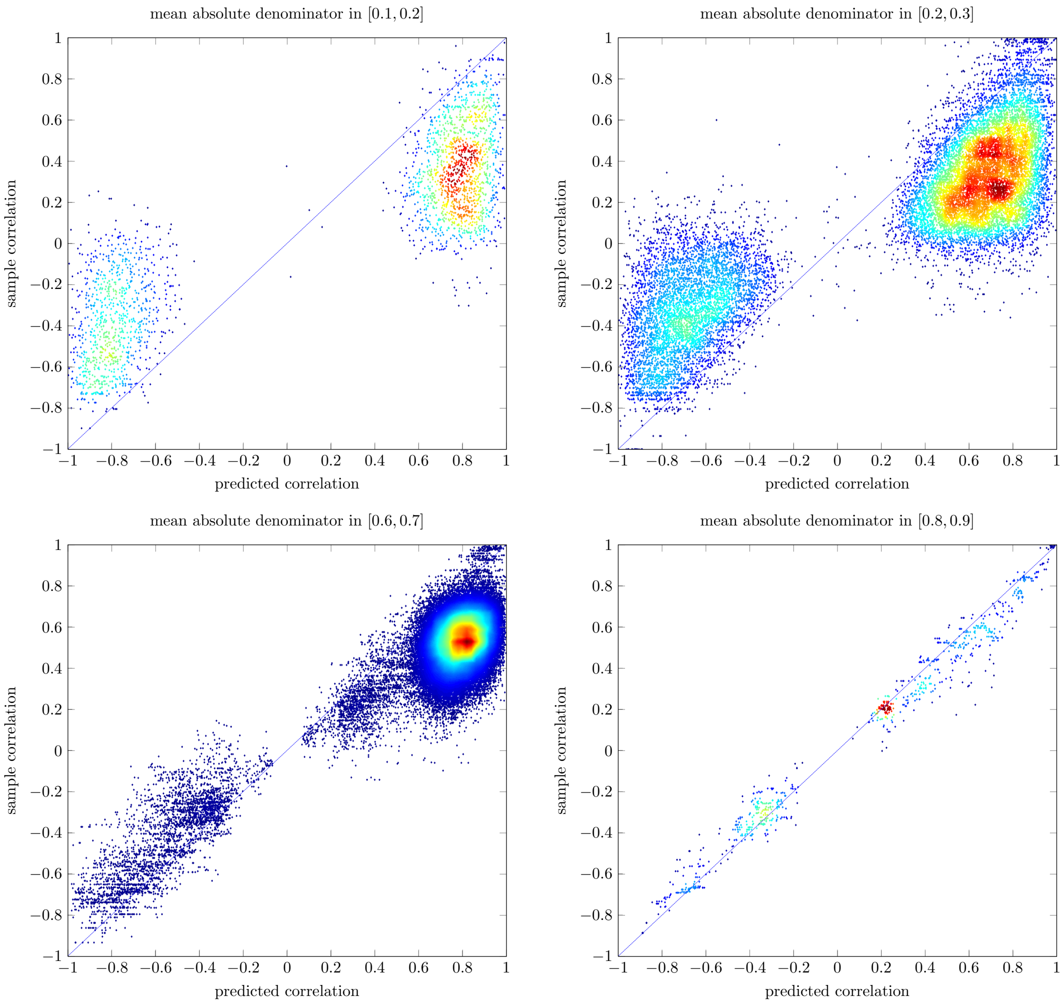


Figure 6.14: **Bias is reduced for larger values of the denominator.**Predicted vs sample correlations over all data sets, grouped by the mean absolute values of the denominators used in their computation: predictions computed based on large correlations have reduced bias. Red regions correspond to higher density areas.

## 6.5 Related Work

Whole sub-fields have been developed to address the problem of integrative analysis, that we review briefly. Meta-Analysis focuses on the co-analysis of studies with similar sampling and experimental design characteristics with the purpose of making inferences about a single association. Meta-Analysis in Statistics (O'Rourke, 2007) combines the results of several studies to address a set of related research hypotheses. While meta-analysis focuses on a pair-wise association of a variable with an outcome of interest, a recent interesting extension addresses the problem of estimating the multivariate associations (for example, in the form of a regression model) with the target variable (Samsa et al., 2005); such methods often appear under the names of meta-regression and univariate

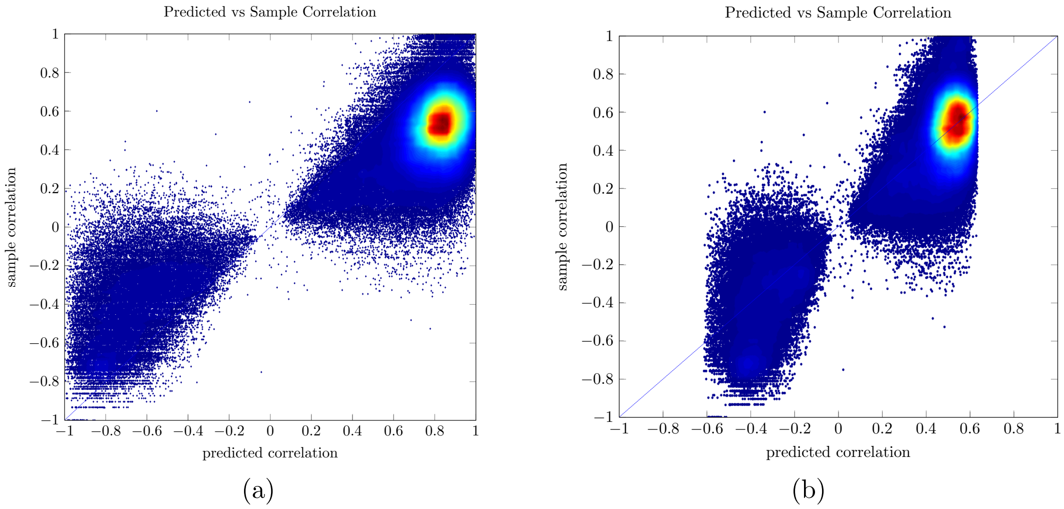


Figure 6.15: **Calibration of the predictions.**Predicted vs sample correlations on all data sets (a) before, and (b) after calibration.

synthesis (Zhou et al., 2009). The main idea of the latter is to assume a parametric form of the regression model and estimate the sufficient statistics from several homogeneous (in terms of being conducted on the same population, experimental conditions, sampling, etc.) studies that may not measure all variables (risk factors in this context). Both statistical matching and meta-analysis's scope does not extend to other sources of heterogeneity of the data sets, such as different experimental conditions.

In Computer Science and Machine Learning, the field of Transfer Learning (Pan and Yang, 2010) represents a main effort in integrative analysis. In Transfer Learning, successful search control strategies, model priors, and other characteristics transfer among different domains and/or tasks. When the task (target) is the same but the domains (populations) are different, this type of Transfer Learning is called Domain Adaptation. In this case, typically one would like to translate the estimated conditional distribution $P_s(Y|X)$ used for prediction in a source distribution to a target distribution $P_t(Y|X)$ that may be different (e.g., has a different marginal class distribution). Given that such methods are typically non-causal based, they cannot transfer to data sets where manipulations have been performed (causal methods could transfer predictive models to manipulated distributions as we show in Tsamardinos and Brown 2008, also shown in Maathuis et al. 2010). In addition, the input space for the predictors $X$ has to be common. When the domain is the same (same distribution), but the tasks (target variables) are different, the type of Transfer Learning is called *Multi-Task Learning*. This type of learning attempts to simultaneously build models for several tasks in an effort to use one for leveraging the performance on the others. Typically this is performed by using a shared representation and learning common induced features. Again, these inferences are limited as they can only combine studies under the same sampling and experimental conditions on the same sets of variables.

Other fields may seem related in a first glance, but are orthogonal to the proposed research. The field of Relational Learning (Getoor and Taskar, 2007) does not really address the problem of learning from different data sets/studies over different samples, rather than a single data set (the one stemming from implicitly propositionalizing the database) in the form of relational tables. Similarly, the field of Distributed Learning (Cannataro et al., 2002) is restricted to designing time and communication-efficient analysis of what is essentially a single data set stored in different locations.

Other related work includes efforts to combine models (that may be developed from different data sets) on the same system but on different scales (Gennari et al., 2008). Typically, such methods involve mechanical models using differential equations and are not concerned with statistical models. In addition, these methods concern vertical integration at different temporal or spatial scales, while INCA proposes a horizontal integration of studies.

To the best of our knowledge, the only approach that can predict the correlation of variables not jointly measured is Statistical Matching (D'Orazio et al., 2006), an integrative analysis procedure for data sets defined over overlapping variable sets. Statistical matching addresses two main tasks named the *micro approach* and *the macro approach*. The micro approach aims to impute the missing values and construct a complete synthetic file, whereas the macro approach aims to identify some characteristics of the joint probability distribution of the variables not jointly observed. Naturally, construction of the synthetic data set premises the estimation of the parameters of the joint distribution. We focus on the macro approach as it presents an alternative to the FTR and MTR.

The problem set up is as follows: variables $\mathbf{Y} \cup \mathbf{X}$ are measured in data set $\mathcal{D}_1$, while variables $\mathbf{Z} \cup \mathbf{X}$ are measured in data set $\mathcal{D}_2$. Thus $\mathbf{X}$ are the commonly measured variables. The goal is to estimate the variances and covariances of $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. The problem cannot be solved without additional assumptions (Rubin, 1974; D'Orazio et al., 2006). Depending on nature of the assumptions, statistical matching is able to produce either intervals or point-estimates for the covariances between $\mathbf{Y}$ and $\mathbf{Z}$. The most common assumptions are discussed and compared with FTR in Section 6.6.

## 6.6 Comparison Against Statistical Matching

The most typical assumption in the literature able to produce point estimates is the Conditional Independence Assumption: $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$. This is an arbitrary assumption that has been long debated. Alternatively, one can limit the shape of the distribution by imposing parametric forms, such as multivariate normality. The latter type of assumptions, for the typical distributions, do not lead to identifiable estimations, but instead provide bounds on the missing covariances. Other approaches do exist that require prior knowledge, for example, Vantaggi (2008) assumes knowledge of structural zeros and Cudeck (2000) of the structure of latent factors; such approaches however, are not directly comparable with FTR and MTR on this task. In this section we briefly present the main theory and techniques used in statistical matching, and then attempt to empirically compare against FTR.

### Statistical Matching Based on the Conditional Independence Assumption

The most common assumption that allows identification of the unknown parameters is the **conditional independence assumption** (CIA): $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$. The conditional independence assumption is usually paired with some parametric assumption. The most common assumption for the shape of a continuous distribution of the variables involved in the model is multivariate normality. In this case, the parameters of the JPD are the mean vector and the covariance matrix. The covariance Matrix for $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ can be written as:

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XZ}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YZ}} \\ \Sigma_{\mathbf{ZX}} & \Sigma_{\mathbf{ZY}} & \Sigma_{\mathbf{ZZ}} \end{bmatrix}$$

where the unknown parameter is $\Sigma_{\mathbf{YZ}}$. The CIA assumption imposes that the covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$ is null, thus,

$$\Sigma_{\mathbf{YZ}} = \Sigma_{\mathbf{YX}}{\Sigma_{\mathbf{XX}}}^{-1}\Sigma_{\mathbf{XZ}}.$$

In case we have standardized variables, and $\mathbf{Y} = \{Y\}$ and $\mathbf{Z} = \{Z\}$, the covariance matrix becomes

$$\Sigma = \begin{bmatrix} \rho_{\mathbf{XX}} & \rho_{\mathbf{X}Y} & \rho_{\mathbf{X}Z} \\ \rho_{Y\mathbf{X}} & 1 & \rho_{YZ} \\ \rho_{Z\mathbf{X}} & \rho_{ZY} & 1 \end{bmatrix}$$

and so

$$\rho_{YZ} = \rho_{Y\mathbf{X}}{\rho_{\mathbf{XX}}}^{-1}\rho_{\mathbf{X}Z}.$$

This formula can be used to produce a prediction $\hat{r}_{YZ}$ for the correlation coefficient of the not commonly observed variables $Y$ and $Z$. Recall that, we assume we are given a data set $\mathcal{D}_1$ on variables $\mathbf{X} \cup Y$ and a data set $\mathcal{D}_2$ on $\mathbf{X} \cup Z$. The parameters $\rho_{\mathbf{X}Y}$ and $\rho_{\mathbf{X}Z}$ can be estimated from $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, while the parameters $\rho_{\mathbf{XX}}$ can be estimated from either or both data sets.

In an applied setting, there is usually also a preprocessing step attempting to identify a subset of the common variables to be used in the matching process. This step serves mainly computational efficiency and interpretability purposes and does not affect the asymptotic properties of the procedure. The main method suggested in D'Orazio et al. (2006) is to disregard all variables in $\mathbf{X}$ that are *independent* with both $Y$ and $Z$. The details are described in Algorithm 12.

**Algorithm 12:** Predict Correlation: Statistical Matching Rule (**SMR**)

```
Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{\mathbf{V} \cup Y\}$ and $\{\mathbf{V} \cup Z\}$, respectively
$\psi_1 \leftarrow \{V \in \mathbf{V} : V \perp\!\!\!\perp Y \mid \emptyset\}$ in $\mathcal{D}_1$
$\psi_2 \leftarrow \{V \in \mathbf{V} : V \perp\!\!\!\perp Z \mid \emptyset\}$ in $\mathcal{D}_2$
$\mathbf{X} \leftarrow \mathbf{V} \setminus (\psi_1 \cap \psi_2)$
Predict $\hat{r}_{YZ} = \hat{\Sigma}_{Y\mathbf{X}}{\hat{\Sigma}_{\mathbf{XX}}}^{-1}\hat{\Sigma}_{\mathbf{X}Z}$
```

Even though the conditional independence assumption seems quite arbitrary, it is intuitively justified in certain cases. When the number of common variables is large it is unlikely that $Y$ provides *additional* information for $Z$, than what $\mathbf{X}$ already provides. In other words, we expect $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$ to hold or hold approximately. Using graphical model theory one can better formalize this intuition:

**Theorem 6.6.1** *Consider a Bayesian network of maximum degree $k$ faithful to a distribution defined over a set of variables $\mathbf{V} = \mathbf{X} \cup Y \cup Z$, $|\mathbf{V}| = N$. Then, the CIA $Y \perp\!\!\!\perp Z \mid \mathbf{X}$ holds if and only if $Y \notin Mb(Z)$, where $Mb(Z)$ is the Markov Boundary of $Z$ in the context of variables $\mathbf{V}$; if $Y$ and $Z$ are chosen at random the probability of the CIA being violated is upper bounded by $k^2/N$.*

**Proof** In a faithful distribution over $\mathbf{V}$, each variable $Y$ has a unique Markov Boundary $Mb(Y)$ (Pearl, 2000) that coincides with the parents, children, and parents of children (spouses) of $Y$ in any network faithful to the distribution. It is also easy to see that $Y \in Mb(Z) \Leftrightarrow Z \in Mb(Y)$. Finally, the $Mb(Y)$ and any of its supersets $d$-separates $Y$ from any other node $Z$. Thus, when $Z \notin Mb(Y)$, then conditioned on the remaining variables (superset of $Mb(Y)$) $Y$ becomes $d$-separated and independent of $Z$. Thus, the CIA holds. Conversely, if $Z \in Mb(Y)$ then it is either a neighbor of $Y$ or a spouse. If it is a neighbor it cannot be made independent of $Y$ conditioned on any subset of the variables (Spirtes et al., 2000). If it is a spouse of $Y$, then conditioned on the remaining variables (which includes the common children) it is $d$-connected to $Y$ and thus dependent. Thus, the CIA does not hold.

Now, the Markov Boundary of $Y$ is a subset of the nodes that are reachable from $Y$ within two edges. If the network has degree at most $k$ the probability that a randomly chosen $Y$ belongs to the Markov Boundary of $Z$ is less than $k^2/N$.

Thus, when the sparsity remains the same, the probability of a violation of the CIA between two randomly selected variables decreases with the number of participating variables $N$. The theoretical results is illustrated in Figure 6.16 on simulated data. The figure shows the results of the statistical matching procedure described in Algorithm 12 for simulated continuous data from a network based on the ALARM network (Beinlich et al., 1989).[6] To recreate the scenario above we generated two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ of 1000 samples each from the distribution of the network. We then applied the statistical matching rule described in Algorithm 12 for each pair of variables, considering that the rest of the variables in the network are jointly measured in both data sets. Finally, we generated a third data set to test the predictions of the method. The pairs of variables are partitioned in two categories: pairs of variables that belong to each other's Markov Boundary, and pairs of variables that do not belong to each other's Markov Boundary. As expected, the results are poorer for the pairs of variables that belong to each other's Markov Boundary, with a mean absolute error of $0.1649 \pm 0.1088$, compared to a mean absolute error of $0.0326 \pm 0.0271$ for pairs that do not belong to each other's Markov Boundary.

In the context of Maximal Ancestral Graphs, defining the Markov Boundary is more complicated and its cardinality cannot be likewise bounded (Pellet and Elisseeff, 2008). Nevertheless, we still expect that, in a sparse network containing a large number of jointly measured variables, the probability that $Y \in Mb(Z)$ is low. We therefore expect that, when the number of common variables is large, the CIA will often hold for randomly-chosen pairs of variables that have not been observed together. If, however, the set of variables measured in common is small, we have no good reason to expect that the conditional independence assumption holds.

### Empirical Evaluation of SMR and FTR-S

In this section, we empirically compare the SMR and FTR-S methods for predicting the correlation $\hat{r}_{YZ}$ between two variables $Y$ and $Z$ never jointly observed. Both SMR and FTR-S procedures provide such predictions, however, they follow different approaches that makes their comparison not straightforward:

- SMR provides a prediction for all cases. FTR-S provides a prediction given it identifies a specific structure that entails a significant correlation.

[6] The ALARM network a well-known network with 37 variables. We used the skeleton of ALARM to simulate a conditional linear gaussian network with random parameters.

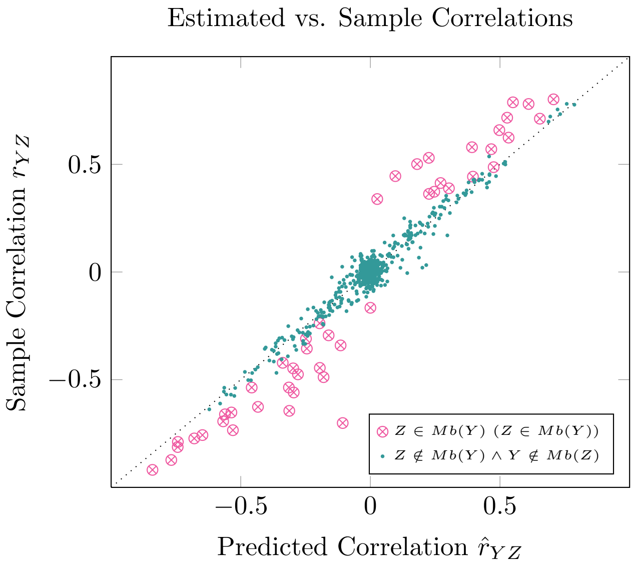


Figure 6.16: **Statistical matching predictions with and without valid assumptions.** Predicted vs actual sample correlations using the Statistical Matching Rule for simulated data from the ALARM network. For each pair of variables, prediction is based upon the subset of the remaining 35 variables that are determined significantly correlated with either $Y$ or $Z$ at level 0.05 . The CIA holds when $Y \notin Mb(Z)$ in which case the mean absolute error is $0.0326 \pm 0.0271$; in contrast, when $Y \in Mb(Z)$ the CIA does not hold and the mean absolute error is $0.1649 \pm 0.1088$.

| Data Sets | SMR$_G$ | SMR$_Q$ | FTR-S |
|---|---|---|---|
| ACPJ | 445121 | 509000 | 0 |
| Breast-Cancer | 436093 | 356000 | 1005 |
| C&C | 5050 | 1000 | 70367 |
| Compactiv | 231 | 1000 | 108 |
| Insurance-C | 3486 | 1000 | 1372 |
| Lymphoma | 180074 | 147000 | 3897 |
| Ohsumed | 124505 | 122000 | 0 |
| Ovarian | 52675 | 43000 | 273456 |
| Wine | 66 | 495 | 4 |
| p53 | 132299 | 108000 | 33934 |

Table 6.4: Number of predictions

- SMR can be applied to sets $X$ with more than two commonly measured variables and get leverage from all available information. FTR-S on the other hand is applicable only when the number of common variables is two.

We applied the SMR method on all continuous data sets, simulating two scenarios. In the first scenario, SMR is applied on two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ defined over a quadruple of variables $\{X, Y, Z, W\}$,

| Data Sets | SMR$_G$ | SMR$_Q$ | FTR-S |
|---|---|---|---|
| ACPJ | 0.01 ± 0.01 | 0.02 ± 0.01 | - |
| Breast-Cancer | 0.11 ± 0.08 | 0.13 ± 0.10 | 0.18 ± 0.13 |
| C&C | 0.05 ± 0.03 | 0.19 ± 0.18 | 0.18 ± 0.13 |
| Compactiv | 0.04 ± 0.06 | 0.19 ± 0.20 | 0.14 ± 0.12 |
| Insurance-C | 0.03 ± 0.08 | 0.09 ± 0.14 | 0.14 ± 0.12 |
| Lymphoma | 0.12 ± 0.09 | 0.14 ± 0.11 | 0.17 ± 0.14 |
| Ohsumed | 0.01 ± 0.02 | 0.02 ± 0.02 | - |
| Ovarian | 0.15 ± 0.10 | 0.16 ± 0.11 | 0.09 ± 0.07 |
| Wine | 0.09 ± 0.10 | 0.15 ± 0.17 | 0.22 ± 0.14 |
| p53 | 0.03 ± 0.05 | 0.07 ± 0.10 | 0.14 ± 0.12 |
| Over data sets | 0.06 ± 0.06 | 0.12 ± 0.11 | 0.16 ± 0.12 |
| Over predictions | 0.07 ± 0.08 | 0.07 ± 0.09 | 0.11 ± 0.10 |

Table 6.5: **Mean Absolute Error (MAE) between the calibrated predictions $\hat{r}_{YZ}$ and sample-estimated $r_{YZ}$.** (average value ± standard deviation). SMR$_G$ refers to the Statistical Matching Rule applied on all pairs of variables in the same group, considering the remaining 48 variables in the group as common variables. SMR$_Q$ is the Statistical Matching Rule applied on quadruples of variables randomly chosen from the same group. Finally, FTR-S consists in the Full Testing Rule modified for estimating the strength of the dependency, see Algorithm 11.

where only $X, W$ are jointly measured in both. The pairs of $\mathcal{D}_1$, $\mathcal{D}_2$ are simulated by considering randomly chosen variable quadruples from each variable group of each data set of Table 6.1; as in all experiments, $\mathcal{D}_1$ and $\mathcal{D}_2$ contain a disjoint third of the original samples. This scenario simulates a case where SMR is applied on low dimensional data; we denote it as *SMR$_Q$*. In this case, *SMR has the same information available for making predictions as FTR-S*. Since the number of possible quadruples is computationally prohibitive, we apply *SMR$_Q$* on 1000 randomly chosen quadruples from each variable group of each data set.[7]

In the second scenario, SMR is applied to data sets of higher-dimensionality. Specifically, we apply SMR to all pairs of variables in the same group (see Section 6.2), considering the remaining 48 variables in the group as the common variables **X**. We name this case *SMR$_G$*. The same leave-one-data-set-out calibration method was used for both SMR cases and FTR-S. Figures 6.17, 6.18, 6.19 and 6.20 plot the predicted vs. the sample estimates of the correlations for *SMR$_G$*, *SMR$_Q$* and FTR-S for all the continuous data sets used in the study. The figures also present the coefficient of determination $R^2$, the percentage of variance explained by the predictions. $R^2$ is also interpreted as the reduction in uncertainty obtained by using a linear function of $\hat{r}$ to predict $r$ vs. predicting $r$ by its expected value $E(r)$. Table 6.7 shows the correlation between predicted and sample estimates for all methods and data sets. Notice that $R^2$ is simply computed as the square of the correlation. Other metrics of performance (Mean Absolute Error and Mean Relative Absolute Error) are also presented in, Tables 6.5 and 6.6, respectively

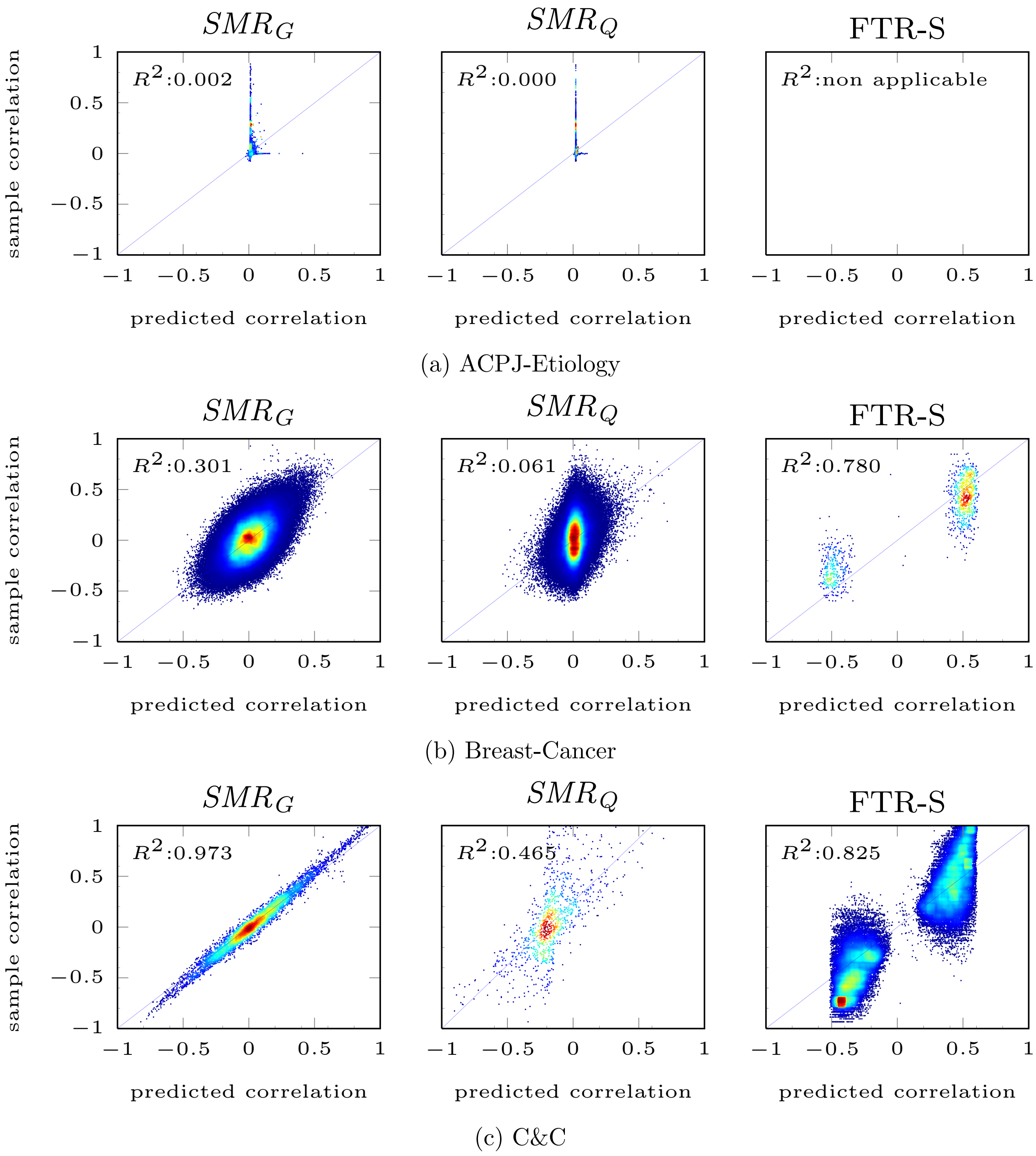
[7] Notice that FTR is typically executed much more efficiently than SMR$_Q$, because of the possible pruning of the search space, for example, if $X$ and $Y$ are independent, there is no need to test whether the rule applies on any quadruples of the form $\langle X, Y, Z, W\rangle$. For the SMR$_Q$ rule instead, one needs to exhaustively consider all quadruples.

| Data Sets | SMR$_G$ | SMR$_Q$ | FTR-S |
|---|---|---|---|
| ACPJ | 13.17 ± 87.17 | 27.22 ± 141.50 | - |
| Breast-Cancer | 5.74 ± 624.51 | 2.79 ± 90.41 | 1.39 ± 4.51 |
| C&C | 1.52 ± 39.16 | 3.53 ± 44.98 | 1.30 ± 16.80 |
| Compactiv | 0.39 ± 1.43 | 1.79 ± 9.39 | 0.46 ± 0.53 |
| Insurance-C | 2.79 ± 11.04 | 2.10 ± 5.15 | 2.44 ± 18.04 |
| Lymphoma | 4.51 ± 182.18 | 3.66 ± 181.90 | 5.77 ± 145.88 |
| Ohsumed | 4.62 ± 30.53 | 7.72 ± 8.95 | - |
| Ovarian | $7.32 \times 10^9 \pm 1.68 \times 10^{13}$ | 0.58 ± 5.51 | 0.20 ± 0.44 |
| Wine | 1.31 ± 2.24 | 1.78 ± 5.65 | 0.38 ± 0.06 |
| p53 | 34.95 ± 7982.92 | 19.86 ± 4544.32 | 4.76 ± 290.58 |
| Over data sets | $7.32 \times 10^9 \pm 1.68 \times 10^{13}$ | 7.10 ± 503.78 | 2.09 ± 59.61 |
| Over predictions | $2.79 \times 10^9 \pm 3.28 \times 10^{12}$ | 14.36 ± 1320.98 | 0.87 ± 87.92 |

Table 6.6: **Mean Relative Absolute Error (MRAE) between the calibrated predictions $\hat{r}_{YZ}$ and sample-estimated $r_{YZ}$.** (average value ± standard deviation) SMR$_G$ refers to the Statistical Matching Rule applied on all pairs of variables in the same group, considering the remaining 48 variables in the group as common variables. SMR$_Q$ is the Statistical Matching Rule applied on quadruples of variables randomly chosen from the same group. Finally, FTR-S consists in the Full Testing Rule modified for estimating the strength of the dependency, see Algorithm 11. For the Ovarian data set the SMR$_G$ rule provides predictions for cases with nearby-zero sample estimated $r_{YZ}$, and these predictions generate extremely high MRAE values. Once excluded such cases, the SMR$_G$ MRAE on the Ovarian data set is 0.54 ± 12.16, while the MRAE averaged over all data sets and over all predictions is 6.95 ± 897.33 and 10.45 ± 2498.28, respectively.

| Data Sets | SMR$_G$ | SMR$_Q$ | FTR-S |
|---|---|---|---|
| ACPJ | 0.05 [0.04; 0.05] | 0.00 [0.00; 0.01] | - |
| Breast-Cancer | 0.55 [0.55; 0.55] | 0.25 [0.24; 0.25] | 0.88 [0.87; 0.90] |
| C&C | 0.99 [0.99; 0.99] | 0.68 [0.65; 0.71] | 0.91 [0.91; 0.91] |
| Compactiv | 0.97 [0.96; 0.98] | 0.49 [0.44; 0.54] | 0.88 [0.83; 0.92] |
| Insurance-C | 0.83 [0.82; 0.84] | 0.47 [0.42; 0.51] | 0.90 [0.89; 0.91] |
| Lymphoma | 0.60 [0.60; 0.60] | 0.32 [0.31; 0.32] | 0.50 [0.47; 0.52] |
| Ohsumed | 0.02 [0.01; 0.03] | 0.01 [0.00; 0.01] | - |
| Ovarian | 0.62 [0.62; 0.63] | 0.50 [0.50; 0.51] | 0.14 [0.14; 0.14] |
| Wine | 0.83 [0.74; 0.90] | 0.58 [0.52; 0.64] | 0.99 [0.47; 1.00] |
| p53 | 0.91 [0.91; 0.91] | 0.45 [0.44; 0.45] | 0.87 [0.87; 0.87] |
| Mean over data sets | 0.64 [0.62; 0.65] | 0.38 [0.35; 0.40] | 0.76 [0.68; 0.77] |
| On all predictions | 0.73 [0.73; 0.73] | 0.58 [0.57; 0.58] | 0.89 [0.89; 0.89] |

Table 6.7: Correlations among predicted $\hat{r}_{YZ}$ and sample-estimated $r_{YZ}$; the 95% confidence intervals are shown in brackets.



(a) ACPJ-Etiology

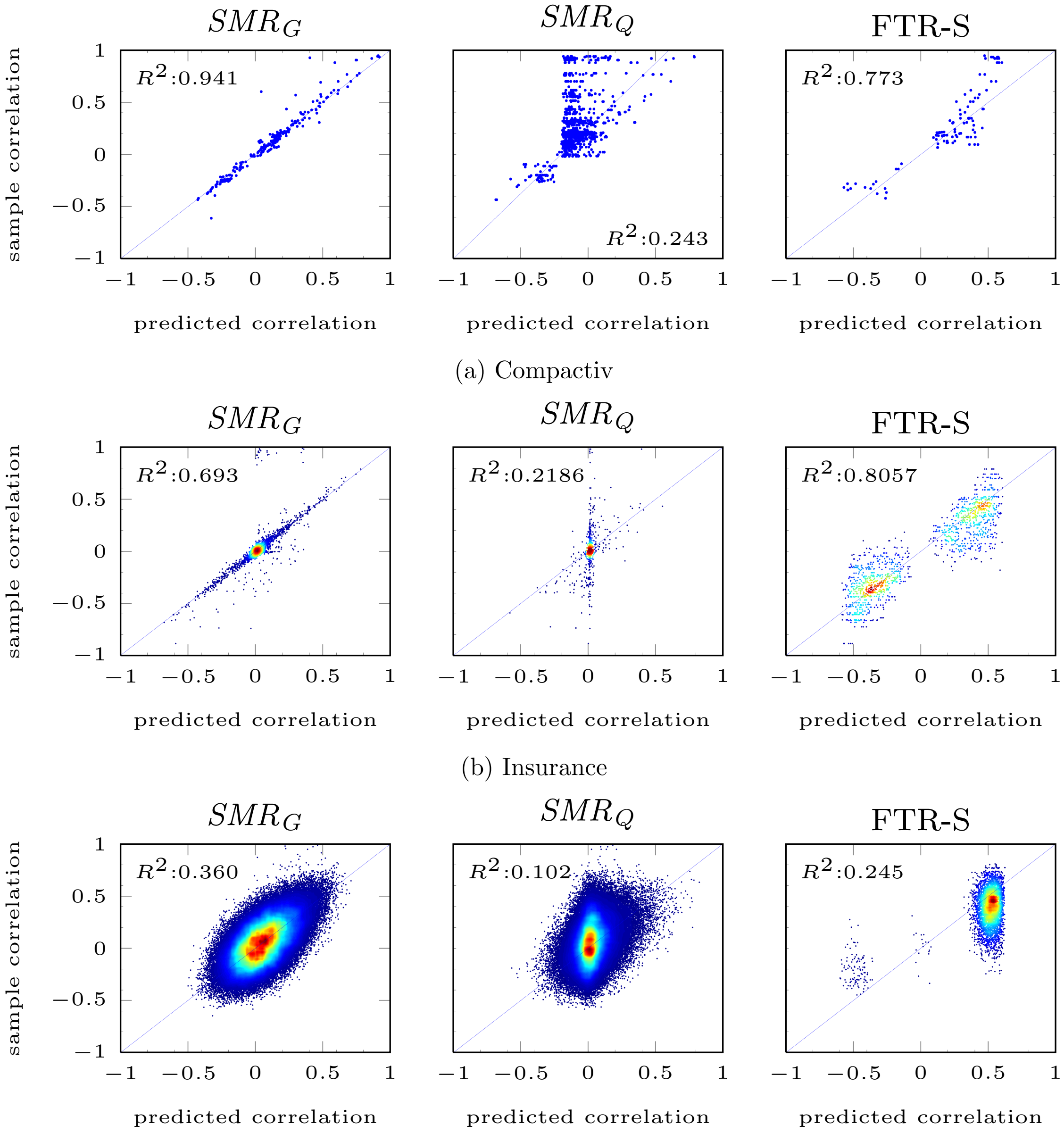
(b) Breast-Cancer

(c) C&C

Figure 6.17: **FTR vs statistical matching (part a).** Predicted vs Sample Correlations for $SMR_Q$, $SMR_G$, FTR-S

### Summary, Interpretation, and Conclusions

The CIA assumption is the most common assumption in statistical matching to produce point-estimates of the unknown distribution parameters. In comparison to FTR-S, we note the following:

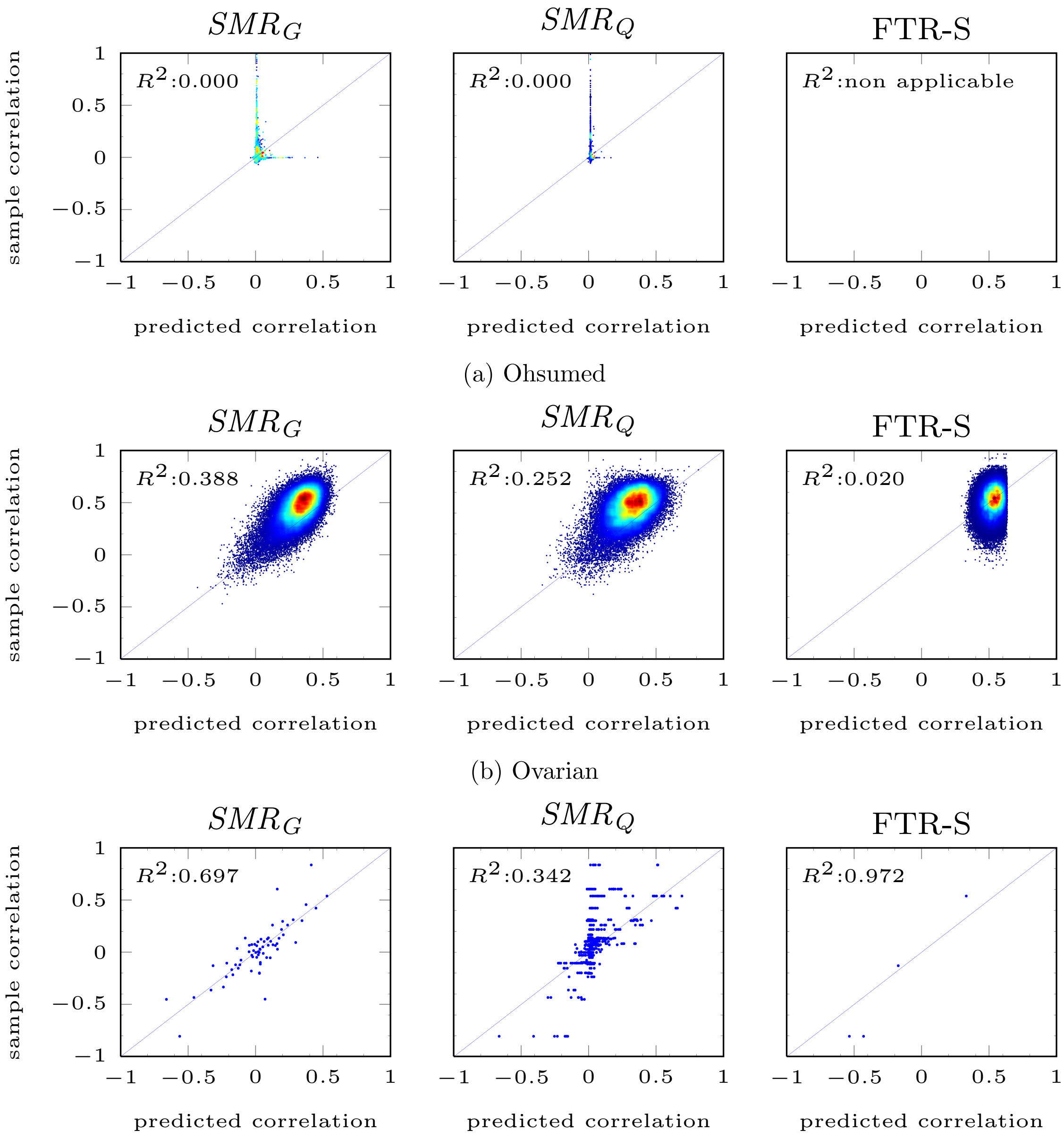


(a) Compactiv

(b) Insurance

(c) Lymphoma

Figure 6.18: **FTR vs statistical matching (part b).** Predicted vs Sample Correlations for $SMR_Q$, $SMR_G$, FTR-S

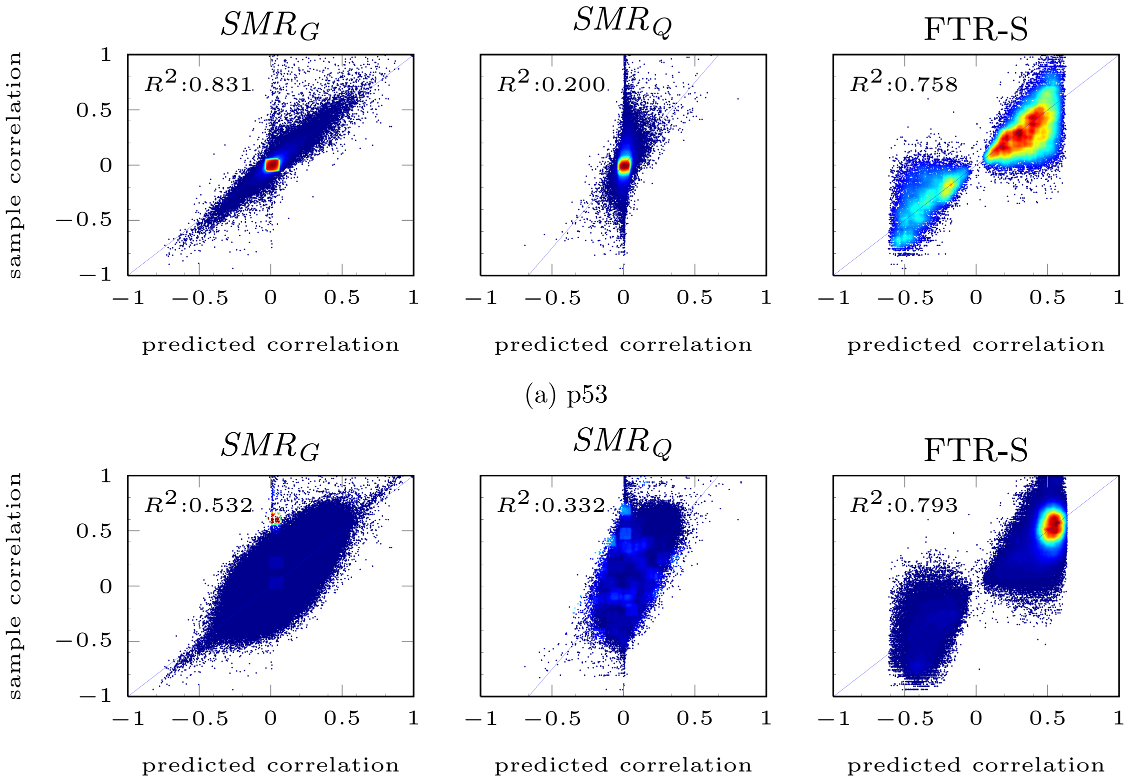


(a) Ohsumed

(b) Ovarian

(c) Wine

Figure 6.19: **FTR vs statistical matching (part c).** Predicted vs Sample Correlations for $SMR_Q$, $SMR_G$, FTR-S



(a) p53

(b) All predictions

Figure 6.20: .
**FTR vs statistical matching (part d).** Predicted vs Sample Correlations for $SMR_Q$, $SMR_G$, FTR-S

- When predictions are based on only 2 common variables, statistical matching based on the CIA (SMR$_Q$) is unreliable in several data sets and particularly the text categorization ones: the correlation of predicted vs. sample estimates in ACPJ, Breast-Cancer, and Ohsumed is less than 0.3 (Table 6.4). In general, SMR tends to predict a zero correlation between the two variables $Y$ and $Z$: the point-clouds in Figures 6.17, 6.18, 6.19 and 6.20 are vertically oriented around zero. While SMR gives a prediction in every case, it is too liberal in its predictions and the CIA is often violated, as expected by Theorem 6.6.1. Over all predictions, the correlation of predicted vs. sample estimates is 0.58.

- When predictions are based on larger sets of common variables statistical matching based on the CIA (SMR$_G$) is more successful. Over all predictions, the correlation of predicted vs. sample estimates is 0.73. The method still fails however, on the text data (ACPJ, Ohsumed) where the predictions are not correlated at all with the sample estimates. On the other hand, FTR-S does not make any predictions on these data sets.

- FTR-S's predictions are highly correlated with sample estimates (0.89 correlation), which is the highest correlation achieved by any of the three methods. However, we point out that these metrics are computed on different sets of predictions and their comparative interpretation is not straightforward.

- FTR-S is a novel alternative to statistical matching based on the CIA. FTR-S predictions are better correlated with the sample estimates of the unknown parameters, particularly when the number of common variables is low; we thus recommend that FTR-S should be preferred than existing statistical matching alternatives making the CIA in such cases.

### Statistical Matching Based on the Assumption of Multivariate Normality

The conditional independence assumption attempts to overcome the lack of joint information of the variables of interest. However, it can often be a misspecified assumption as pointed out in the literature (D'Orazio et al., 2006) and the simulated results above. An alternative approach, is to limit oneself to an assumption involving only the shape of the distribution. The most common distributional assumption adopted by statistical matching techniques for continuous variables is multivariate normality. Of course, multivariate normality alone does not allow the estimation of the parameters of the model. It does, however, impose some constraints on the parameters. These constraints stem from the positive semi-definiteness of the covariance matrix in multivariate normal distributions, thus, they naturally apply to any distribution with a positive semi-definite covariance matrix.

Let us consider again *standardized* variables $\{\mathbf{X}, Y, Z\}$ and assume their joint is distributed as multivariate normal with correlation / covariance matrix $\Sigma$ (which is symmetric)

$$\Sigma = \begin{bmatrix} \rho_{\mathbf{XX}} & \rho_{\mathbf{X}Y} & \rho_{\mathbf{X}Z} \\ \rho_{Y\mathbf{X}} & 1 & \rho_{YZ} \\ \rho_{Z\mathbf{X}} & \rho_{ZY} & 1 \end{bmatrix}.$$

The unknown quantity in the problem is parameter $\rho_{YZ}$. One can start from the requirement that $\Sigma$ must be positive semi-definite to prove that $\rho_{YZ}$ must lie within the interval $C \pm \sqrt{(D)}$ (Moriarity and Scheuren, 2001), where

$$C = \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{YX_i} \times B^{i,j} \times \rho_{ZX_j}$$

and

$$D = [1 - \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{YX_i} \times B^{i,j} \times \rho_{YX_j}] \times [1 - \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{ZX_i} \times B^{i,j} \times \rho_{ZX_j}]$$

where $p$ is the cardinality of set $\mathbf{X}$, and $B$ is the inverse of $\rho_{\mathbf{XX}}$, and $B^{i,j}$ is $B$'s $i, j$ element. This constraint is equivalent stating that the partial correlation $\rho_{YZ|\mathbf{X}}$ parameter can range freely in the interval [-1, 1]. Instead, the CIA specifies that $\rho_{YZ|\mathbf{X}} = 0$, that is, the mid-point of the interval.

The formula above can be applied to quadruples of variables to produce bounds for the unknown parameter $\rho_{YZ}$. The usefulness of such a prediction depends, of course, on the length of the predicted interval. In case the interval does not include 0, we may also say that the method *predicts an unconditional independence for* $Y$ *and* $Z$. This procedure is described in Algorithm 13. In practice, we apply Algorithm 13 using the sample estimates $\hat{r}$ in place of the unknown population parameters

$\rho$. The sample estimates are the maximum likelihood ones. The uncertainty of the estimation could be considered in the computation of the intervals by considering the worst case over all correlation estimates $\hat{r}$ that belong in the 95% confidence interval of their corresponding $\rho$. However, in this case the algorithm would produce wider intervals and thus fewer predictions.

**Algorithm 13:** Predict Dependency and Its Strength: Multivariate Normality Rule (**MNR**)

```
Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X, Y, W\}$ and $\{X, Z, W\}$, respectively
Compute sample correlation matrix $\Sigma$ (except unknown quantity $\rho_{YZ}$) ;
$MNI \leftarrow [C - \sqrt{(D)}, C + \sqrt{(D)}]$;
if $0 \notin MNI$ then
    Predict $Y \not\perp\!\!\!\perp Z \mid \emptyset$ ;
end
Predict $\hat{r}_{XY} \in MNI$
```

### Empirical Evaluation and Comparison of MNR and FTR

In order to evaluate how often MNR provides a prediction, we applied Algorithm 13 on real data. Applying Algorithm 13 on all possible combinations of four variables is prohibitive. Thus, to evaluate the MNR we randomly sampled 1000 quadruples from each group of 50 variables in each data set, for all data sets with continuous variables; For the Wine data set we generated all possible 495 quadruples out of its 12 variables.

Table 6.8 reports MNR performances on the randomly chosen quadruples. The columns of the table present the total number of randomly chosen quadruples (1000 × the number of chunks, except for the Wine data set), the number of predictions made by MNR on these random quadruples, the accuracies $Acc^{MNR}$ and $Acc^{FTR}$ at threshold $t = 0.05$. We then calculate (project) the *expected* number of predictions by the MNR rule, had it been applied on all possible quadruples. The final column presents the ratio of the number of predictions by the FTR rule over the *expected* number of predictions made by the MNR rule on all possible quadruples.

First, notice that MNR, similarly to FTR, does not provide any predictions for the text data sets ACPJ and Ohsumed data sets. Second, the rule is in general, highly accurate and on par with FTR. The most important observation however, is that the MNR does not outperform FTR in the number of predictions. The number of predictions made by FTR ranges from about 25% to 50% of those made by MNR (in four out of eight data sets) to 4 to 6 times more than MNR in the remaining data sets.

To examine whether the predictions of MNR rule overlap with those of FTR, we applied the MNR rule on the quadruples where FTR makes a prediction. The comparison is shown in Table 6.9. *MNR is able to predict a dependence only for* 1% *to* 25% *of FTR predictions.* The results in both Tables 6.8 and 6.9 clearly indicate that the two methods share only a small subset of common predictions, and thus neither method subsumes the other.

### Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:

| Data Set | # rand. quads sampled | #MNR predictions on sampled quads | $ACC^{MNR}$ | $ACC^{FTR}$ | #FTR predictions / #expected MNR predictions on all quads |
|---|---|---|---|---|---|
| Breast-Cancer | 356000 | 2 | 0.50 | 0.84 | 3.98 |
| C&C | 1000 | 45 | 1.00 | 0.96 | 0.02 |
| Compactiv | 1000 | 30 | 1.00 | 1.00 | 0.62 |
| Insurance-C | 1000 | 4 | 0.75 | 0.97 | 0.24 |
| Lymphoma | 147000 | 12 | 0.67 | 0.82 | 2.79 |
| Ovarian | 43000 | 391 | 0.99 | 0.99 | 5.99 |
| p53 | 108000 | 39 | 1.00 | 0.97 | 5.19 |
| Wine | 495 | 7 | 1.00 | 1.00 | 0.57 |

Table 6.8: **Comparison between FTR vs. MNR in predicting unconditional dependencies on randomly sampled quadruples.** The columns are: the data set name, the total number of randomly sampled quadruples (1000 × the number of chunks, except for the Wine data set), the number of predictions made by MNR on those, the accuracies $Acc^{MNR}$ and $Acc^{FTR}$ at threshold $t = 0.05$. The final column presents the ratio of the number of predictions by the FTR rule over the *expected* number of predictions made by the MNR rule on all possible quadruples. The number of predictions made by FTR ranges from about 25% to 50% of those made by MNR to 4 to 6 times more than MNR.

| Data Set | #FTR predictions | #MNR predictions restricted to cases FTR makes a prediction | % common predictions | $ACC$ of both MNR and FTR |
|---|---|---|---|---|
| Breast-Cancer | 1833 | 32 | 0.02 | 1.00 |
| C&C | 99241 | 10640 | 0.11 | 1.00 |
| Compactiv | 135 | 28 | 0.21 | 1.00 |
| Insurance-C | 1839 | 15 | 0.01 | 1.00 |
| Lymphoma | 7712 | 681 | 0.09 | 0.97 |
| Ovarian | 539165 | 59327 | 0.11 | 1.00 |
| p53 | 46647 | 413 | 0.01 | 1.00 |
| Wine | 4 | 1 | 0.25 | 1.00 |

Table 6.9: **Comparison between FTR vs. MNR in predicting unconditional dependencies on the cases where both rules apply.**

- It is possible to predict the presence of dependencies and bound their strength with distributional assumptions other than faithfulness, such as multivariate normality.
- The sets of predictions entailed by assuming faithfulness (FTR) and multivariate normality (MNR) do not overlap to a significant degree and neither method subsumes the other and they could be considered complementary. For example, the MNR makes a prediction only in the 1% to 25% of cases where FTR applies. In addition, in some data sets MNR makes only 2% of the number of FTR predictions, while in others MNR makes 6 times more predictions.

*Chapter 7*

# Discussion

*This chapter presents an overview of the thesis' contributions to the field, and discusses the challenges and future directions of causal analysis.*

This thesis tries to address the issue of integratively analyzing data sets that can be heterogeneous in terms of measured variables and experimental conditions. We argue that causality connects all observed data sets to the underlying causal structure, and propose the approach of Integrative Causal Analysis (INCA), which attempts to identify one or all causal models that are consistent with *all* available data sets. In this framework, we propose a set of INCA algorithms that address the problem from different perspectives: from COmbINE, that solves the general problem for up to 100 variables, to FTR-S, a local and extremely conservative algorithm that opportunistically works on two overlapping data sets and a total of four variables.

## 7.1 Summary of contributions

Integrative analysis of multiple data sets is a goal of several subfields of statistics and machine learning. A more detailed review of relevant methods is presented in Section 6.5. However, different experimental designs cannot be formally represented outside the framework of causality. Thus, co-analyzing data sets coming from different manipulated distributions has been addressed mainly within the field of causal discovery. Related work in the field of causal discovery is discussed in Section 3.4.

This thesis proposes several algorithms that address different causal discovery tasks. The empirical results show that the proposed algorithms' predictions are reasonable, indicating the potential of the approach. But more importantly, what we consider as the main contribution of this thesis to the scientific field of causal discovery lies in several concepts that underly the algorithms and can help pioneer promising directions in the field of causal discovery.

Overall, this thesis makes the following contributions to the field of causal analysis from multiple heterogeneous data sets:

- A **thorough analysis of causal models under causal insufficiency**, and a detailed comparison between Maximal ancestral graphs and Semi-Markov Causal Models. While the two causal graphical models are the most common representations of causally insufficient systems, a direct comparison of the two like the one presented in Chapter 2 had not been attempted. SMCMs admit a straight-forward interpretation, but no learning algorithm exists in the general case. On the other hand, invariant characteristics of MAGs can be identified using constraint-based methods, but their very complicated semantics do not always allow predicting the effect of manipulations . Thus, it is important that the correspondence between the two types of models is understood.

- The **introduction of SAT-based causal analysis**. The conversion of a causal discovery problem to a SAT instance makes the methods easily extendable to other inference tasks. For example, for the particular problems discussed in this thesis, one could opt to obtain all the SMCMs that are possibly underlying for a given data set; there is no other known procedure for this task. Alternatively, one could easily query whether there are solution models with certain structural characteristics of interest (e.g., a directed path from $A$ to $B$); this is easily done by imposing additional SAT clauses expressing the presence of these features. Incorporating certain types of prior knowledge such as causal precedence information can also be achieved by imposing additional path constraints. In essence, any information that can be expressed as graph constraints on the causal structure can be added to the SAT formula. In addition, using satisfiability instead of custom constraint–satisfaction instances offers the leverage from the efficiency of state-of-the-art solvers, and connects the algorithms' efficiency to the ever-growing research on SAT solvers. SAT-based causal analysis, introduced in Triantafillou et al. (2010) has since been adopted by different causal discovery groups (Claassen and Heskes, 2010b; Hyttinen et al., 2013).

- It proposes **query-based causal discovery to avoid the explosion of possible solutions**. The first algorithms dealing with the problem of causal structure learning from overlapping variable sets (Tillman, 2009) attempted to identify *all* possible solutions, thus limiting their scalability since the number of possible solutions is at least exponential to the input size. We propose that the algorithms output instead a summary of the structural characteristics of the underlying SMCM, distinguishing between the characteristics that are identifiable from the data (e.g., causal relations that are postulated as present), and the ones that are not (e.g., relations that could be present or not).

- It introduces a **new algorithm for estimating posterior probabilities in networks learnt using constraint-based methods**. The algorithm fits a probability density function to the a list of p-values computed during the search stage of the algorithm, and therefore has the significant advantage of having no computational overhead, so it can practically scale to any input size. In addition, the algorithm can be used with any constraint-based algorithm and any type of data, provided a suitable test of conditional independence. Despite its scalability, the method is shown to produce probability estimates that are comparable, if not better, to more expensive exhaustive Bayesian methods. Equipping constraint-based causal discovery algorithms with a method that can provide some measure of confidence on their output improves their usability. Estimating posterior probabilities based on p-values can be of use in several causal discovery tasks, including conflict resolution, improving orientations, and experiment selection.

- It advocates that **being local and conservative improves the application of causal–discovery methods to real–world data sets**. While constraint–based algorithms are fast and scalable, they are typically sensitive to error propagation. In addition, since they employ causal assumptions (particularly faithfulness and acyclicity) to limit the search space and accelerate the search, they are very susceptible to violations of these assumptions. These weaknesses question their applicability to real data sets, where noise and violation of the assumptions is very common. Local and conservative algorithms as an antidote to these problems is investigated in Chapter 6, where we identify a local rule where the INCA idea provides testable predictions. Specifically, it predicts the presence and strength of an unconditional dependence, and a chain-like causal structure (entailing several additional conditional dependencies). The idea is then implemented in three versions of increasing conservatism: the Minimal-Testing Rule (MTR), the Full-Testing Rule (FTR) and FTR-S that additionally predicts the strength of the dependence (and filters out additional predictions). The results indicate that exhaustively testing *all* conditional independencies and keeping only conservative predictions increases the validation of the rules in hold–out test–sets.

- **A proof of concept that causal assumptions can make testable predictions** and can be exploited for novel statistical inferences. In the experiments in Chapter 6, the causal semantics of the models are not employed to predict the effect of manipulations; their ability to represent independencies, based on the assumption of Markov condition and faithfulness is. The results support that the assumptions often hold to a good degree of approximation in many real systems. While this is not a direct proof in favor of the causal semantics of the models, we do note that both assumptions have been inspired by theories of probabilistic causality.

## 7.2 Limitations and Future Work

While inducing causal models from observational data has been long debated (Pearl, 2000; Spirtes et al., 2000; Pearl, 2009), there has been a growing enthusiasm for causal Bayesian networks and related methods over the past few years, particularly in the domain of molecular biology, as a result of: (a) the success of applications in machine learning tasks like feature selection (see for example Aliferis et al., 2010) (b) the bulk of public data sets produced, driven by the exponential increase in the capacity of data–production technologies and (c) some very successful applications of causal Bayesian networks in molecular biology (e.g. Sachs et al., 2005; Schadt et al., 2005; Maathuis et al., 2010) that advertised automatic causal discovery from high-throughput data.

However, significant progress is needed before successful *causal* predictions in real applications become the norm. Application of causal discovery methods may be more challenging than it initially appears to be due to the implicit possible violations of the methods' underlying assumptions. Apart from (some of the) causal assumptions described in Chapter 1, each causal discovery algorithm makes additional data–related assumptions, including independent, identically distributed data, limitations in measurement error, measurements representing random variables, and parametric assumptions related to the statistics used for independence testing or scoring. However, within the machine–learning community, algorithms are typically tested in ideal settings: valid assumptions, sparse generating networks, abundant samples, data types and parametric assumptions for which appropriate statistics are available.

It is therefore important to investigate the extent to which possible violations affect the algorithms' outcomes. This has been done for some of the assumptions (Chu et al., 2003; Ramsey et al., 2006). Moreover, algorithms that work with relaxed sets of assumptions are also becoming popular

(Richardson, 1996; Shimizu et al., 2006; Hoyer et al., 2008; Lacerda et al., 2008; Hoyer et al., 2009; Peters et al., 2011). A detailed review of such methods is presented in Eberhardt (2013).

Other approaches can be helpful in increasing the applicability of causal discovery methods in real data sets. In Chapter 6, we advocated that being local and conservative increases the success of causally–inspired predictions. Maathuis et al. (2010) also follow a local approach to successfully identify lower bounds of causal effects.

Benchmarking could also help increase the reliability of causal methods. Unfortunately, in most real applications the *causal* ground truth is unknown. Hopefully however, the increasing amount and decreasing cost of data production will eventually lead to accessible and annotated data repositories where the reproducibility and predictions of causal algorithms can be tested.

Causal discovery methods can provide novel findings and useful insights into the problem at hand. To do so however, scientists have to make sure that the assumptions, limitations and technicalities of the methods, as well as the idiosyncrasies of each specific application, are duly taken into account. Thus, another important step towards increasing the applicability of causal discovery methods is to have scientists that are knowledgeable in both areas.

Despite these open challenges, causal discovery methods constitute a valuable ally in the scientific process. Scientists can save significant time and money normally spent in intensive experiments, and use automated causal discovery applied on existing data as a guide for scientific design. Taking into account the increasing interest and the recent advances in causal discovery methods, we expect that automated causal discovery will soon become part of the standard data analysis arsenal.

# Bibliography

B Abramson, J Brown, W Edwards, A Murphy, and RL Winkler. Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71, 1996.

J Alcalá-Fdez, L Sánchez, S García, M J Jesus, S Ventura, J M Garrell, J Otero, C Romero, J Bacardit, V M Rivas, J C Fernández, and F Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, 2009.

RA Ali, TS Richardson, and P Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, October 2009.

CF Aliferis, A Statnikov, I Tsamardinos, S Mani, and X Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part i : Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:235–284, 2010.

Y Aphinyanaphongs, AR Statnikov, and CF Aliferis. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *JAMIA*, 13(4):446–455, 2006.

A Balke and J Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, 1997.

IA Beinlich, HJ Suermondt, RM Chavez, and GF Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Second European Conference on Artificial Intelligence in Medicine*, volume 38, pages 247–256. Springer-Verlag, Berlin, 1989.

SC Bendall, EF Simonds, P Qiu, El-ad D Amir, PO Krutzik, R Finck, RV Bruggner, R Melamed, A Trejo, OI Ornatsky, RS Balderas, SK Plevritis, K Sachs, D Peér, SD Tanner, and GP Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

J Binder, D Koller, S Russell, and K Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29, 1997.

JA Blackard and DJ Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

B Bodenmiller, ER Zunder, R Finck, TJ Chen, ES Savig, RV Bruggner, EF Simonds, SC Bendall, K Sachs, PO Krutzik, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology*, 30(9):858–867, 2012.

G Borboudakis, S Triantafillou, and I Tsamardinos. Tools and algorithms for causally interpreting directed edges in maximal ancestral graphs. In *Sixth European Workshop on Probabilistic Graphical Models(PGM)*, 2012.

C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

M Cannataro, D Talia, and P Trunfio. Distributed data mining on the grid. *Future Gener. Comput. Syst.*, 18:1101–1112, October 2002.

T Chu, C Glymour, R Scheines, and P Spirtes. A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152, 2003.

T Claassen and T Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 1–9, 2010a.

T Claassen and T Heskes. Learning causal network structure from multiple (in) dependence models. In *Proc. of the Fifth European Workshop on Probabilistic Graphical Models (PGM)*, pages 81–88, 2010b.

T Claassen and T Heskes. A Bayesian Approach to Constraint Based Causal Inference. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 207–217, 2012a.

T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 2992–2996, 2012b.

D Colombo, MH Maathuis, M Kalisch, and TS Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 02 2012.

TP et al. Conrads. High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*, 11(2):163–78, 2004.

GF Cooper and Ch Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, volume 10, pages 116–125, 1999.

P Cortez, Ao Cerdeira, F Almeida, T Matos, and J Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009. ISSN 01679236.

R Cudeck. An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, 65(4):539–546, 2000.

David Danks. Scientific coherence and the fusion of experimental results. *The British journal for the philosophy of science*, 56(4):791–807, 2005.

SA Danziger, R Baronio, L Ho, L Hall, K Salmon, GW Hatfield, P Kaiser, and RH Lathrop. Predicting positive p53 cancer rescue regions using most informative positive (MIP) active learning. *PLoS Computational Biology*, 5(9):12, 2009.

M D'Orazio, MD Zio, and M Scanu. *Statistical Matching: Theory and Practice*. Wiley, 2006.

D Eaton and K Murphy. Bdagl: Bayesian dag learning. `http://www.cs.ubc.ca/~murphyk/Software/BDAGL/`, 2007a.

Daniel Eaton and Kevin P Murphy. Exact bayesian structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 107–114, 2007b.

F Eberhardt. All of causal disovery. In *Workshop in Case Studies of Causal Discovery with Model Search, Carnegie Mellon University*, 2013. URL `http://www.hss.cmu.edu/philosophy/casestudiesworkshop.php`.

F Eberhardt and R Scheines. Interventions and causal inference. *Philosophy of science*, 74(5): 981–995, 2007.

N Eén and N Sörensson. An extensible SAT-solver. In *Theory and Applications of Satisfiability Testing*, pages 333–336, 2004.

C Elkan. Magical thinking in data mining: lessons from CoIL challenge 2000. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 426–431, 2001.

RJ Evans and TS Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the 26th International Conference on Uncertainty in Artificial Intelligence*, 2010.

RJ Evans and TS Richardson. Marginal log-linear parameters for graphical markov models. *arXiv preprint arXiv:1105.6075*, 2011.

RA Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

A Frank and A Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

N. Friedman and D. Koller. Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50(1-2):95–125, 2003.

N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks. In *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*, pages 196–205, 1999.

D Geiger and D Heckerman. Learning gaussian networks. In *Proceedings of the 10th Annual Conference on Uncertainty in Artificial Intelligence*, pages 235–243, 1994.

JH Gennari, ML Neal, BE Carlson, and DL Cook. Integration of multi-scale biosimulation models via light-weight semantics. *Pacific Symposium On Biocomputing*, 425:414–25, 2008.

L Getoor and B Taskar. *Introduction to Statistical Relational Learning*, volume L. The MIT Press, 2007.

CP Gomes, B Selman, N Crato, and H Kautz. Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. *Journal of automated reasoning*, 24(1-2):67–100, 2000.

HA Guvenir and I Uysal. Bilkent University function approximation repository, 2000. URL `http://funapp.cs.bilkent.edu.tr`.

I Guyon, S Gunn, M Nikravesh, and L Zadeh. *Feature Extraction, Foundations and Applications*. Springer–Verlag, Berlin, Germany, 2006a.

I Guyon, A Saffari, G Dror, and J Buhmann. Performance prediction challenge. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1649–1656, 2006b.

A Hauser and P Bühlmann. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *JMLR*, 13, August 2012.

PO Hoyer, S Shimizu, AJ Kerminen, and M Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362 – 378, 2008.

PO Hoyer, D Janzig, JM Mooij, J Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696. 2009.

A Hyttinen, F Eberhardt, and PO Hoyer. Learning linear cyclic causal models with latent variables. *JMLR*, 13:3387–3439, 2012a.

A Hyttinen, F Eberhardt, and PO Hoyer. Causal discovery of linear cyclic models from multiple experimental data sets with overlapping variables. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012b.

A Hyttinen, PO Hoyer, F Eberhardt, and M Järvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

S Itani, M Ohannessian, K Sachs, GP Nolan, and MA Dahleh. Structure learning in causal cyclic networks. In *JMLR Workshop and Conference Proceedings*, volume 6, pages 165 – 176, 2010.

Th Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. (The Kluwer International Series in Engineering and Computer Science)*. Springer, 2002.

RM Karp. Reducibility Among Combinatorial Problems. In *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972. ISBN 9783540682745.

SH Kim. Markovian combination of subgraphs of DAGs. In *Proceedings of The 10th IASTED International Conference on Artificial Intelligence and Applications*, pages 90–95, 2010.

SH Kim and S Lee. *New Developments in Robotics, Automation and Control*. In-Tech, Vienna, Austria, 2008.

M Koivisto. Advances in exact Bayesian structure discovery in Bayesian networks. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 241–248, 2006.

M Koivisto and K Sood. Exact Bayesian structure discovery in Bayesian networks. *JMLR*, 5: 549–573, 2004.

A Kuegel. Improved exact solver for the weighted max-sat problem. In *Workshop Pragmatics of SAT*, 2010.

G Lacerda, P Spirtes, J Ramsey, and P Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008.

MH Maathuis and D Colombo. A generalized backdoor criterion. *Annals of statistics*, to appear.

MH Maathuis, M Kalisch, and P Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164, 2009.

MH Maathuis, D Colombo, M Kalisch, and P Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010. ISSN 15487105.

S Mani and GF Cooper. Causal discovery using a Bayesian local causal discovery algorithm. *Medinfo 2004*, 11:731–735, 2004.

S Meganck, S Maes, P Leray, and B Manderick. Learning semi-markovian causal models using experiments. In *Third European Workshop on Probabilistic Graphical Models(PGM)*, 2006.

JM Mooij and T Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 2013.

C Moriarity and F Scheuren. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17:407–422, 2001.

K Murphy. Active learning of causal bayes net structure. Technical report, UC Berkeley, 2001.

K O'Rourke. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12):579–582, 2007.

SJ Pan and Q Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

J Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 0-934613-73-7.

J Pearl. *Causality: Models, Reasoning and Inference*, volume 113 of *Hardcover*. Cambridge University Press, 2000.

J Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

K Pearson. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 50(50):157–175, 1900.

CS Peirce and J Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885.

JP Pellet and A Elisseeff. Finding latent causes in causal networks: an efficient approach based on Markov blankets. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, 2008.

J. Pena, T. Kocka, and J. Nielsen. Featuring multiple local optima to assist the user in the interpretation of induced Bayesian Network models. In *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 1683–1690, 2004.

J Peters, J Mooij, D Janzing, and B Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.

J Ramsey, P Spirtes, and J Zhang. Adjacency faithfulness and conservative causal inference. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

TS Richardson. *Feedback Models: Interpretation and Discovery.* PhD thesis, Carnegie Mellon, 1996.

TS Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

TS Richardson and P Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4): 962–1030, 2002.

TS Richardson, JM Robins, and I Shpitser. Nested markov properties for acyclic directed mixed graphs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, page 13. 2012.

A Rosenwald et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N Engl J Med*, 346(25):1937–1947, 2002.

DG Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69:467–474, 1974.

K Sachs, O Perez, D Pe'er, DA Lauffenburger, and GP Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

K Sadeghi. *Graphical representation of independence structures.* PhD thesis, Oxford University, 2012.

G Samsa, G Hu, and M Root. Combining information from multiple data sources to create multivariable risk models: Illustration and preliminary assessment of a new method. *Journal of Biomedicine and Biotechnology*, 2005(2):113–123, 2005.

EE Schadt, J Lamb, X Yang, J Zhu, S Edwards, D GuhaThakurta, SK Sieberts, S Monks, M Reitman, C Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

T Sellke, MJ Bayarri, and JO Berger. Calibration of $\rho$ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

S Shimizu, PO Hoyer, A Hyvärinen, and A Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(2):2003–2030, 2006.

S Shimizu, T Inazumi, Y Sogawa, A Hyvarinen, Y Kawahara, T Washio, PO Hoyer, and K Bollen. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

I Shpitser, R Evans, TS Richardson, and JM Robins. Sparse nested markov models with log-linear parameters. In *Proceedings of the 29h Conference on Uncertainty in Artificial Intelligence*, pages 576–585. 2013.

ME Smoot, K Ono, J Ruscheinski, PL Wang, and T Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

A Spanos. Revisiting the omitted variables argument: Substantive vs. statistical adequacy. *Journal of Economic Methodology*, 13(2):179–218, 2006.

P Spirtes and TS Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

P Spirtes, C Glymour, and R Scheines. *Causation, Prediction, and Search.* MIT Press, second edition, January 2000.

JD Storey and R Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440, 2003.

J. Tian and R. He. Computing posterior probabilities of structural features in Bayesian networks. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 538–547, 2009.

J Tian and J Pearl. On the identification of causal effects. Technical Report R-290-L, UCLA Cognitive Systems Laboratory, 2003.

RE Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048. ACM, 2009.

RE Tillman and P Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 3–15, 2011.

RE Tillman, D Danks, and C Glymour. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems (NIPS*, 2008.

S Tong and D Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, pages 863–869, 2001.

S Triantafillou, I Tsamardinos, and IG Tollis. Learning causal structure from overlapping variable sets. In *Proceedings of Artificial Intelligence and Statistics*, volume 9, 2010.

I Tsamardinos and G Borboudakis. Permutation testing improves Bayesian network learning. In *ECML PKDD*, pages 322–337, 2010.

I Tsamardinos and LE Brown. Bounding the false discovery rate in local Bayesian network learning. In *Proceedings of the 23rd Conference on Artificial Intelligence (AAAI)*, pages 1100–1105, 2008.

I Tsamardinos, LE Brown, and CF Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

I Tsamardinos, S Triantafillou, and V Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *The Journal of Machine Learning Research*, 98888:1097–1157, 2012.

G Tsoumakas, I Katakis, and I Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 1–20, 2010.

B van der Zander, M Liskiewicz, and J Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the 30th International Conference on Artificial Intelligence and Statistics*, 2014.

B Vantaggi. Statistical matching of multiple sources: A look through coherence. *Int. J. Approx. Reasoning*, 49(3):701–711, 2008.

T Verma and J Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th conference on uncertainty in Artificial Intelligence*, pages 220–227, 1990.

Y et al. Wang. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, 2005.

S Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

J Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, PhD thesis, Carnegie Mellon University, 2006.

J Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008a.

J Zhang. Causal Reasoning with Ancestral Graphs. *Journal of Machine Learning Research*, 9(1): 1437–1474, 2008b.

XH Zhou, N Hu, G Hu, and M Root. Synthesis analysis of regression models with a continuous outcome. *Statistics in Medicine*, 28(11):1620–1635, 2009.

DW Zimmerman, BD Zumbo, and RH Williams. Bias in estimation and hypothesis testing of correlation. *Psicoliogica*, 24:133–158, 2003.