Learning Bayesian Networks using Minimum Free Energy Principle

Takashi Isozaki

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy in Engineering

GRADUATE SCHOOL OF INFORMATION SYSTEMS

THE UNIVERSITY OF ELECTRO-COMMUNICATIONS

MARCH 2010

Learning Bayesian Networks using Minimum Free Energy Principle

APPROVED BY SUPERVISORY COMMITTEE:

CHAIRPERSON:	Associate Professor Maomi Ueno
MEMBER:	Professor Toshio Okamoto
MEMBER:	Professor Haruhisa Takahashi
MEMBER:	Associate Professor Tomohiro Ogawa
MEMBER:	Associate Professor Yasuyuki Tahara

自由エネルギー最小原理に基づくベイジアンネットワーク学習

磯崎 隆司

概要

ベイジアンネットワークは確率グラフィカルモデリングの一種であり，結合確率分布を条件付き確率分布の積で表し，グラフ上では確率変数をノードとし依存関係を有向辺で表す．ベイジアンネットワークの学習とは一般には条件付き確率分布の推定に相当するパラメータ学習と，依存関係を推定する構造学習からなる．

標準的な学習法である最尤推定では過学習が問題であり，一般にベイズ推定が用いられるが，近年ベイジアンネットワークの学習結果がベイズ推定における事前分布のハイパーパラメータに強く依存することが明らかになっている．そのため事前知識がない場合にはハイパーパラメータを最適化することが必要であるが，理論的にも実際的にも難しい問題となっている．

そこで本論文では，ベイジアンネットワークの学習において，ベイズ推定とは異なる枠組みを提案する．それは熱力学において現れた概念である自由エネルギー最小原理に基づく．著者は尤度最大とエントロピー最大とのトレードオフが，熱力学における内部エネルギー最小とエントロピー最大とのトレードオフのメタファーとして捉えられることに着目した．しかしながら熱力学において重要な物理量である温度をどのように扱うかは一見不定である．本研究では，上記のメタファーを推し進めることで熱力学における温度と統計的学習の問題におけるデータ数との関係を仮定することによって，自由エネルギー最小原理を有効に利用できると考え，データ温度仮説を導入する．

本研究ではまず，このデータ温度仮説と自由エネルギー最小原理を用いてベイジアンネットワークのパラメータ学習を定式化する．この手法の効果を調べるために標準的なベンチマークデータを用いて実験を行なったところ，理論的あるいは実験的に推奨されているハイパーパラメータを使ったベイズ推定と同等あるいはそれ以上という有効性を示し，さらには提案手法に現れたハイパーパラメータの設定に関して，高い学習精度を維持できる共通の範囲があることを示した．その意味で提案手法はロバストであることがわかり，このハイパーパラメータを固定化して利用できる可能性を示した．

次に，パラメータ学習で提案した手法を基礎としてベイジアンネットワークの制約ベース・アプローチと呼ばれる構造学習に対しても自由エネルギー最小原理を適用する．このアプローチでは統計的仮説検定を用いて条件付き独立性を判定するステップを有するが，本論文では自由エネルギー最小原理に基づく条件付き独立性を表す不等式を導く．また提案手法は漸近領域では古典的な G^2 検定に近づくことを示し，古典的な検定方法の一つの拡張とみなせることになる．

構造学習においてもシミュレーションデータと標準的ベンチマークデータとを用いてサンプリングデータから元の構造を復元できるかを調べた．本手法を代表的な構造学習アルゴリズムに埋め込み，標準的な G^2 検定を用いた場合と比較した結果，シミュレーションデータでは有向辺の向きについて，ベンチマークデータではそれに加えて有向辺の有無についても，想定通り十分にはないデータ数において優位性を示した．さらに複雑なネットワーク構造に対しては，従来では少ないと考えられていない程度のデータ数においても効果が大きいことを示した．

ABSTRACT

Learning Bayesian Networks using
Minimum Free Energy Principle
by
Takashi Isozaki
Doctor of Philosophy in Engineering
The University of Electro-Communications
Chairperson: Associate Professor Maomi Ueno

Bayesian networks (BNs) are representative causal models and are expressed as directed acyclic graphs (DAGs) in which random variables and their dependencies are associated, respectively, with nodes and directed edges. Qualitative relations are expressed as their structures and quantitative relations are expressed as their parameters. Therefore, learning BNs require two steps of parameters and structures. Learning BN algorithms are anticipated as causal mining tools from data.

Although the maximum likelihood (ML) principle is widely used for learning, we often suffer from shortages of the data size because BNs need many data for processes that are used to deal with combined multivariate systems, and because ML estimation often falls into overfitting to a small data size. The maximum entropy (ME) principle, in contrast, states that probability distributions should be states of maximizing their entropies for no information. In fact, the mixture states of these two principles should be realized for the actual available data size. Bayesian methods, which involve prior distributions, are effective for avoiding overfitting. This prior has hyperparameters that can be interpreted as prior imaginary instances, which can easily realize the ML and ME principles with corresponding data size. However, learning performances of BNs are known to be highly sensitive to the values of hyperparameters, and it is difficult to decide the optimal values.

We specifically examine Helmholtz free energies and the principle of minimizing them as a metaphor of the tradeoff between the ML and the ME principles, for use in an alternative approach to learning. The minimum free energy (MFE) principle originates from thermodynamics, which maintains balances between minimum internal energies and maximum entropies under a given temperature in thermodynamical systems. Consequently, the author proposes an approach from a thermodynamical view for learning BNs, which is especially effective even for insufficient data. The “Data Temperature” assumption is important; it provides a meaning of temperature in use of free energies for statistical sciences. Internal energies, entropies, and temperature are defined and applied for learning parameters and structures of BNs. This approach can treat the ML and the ME principles in a unified manner of the MFE principle with varying data size.

In experiments of parameter learning with real-world datasets, our approach is superior to the Bayesian method with some values of hyperparameters recommended in recent studies, and shows non-sensitivity to the selection of hyperparameters involved in our method. In simulations and experiments using real-world datasets for structure learning, the proposed method notably improves the performance of the PC algorithm, which is a representative structure learning algorithm in terms of the direction and existence of edges for insufficient data.

ACKNOWLEDGMENTS

I would like to extend my deepest appreciation to the many people who engaged and supported me in graduate studies at the University of Electro-Communications (UEC).

First, I would like to thank Maomi Ueno for supervising this dissertation. I sincerely appreciate his diverse and continued support throughout the period of my time as a doctoral student. Additionally, he advised me appropriately on how to proceed with these studies. Furthermore, his tolerant mind provided me with opportunity to pursue my academic interests with complete freedom. Therefore, I was able enjoy the doctoral research activities and accomplish the work of exploring a new thermodynamical approach using my background knowledge.

I would like to thank Tomohiro Ogawa and Yasuyuki Tahara of UEC, who are some of my committee members, for proofreading the draft of this thesis and providing useful comments. Additionally, I thank Dr. Ogawa for discussion from the perspective of information theory. I also appreciate the other committee members, including Toshio Okamoto and Haruhisa Takahashi, for valuable comments. I also appreciate my many colleagues at UEC, including Mimpei Morishita, who has especially shared with me much time in enjoying discussion and offering mutual encouragement; I would like to thank all other colleagues as well.

I also appreciate the important work of many Japanese researchers of Bayesian networks including Joe Suzuki, Manabu Kuroki, Yoichi Motomura, Shin-ichi Minato, Taisuke Sato, and Takashi Washio for encouraging these studies and for providing valuable comments on the examinations described herein.

I also acknowledge Toru Tanaka, Toru Yamasaki, Takeshi Yoshioka, and Kengo Shinozaki of Fuji Xerox Co., Ltd. for providing the opportunity to study in the doctor course. I would like to thank Masaki Kyojima, my supervisor in the office, who has understood my passion for my studies, and who has supported and encouraged me throughout the period of doctoral study. I was able not only to enjoy this research but to finish writing this thesis early because he has been such an effective manager. I thank the members of the project in the office, who have understood and supported me.

Hirotsugu Kashimura accommodated me in his machine learning team at Fuji Xerox several years ago, which is appreciated; he allowed me to begin research into Bayesian networks. I thank Noriji Kato for supervising me in a project of semi-supervised recommendation agents and for supporting me in an important part of the thesis studies by elaborating my draft paper related to parameter learning method presented in this dissertation, Hiroshi Okamoto is acknowledged for advising me in academic research attitudes and encouraging me over the years, and all the other members for sharing time

with me together in accomplishing our team project plans.

I am grateful to all the members of the optical and electrical device projects at Fuji Xerox in which I was engaged. I was able to obtain an intuitive sense of material and physical science throughout the activities in these research projects, where I investigated behaviors of invisible impurities, invisible light wave and electrons in ferroelectric and quantum devices. I believe all the activities implicitly influenced the work presented in this dissertation.

I am particularly indebted to Shin Takagi of Fuji Tokoha University, who was the supervisor in my master course of theoretical physics at Tohoku University. I believe the training given by him for researchers has strongly supported me as a solid background.

I wish to thank my parents, Shotaro and Ikuko, and brother, Keiichiro, for their continued encouragement and support.

But certainly not least, I would like to thank my wife Misako, for supporting and for encouraging me, and for overcoming various difficulties together with me. I am at a loss to express my deepest appreciation for her. Finally, I thank my son Kentaro for always giving me blissful moments with his smile.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human beings often discover causality among observational events in daily life. Scientists and engineers are often engaged in pursuing cause–effect relations to find a new law of nature or a new trigger for improving industrial products and services. However, they have not yet discovered many cause–effect relations that are hidden even among large amounts of data. Several examples can be given readily, such as symptom–disease, gene defect – disease, factory environment data-yield ratios, and abnormal changes in global geoenvironmental data. Effective causal data mining techniques would yield meaningful fruit from any of those pools of data if humans' modes of causal discovery could be represented in computer systems because such systems present too much data to manage in the world without use of computers. There are many such data from which cause–effect relations could be mined as bioinformatics, clinical, POS, geoenvironmental, survey data on social sciences, and so on.

These cause–effect relations probably involve many events, which form network structures of the relations. In addition, because almost every event occurs under many complicated conditions including undetectable elements, we cannot definitely describe relations of cause–effect anymore. Therefore we have described those with the language of probability, and have used statistical analyses for estimation. Based on these contexts, we have adopted probabilistic graphical models [Pearl, 1988, 2000; Spirtes et al., 2000; Koller and Friedman, 2009], that have recently been hot research areas in artificial intelligence and its sub-domains: machine learning and data mining.

Among these models, we chose to examine a class of models specifically: directed acyclic graphical models, which are called *Bayesian networks (BNs)* [Pearl, 1988], representing qualitative aspects of direct dependency as directed edges and quantitative aspects of them as parameters (usually conditional probability distributions). The reasons are as follows:

- BNs have directed edges that can express the cause–effect relations.
- BNs are intuitively comprehensive for displaying the relations.
- BNs are based on the concrete probability theory for representing uncertainty.
- Regarding constructed BNs, probabilistic reasoning can be conducted using some established algorithms [Kim and Pearl, 1983; Pearl, 1986; Lauritzen and Spiegelhalter, 1988].
- Actually, BNs are important tools for estimating causal effects using interventions and *do-calculus* [Pearl, 1995].

We can obtain a tool of constructing and mining causal models from observational data if learning structures and parameters of BNs from those data are established. Many studies have been conducted for learning cause–effect models from observational data using BNs during the decades. The results show that learning causal models have remained matters of research, partly because they deal appropriately with insufficient training data. Consequently, we have addressed the issue of learning BNs that can use insufficient data.

The maximum likelihood (ML) principle is the main principle of estimation in statistical science, which states that we should estimate parameters that regenerate the obtained data. However, the estimation based on the ML is likely to fall into overfitting of the finite data. The maximum entropy (ME) principle is often used for treating sparse data such as in natural language processing, which states that one should estimate parameters so that states of those take the values maximizing entropy if we have no information. Bayesian methods have recently attracted much attention from researchers not only because of their incorporation of a prior background knowledge but also because they have a smoothing effect with prior distributions, which can avoid overfitting and include ME principles with some conditions. These prior distributions have hyperparameters that can be interpreted as prior imaginary instances (we designate it as α in this dissertation). Some studies have used hyperparameters such as $\alpha = 1$ (meaning uniform prior distributions) or $\alpha = 0.5$ when no prior knowledge exists. However, it remains controversial to decide hyperparameters of prior distributions in theoretical perspectives (meaning noninformative prior, also called Jeffrey's prior) [Gelman et al., 2004; Robert, 2007]. Furthermore, recent studies of Bayesian approaches clarified as a critical issue to decide hyperparameters because accuracy in learning BNs is heavily dependent on the selection of hyperparameters [Silander et al., 2007], and it is difficult to find optimal values.

Another approach to avoid overfitting to data is using *the principle of minimum free energy (MFE)*, which has its roots in thermodynamics. The free energy consists of internal energy, entropy and temperature. In thermodynamics, MFE maintains balances between minimizing internal energies and maximizing entropies at some constant temperature. In recent years, the MFE principle and similar concepts have been used in widely various areas of statistical science. Nevertheless, to our knowledge, the meaning of temperature has not been established to date. Consequently, temperature is treated as fixed parameters or free parameters decided in each dataset. We noticed that the approach using MFE principle is attractive because of its intuitive comprehension of the role described above. It also is inadequately investigated especially in the role of temperature for use in statistical sciences. Then we addressed this approach in this dissertation, where, differently from the other approaches, we define internal energy and entropy, and then provide a meaning of temperature in statistical science. Learning BNs has two steps of the structures and parameters, as described above. We hereby propose a new unified framework of learning parameters and structures of BNs, which includes the ML and the ME principles.

Thesis Overview

This thesis comprises the following chapters. Chapter 2 introduces Bayesian networks, which are the basis of the thesis. Herein, the model properties are described in graphical and probabilistic aspects in some detail; then we describe Reichenbach's theory, which connects statistical patterns with causal patterns, which is a foundation on which Bayesian network models are used as causal network models. Subsequently, methods of learning parameters and structures are described using standard approaches. These descriptions are necessary in later chapters. In Chapter 3, the minimum free energy (MFE) principle is introduced, which originates from thermodynamics and provides an idea for utilizing MFE principle effectively in statistical science. The idea is the "Data Temperature" assumption, which, although simple, provides a meaning or interpretation of temperature in statistical sciences. Then we apply a combination of the MFE principle and the "Data Temperature" model to parameter learning of BNs for an alternative of the Bayesian–Dirichlet methods. A comparative study using the Bayesian methods is also conducted using repository datasets that are typically used in this research area. In Chapter 4, which is a highlight of the thesis, it is applied, in turn, to structure learning BNs. In the approach of structure learning, hypothesis testing of conditional independencies is needed; its impracticality has been a disadvantage of the approach for especially insufficient data size. We describe both null and opposite hypotheses using the MFE and derive a new condition of the conditional independence: our method provides another framework for hypothesis testing. Additionally, it is shown theoretically to have a preferable property in asymptotic region. A simulation study and experiments using real-world datasets show its effectiveness through comparison with the classical tests, in which two testing methods were embedded in a representative structural learning algorithm. Finally, Chapter 5 gives final conclusions and presents future plans for additional research. Throughout the dissertation, we assume that the following conditions are satisfied: all random variables have discrete states, and there are no missing data, no latent variable models, no selection bias, and no prior background knowledge.

Chapter 2

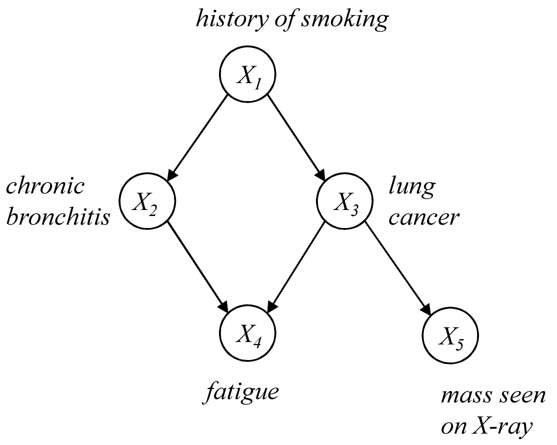
Bayesian Networks

This chapter presents our research basis—Bayesian networks— in some detail before proceeding to our approach. The reader is assumed to be familiar with the basis of probability theory (including joint probability, conditional probability and the Bayes' theorem), and statistics (including the maximum likelihood principle, hypothesis tests, χ^2-tests). After describing an example of Bayesian networks, the preliminary concepts are described along with notation related to the thesis, some graph theory, and conditional independence and dependence in probability distributions. Next the Markov condition, minimality condition and then Bayesian networks (BNs) are defined. Subsequently, we provide some properties of BNs related to the thesis. Then parameter learning methods are described—the maximum likelihood and Bayesian inference. The structure learning methods are also explained— constraint-based approaches and score-search approaches. For more details including some omitted proofs, please refer to the references or some seminar texts [Pearl, 1988, 2000; Spirtes et al., 2000; Jensen and Nielsen, 2007; Cowell et al., 1999; Neapolitan, 2004].

2.1 An Example

An instance of a Bayesian network can be described as Cooper [1999], as shown in Fig. 2.1, which represents dependence–independence relations in some diseases and symptoms. Five binary random variables are given as $\{X_1, X_2, X_3, X_4, X_5\}$, where X_1 signifies a history of smoking, X_2 denotes chronic bronchitis, X_3 stands for lung cancer, X_4 represents fatigue, and X_5 denotes a mass viewed on an X-ray image. The Bayesian network in Fig. 2.1 has the joint probability as the following factorization form based on the structure.

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2, X_3)P(X_5 \mid X_3).$$



$P(X_1\text{=no})=0.80$

$P(X_2\text{=absent} \mid X_1\text{=no})=0.95$
$P(X_2\text{=absent} \mid X_1\text{=yes})=0.75$

$P(X_3\text{=absent} \mid X_1\text{=no})=0.99995$
$P(X_3\text{=absent} \mid X_1\text{=yes})=0.997$

$P(X_4\text{=absent} \mid X_2\text{=absent}, X_3\text{=absent})=0.95$
$P(X_4\text{=absent} \mid X_2\text{=absent}, X_3\text{=present})=0.50$
$P(X_4\text{=absent} \mid X_2\text{=present}, X_3\text{=absent})=0.90$
$P(X_4\text{=absent} \mid X2\text{=present}, X_3\text{=present})=0.25$

$P(X_5\text{=absent} \mid X_3\text{=absent})=0.98$
$P(X_5\text{=absent} \mid X_3\text{=present})=0.40$

Figure 2.1: Bayesian network with five binary nodes with a conditional probability table.

The edges in the figure represent relative direct influences among variables. A conditional probability table is attached with the graph; that is, a Bayesian network denotes direct influences as the directed edges and their quantities as the conditional probability distributions. It seems reasonable to represent these relations quantitatively as (conditional) probability distributions because those relations such as in Fig. 2.1 cannot be described in deterministic language because of the existence of latent conditions that are too complicated to represent explicitly. In addition, the Bayesian network in Fig. 2.1 has only 11 independent probabilities. If one deals with the joint probabilities of these variables, then one should estimate $2^5 - 1\,(= 31)$ independent probabilities. Furthermore, the parameters of such joint probability distributions become more numerous exponentially with the number of variables. For that reason, this representation reduces the parametric space that must be estimated. Furthermore, it can be readily understood that the edges represent direct influences of a variable on other variables in the Bayesian network.

2.2 Preliminaries

We explain the symbols used throughout the thesis. General variables are denoted with upper-case letters (e.g., X, Y, Z), whereas the states or values of the variables are denoted with lower-case letters (e.g., x, y, z). Variable sets are denoted as upper-case bold-faced letters (e.g., $\boldsymbol{Z}$, $\boldsymbol{V}$), and an assignment of the states or values to each variable in the given set by lower-case bold-faced letters (e.g., $\boldsymbol{z}$, $\boldsymbol{\pi}$). However, we present exceptional notations for some symbols. In this thesis, we deal with only complete datasets and discrete random variables. We use the terms node, vertex, and variable interchangedly along with the terms edge and arc.

We first review some graph theory [Neapolitan, 2004]. A directed graph is a pair $(\boldsymbol{V}, \mathrm{E})$, where $\boldsymbol{V}$ is a finite, nonempty set whose elements are called nodes or vertices, and where E is a set of ordered pairs of distinct elements of $\boldsymbol{V}$. Elements of E are called edges or arcs; if $(X, Y) \in \mathrm{E}$, then it is said that an edge exists from X to Y or from Y to X, and that X and Y are adjacent. Presuming a set of nodes $\{X_1, X_2, \ldots X_k\}$, where $k \geq 2$, such $(X_{i-1}, X_i) \in \mathrm{E}$ for $2 \leq i \leq k$. The set of edges connecting the k nodes is called a path from X_1 to X_k $(k > 1)$. We denote an undirected path in a graph as a sequence of nodes that are adjacent in the graph. Here the directed cycle indicates a path from a node to itself. A directed graph $\mathbb{G}$ is called a directed acyclic graph (DAG) if it contains no directed cycles. Given a DAG $\mathbb{G} = (\boldsymbol{V}, \mathrm{E})$ and nodes X and Y in $\boldsymbol{V}$, Y is a parent of X if there is an edge from Y to X, Y is called a descendent of X and X is called an ancestor of Y if there is a path from X to Y, and Y is called a nondescendent of X if Y is not a descendent of X.

Next, we define conditionally independent and dependent for probability distributions:

Definition 2.1 *Presuming a probability distribution P of the random variables in some set $\boldsymbol{V}$, two variables $X \in \boldsymbol{V}$ and $Y \in \boldsymbol{V}$ are conditionally independent given $\boldsymbol{Z} \subset \boldsymbol{V}$, which we designate as*

$$Ind\,(X; Y|\boldsymbol{Z}), \tag{2.1}$$

or

$$X \perp\!\!\!\perp Y \,|\, \boldsymbol{Z}, \tag{2.2}$$

if $\forall x, y, \boldsymbol{z}$ where $P(\boldsymbol{z}) > 0$,

$$P(x, y \,|\, \boldsymbol{z}) = P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z}), \tag{2.3}$$

or we designate it for short as

$$P(X, Y|\boldsymbol{Z}) = P(X|\boldsymbol{Z}) P(Y|\boldsymbol{Z}). \tag{2.4}$$

We define and denote dependence as

$$Dep\,(X; Y|\boldsymbol{Z}) := \neg\, Ind\,(X; Y|\boldsymbol{Z}), \tag{2.5}$$

or

$$X \not\!\perp\!\!\!\perp Y \,|\, \boldsymbol{Z}. \tag{2.6}$$

We define and denote (marginal) independence when $Ind(X;Y|\boldsymbol{Z})$ *for* $\boldsymbol{Z}$*:* $\boldsymbol{Z} = \varnothing$*, as* $Ind(X;Y)$ *and* $P(X,Y) = P(X)P(Y)$*. Then* $Dep(X;Y) := \neg\, Ind(X;Y)$*. In addition, for some sets* $\boldsymbol{X}, \boldsymbol{Y}$ *and* $\boldsymbol{Z}$*, we designate*

$$Ind(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}), \tag{2.7}$$

if $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ *where* $P(\boldsymbol{z}) > 0$*,* $P(\boldsymbol{x}, \boldsymbol{y} \,|\, \boldsymbol{z}) = P(\boldsymbol{x} \,|\, \boldsymbol{z})P(\boldsymbol{y} \,|\, \boldsymbol{z})$*.*

2.3 Bayesian Networks

2.3.1 Bayesian Networks

First, we state the following definition, called the Markov condition:

Definition 2.2 (Markov Condition) *Presuming a joint probability distribution* P *of the random variables in some set* $\boldsymbol{V}$ *and a DAG* $\mathbb{G} = (\boldsymbol{V}, E)$*. We say that* $(\mathbb{G}, P)$ *satisfies the Markov condition if, for each variable* $X \in \boldsymbol{V}$*,* $\{X\}$ *is conditionally independent of the set of all its nondescendents given the set of all its parents. Therefore, if we designate the sets of parents and nondescendents of* X *as* $Pa(X)$ *and* ND_X *respectively, then*

$$Ind(X; ND_X \setminus Pa(X) \,|\, Pa(X)).$$

We then introduce and define the minimality condition as follows.

Definition 2.3 (Minimality Condition) *Presuming a joint probability distribution* P *of the random variables in some set* $\boldsymbol{V}$ *and a DAG* $\mathbb{G} = (\boldsymbol{V}, E)$*, it is said that* $(\mathbb{G}, P)$ *satisfies the minimality condition if the following two conditions hold:*

1. $(\mathbb{G}, P)$ *satisfies the Markov condition.*
2. *The resultant DAG no longer satisfies the Markov condition with* P *if we remove any edge from* $\mathbb{G}$*.*

The following definition of Bayesian networks can be stated validly:

Definition 2.4 (Bayesian networks) *Let* P *be a joint probability distribution of the random variables in a set* $\boldsymbol{V}$ *and* $\mathbb{G}$ *be a DAG. We define* $(\mathbb{G}, P)$ *as a Bayesian network (BN) if* $(\mathbb{G}, P)$ *satisfies the Markov and the minimality condition.*

It is readily proven that a joint probability distribution of the BN $(\mathbb{G}, P)$ that has n variables $\{X_1, \ldots, X_n\} \in \boldsymbol{V}$ can be factorized as the following products of the conditional probability distributions, as

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \,|\, Pa(X_i)). \tag{2.8}$$

A term *Bayesian networks* was named in Pearl [1985] to emphasize some aspects such as the subjective nature of input information, the dependence on the Bayes' theorem of updating information, and distinction between causal and evidential modes of reasoning [Pearl, 2000].

2.3.2 d-Separation

Next we describe an important DAG property called "d-separation" (Pearl [1988]; the d denotes *directional*), which plays a major role in the domain of Bayesian networks. First, we define a term collider:

Definition 2.5 *Presuming that we have a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, then a node $Z \in \boldsymbol{V}$ on an undirected path ρ is a collider if and only if two distinct incoming edges exist into Z.*

If we have $\{X, Y, Z\} \in \boldsymbol{V}$ and Z is a collider and X and Y are not adjacent, then we designate the graph as $X \rightarrow Z \leftarrow Y$, the structure of which is called a *v-structure* and Z is called an unshielded collider.

The following definition is given [Pearl, 1988]:

Definition 2.6 *Letting $\mathbb{G} = (\boldsymbol{V}, E)$ be a DAG, with variables $\{X, Y\} \in \boldsymbol{V}$, and $\boldsymbol{Z} \subset \boldsymbol{V}$, an undirected path ρ between two distinct nodes X and Y is blocked using a set of nodes $\boldsymbol{Z}$ if there is a node W on ρ for which one of the following two conditions hold:*

- *W is not a collider and $W \in \boldsymbol{Z}$, or*
- *W is a collider and neither W and its descendents are in $\boldsymbol{Z}$*

We then define "d-separation":

Definition 2.7 (d-Separation) *Presuming a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, $\{X, Y\} \in \boldsymbol{V}$, and $\boldsymbol{Z} \subset \boldsymbol{V}$, then two nodes X and Y are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}$ if and only if every undirected path between two distinct nodes X and Y is blocked by $\boldsymbol{Z}$. That is denoted by $Dsep^{\mathbb{G}}$. Two nodes are d-connected if they are not d-separated. We denote the d-separation as $Dsep^{\mathbb{G}}(X; Y|\boldsymbol{Z})$. We write $Dsep^{\mathbb{G}}(X; Y)$ if $\boldsymbol{Z} = \varnothing$.*

Furthermore, we define d-separation for subsets in nodes as follows.

Definition 2.8 *Presuming that we have a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, and $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are mutually disjoint subsets of $\boldsymbol{V}$, it can be said $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}$ if for every $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$. We write $Dsep^{\mathbb{G}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z})$. If $\boldsymbol{Z} = \varnothing$, then we write $Dsep^{\mathbb{G}}(\boldsymbol{X}; \boldsymbol{Y})$.*

Using the three lemmas provided in Appendix A, the following theorem can be proven, which states that d-separation identifies all and only the conditional independences that are necessary for $\mathbb{G}$ by the Markov condition.

Definition 2.9 *Presuming $\boldsymbol{V}$ as a set of random variables, and $\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{Z}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2$, and $\boldsymbol{Z}_2 \subset \boldsymbol{V}$, then if and only if $Ind\,(\boldsymbol{X}_2; \boldsymbol{Y}_2|\boldsymbol{Z}_2)$ hold for all probability distributions P in $\boldsymbol{V}$ and $Ind\,(\boldsymbol{X}_1; \boldsymbol{Y}_1|\boldsymbol{Z}_1)$ also does, can we we say that $Ind\,(\boldsymbol{X}_1; \boldsymbol{Y}_1|\boldsymbol{Z}_1)$ and $Ind\,(\boldsymbol{X}_2; \boldsymbol{Y}_2|\boldsymbol{Z}_2)$ are equivalent.*

The following definition is necessary before stating the main theorem.

Definition 2.10 *We say conditional independency $Ind\,(X; Y|\boldsymbol{Z})$ is identified by d-separation in $\mathbb{G}$ if one of the following holds:*

- *$Dsep^{\mathbb{G}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z})$.*
- *$\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are mutually disjoint, $\boldsymbol{X'}$, $\boldsymbol{Y'}$, and $\boldsymbol{Z'}$ are mutually disjoint, $Ind\,(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z})$ and $Ind\,(\boldsymbol{X'}; \boldsymbol{Y'}|\boldsymbol{Z'})$ are equivalent, and we have $Dsep^{\mathbb{G}}(\boldsymbol{X'}; \boldsymbol{Y'}|\boldsymbol{Z'})$.*

Theorem 2.1 (Verma and Pearl [1988]; Geiger et al. [1990]) *Based on the Markov condition, a DAG $\mathbb{G}$ entails all and only those conditional independencies that are identified by d-separation in $\mathbb{G}$.*

It is necessary to be careful to interpret Theorem 2.1 correctly. A particular distribution P might exist that satisfies the Markov condition with $\mathbb{G}$, and which has conditional independencies that are not identified by d-separation. Such distributions have particular conditional probability distributions to cancel their mutual dependencies.

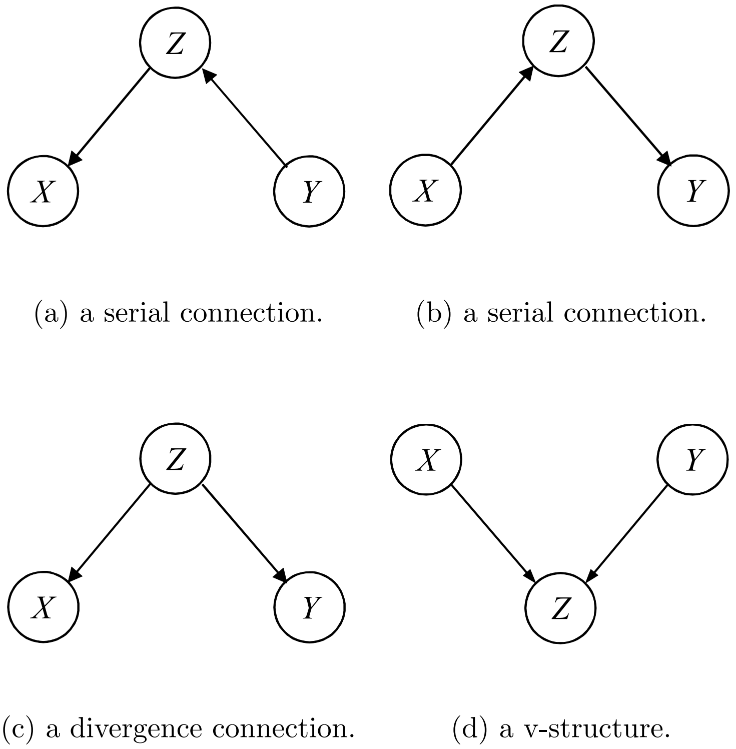
According to the Theorem 2.1, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}$, the Markov condition entails $Ind\,(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z})$. Therefore, it is said that, if $(\mathbb{G}, P)$ satisfies the Markov condition, then $\mathbb{G}$ is an independence map of P, also called an I-map of P. Then Bayesian networks are also defined as minimal I-maps [Pearl, 1988].

2.3.3 Markov Equivalence

Many DAGs are equivalent in the sense that they have the same d-separations: they cannot be distinguished by d-separation.

Definition 2.11 *Letting $\mathbb{G}_1 = (\boldsymbol{V}, E_1)$ and $\mathbb{G}_2 = (\boldsymbol{V}, E_2)$ be two DAGs containing the same set of nodes $\boldsymbol{V}$, then $\mathbb{G}_1$ and $\mathbb{G}_2$ are called Markov equivalent if, for every three mutually disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}_1$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}_2$. That is*

$$Dsep^{\mathbb{G}_1}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}) \Longleftrightarrow Dsep^{\mathbb{G}_2}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}).$$



(a) a serial connection.

(b) a serial connection.

(c) a divergence connection.

(d) a v-structure.

Figure 2.2: Simple examples of Markov equivalence. The three graphs shown in panels (a), (b), and (c) are Markov equivalent, although (d) is not equivalent.

For example, Fig. 2.2 shows simple graphs, with three panels (a), (b), and (c) showing Markov equivalence. However, (d) is not equivalent to the preceding three graphs. The types of Figs. 2.2 (a) and 2.2 (b) are called *serial connections* and Fig. 2.2 (c) is *divergence connections*, whereas the type of graph in Fig. 2.2 (d) is called *convergence connections* or also called *v-structures*. We denote the Markov equivalence class as an undirected graph. For example, Fig. 2.3 represents the Markov equivalence class in Figs. 2.2 (a)–2.2 (c). We designate this as a DAG pattern. In fact, Chickering [1995] provides an efficient method for testing whether two structures are in the same Markov equivalence class.

The previous definition is related to graph properties only. However, it is applied in probability distributions because of the following theorem:

Theorem 2.2 *Two DAGs are Markov equivalent if and only if they entail the same conditional independencies based on the Markov condition.*

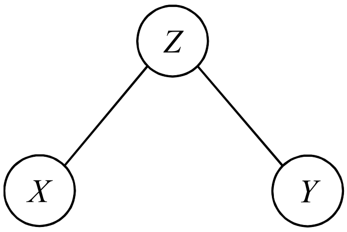


Figure 2.3: The DAG pattern represents the Markov equivalence class in Figs. 2.2 (a)–2.2 (c).

The proof follows immediately from Theorem 2.1. We describe a theorem that shows how to identify Markov equivalence. Three lemmas described in Appendix A.2 are necessary for its proof.

Theorem 2.3 (Pearl et al. [1989]; Verma and Pearl [1990]) *Two DAGs $\mathbb{G}_1$ and $\mathbb{G}_2$ are Markov equivalent if and only if they have the same links (edges without direction) and the same set of v-structures.*

Theorem 2.3 provides a simple means to represent a Markov equivalent class with a single graph: a Markov equivalent class can be represented by a graph that has the same links and the same v-structures as the DAGs in the class. Any direction can be assigned to the undirected edges in this graph, which does not create a new v-structure or a directed cycle, yields a member of the equivalent class. Therefore we define a DAG pattern for a Markov equivalent class to be the graph that has the same links as the DAGs in the equivalent class and has oriented all and only the edges common to all the DAGs in the equivalent class. All DAGs in the same Markov equivalence class have the same d-separations. Therefore, d-separation can be defined for DAG patterns:

Definition 2.12 *Let $\mathbb{G}p$ be a DAG pattern whose nodes are the elements of $\boldsymbol{V}$, and let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be mutually disjoint subsets of $\boldsymbol{V}$. We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathbb{G}p$ if X and Y are d-separated by $\boldsymbol{Z}$ in any (and therefore every) DAG $\mathbb{G}$ in the Markov equivalence class represented by $\mathbb{G}p$.*

2.3.4 Faithfulness Condition

The Markov condition implies only independencies; it does not entail any dependencies. In general, it is desirable that an edge means that a direct dependency exists. The *Faithfulness Condition* entails this.

Definition 2.13 *Presuming that we have a joint probability distribution P of the random variables in some set $\boldsymbol{V}$ and a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, it can be said that $(\mathbb{G}, P)$ satisfies the faithfulness condition if, based on the Markov condition, $\mathbb{G}$ entails all and only conditional independencies in P. That is, the following two conditions hold:*

- *$(\mathbb{G}, P)$ satisfies the Markov condition (This means $\mathbb{G}$ entails only conditional independencies in P.)*
- *All conditional independencies in P are entailed by $\mathbb{G}$, based on the Markov condition.*

When $(\mathbb{G}, P)$ satisfies the faithfulness condition, it is said that P and $\mathbb{G}$ are mutually faithful, and that $\mathbb{G}$ is a *perfect map* of P. The following theorems establish a criterion for recognizing faithfulness. In addition, those theorems are bases of the constraint-based structure learning algorithms, the type of which is an important key of this dissertation.

Theorem 2.4 (Geiger and Pearl [1988a]; Verma and Pearl [1990]) *Presuming a joint probability distribution P of the random variables in some set $\boldsymbol{V}$ and a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, then $(\mathbb{G}, P)$ satisfies the faithfulness condition if and only if all and only conditional independencies in P are identified by d-separation in $\mathbb{G}$.*

The proof follows immediately from Theorem 2.1.

We can make a very specific distribution as an example that is unfaithful to the DAG [Neapolitan, 2004]. See Appendix B. However, Spirtes et al. [2000] shows that almost all assignments of conditional probabilities generate distributions that are faithful to the DAG. Meek [1995b] extends this theorem to the case of discrete variables.

The following theorem yields the result if P is faithful to some DAG; then P is faithful to an equivalence class of DAGs:

Theorem 2.5 *If $(\mathbb{G}, P)$ satisfies the faithfulness condition, then P satisfies this condition with all and only those DAGs that are Markov equivalent to $\mathbb{G}$. Furthermore, if $\mathbb{G}p$ is the DAG pattern corresponding to this Markov equivalence class, then the d-separations in $\mathbb{G}p$ identify all and only conditional independencies in P. It can be said that $\mathbb{G}p$ and P are faithful to each other, and that $\mathbb{G}p$ is a perfect map of P.*

The proof follows immediately from Theorem 2.4.

We say that a distribution P admits a faithful DAG representation if P is faithful to some DAG and DAG pattern. A unique DAG pattern exists with which P is faithful because of the previous theorem if P admits a faithful DAG representation. Therefore, in principle, a DAG pattern can be found whenever P admits a faithful DAG representation. Methods for finding such patterns are explained in section 2.5.

We stated that, under faithfulness condition, an edge between two nodes means that a direct dependency exists between the nodes. The following theorem includes this result and more [Pearl, 1988].

Theorem 2.6 (Verma and Pearl [1990]) *Presuming a joint probability distribution P of the random variables in some set of $\boldsymbol{V}$ and a DAG $\mathbb{G} = (\boldsymbol{V}, E)$, then if P admits a faithful DAG representation, $\mathbb{G}p$ is the DAG pattern faithful to P if and only if the following two conditions hold:*

1. *X and Y are adjacent in $\mathbb{G}p$ if and only if there is no subset $\boldsymbol{Z} \subseteq \boldsymbol{V}$ such that $Ind(X;Y|\boldsymbol{Z})$. That is, X and Y are adjacent if and only if a direct dependency exists between X and Y.*

2. *$X - U - Y$ is a v-structure in $\mathbb{G}p$ if and only if $U \in \boldsymbol{Z}$ implies $\neg Ind(X;Y|\boldsymbol{Z})$.*

That proof is presented in Appendix A.3.

A DAG cannot be faithful to a distribution without satisfying the minimality condition with the distribution. However, a DAG can satisfy the minimality condition with the distribution without satisfying the faithfulness condition [Neapolitan, 2004].

A serial connection and a divergence connection, which are shown in Figs. 2.2 (a)–2.2 (c) have the same conditional independence relations $Ind(X;Y|Z)$ if each probability distributions is faithful to each corresponding DAG. In addition, a v-structure depicted in Fig. 2.2 (d) has a marginal independence relation $Ind(X;Y)$ if the probability distribution is faithful to the DAG.

2.3.5 Causal Bayesian Networks

The BNs represent probability distributions, which do not seem to represent causation. In fact, the existence of the Markov equivalence class tells us that statistically indistinguishable models of DAGs exist, as shown in Figs. 2.2 (a)–2.2 (c). However, according to discussions developed by Reichenbach [1956], the following can be said [Pearl, 2000]. In general, time ordering is considered to be the necessary condition for defining causality, which should be an important clue for distinguishing causal relations and others. Temporal information alone, however, cannot distinguish causal patterns from spurious correlations caused by unknown factors. Reichenbach, a philosopher, analyzed statistical associational patterns that represent causal organizations, which can be given meaningful interpretation only in terms of causal directionality: three events A, B, and C are assumd, with the presumption that A and B are dependent, B and C are dependent, and A and C are independent. These relations are usually interpreted as A and C as two independent causes and B as their common effect, that is, $A \rightarrow B \leftarrow C$. This problem says that there exist dependent patterns with causal directions despite a lack of temporal information.

Reichenbach [1956] also derived the relations between probability distributions and statistical cause–effect relations using directed edges representing *the direction of time* from older to newer events, and *causal networks* he called[1] from discussions of statistical patterns of events: Two variables X and Y, and their common cause Z have the same graph depicted in Fig. 2.2(c) and have the same conditional independence relation: $P(X, Y|Z) = P(X|Z)P(Y|Z)$, which indicates $\mathit{Ind}(X;Y|Z)$. A causal path $X \rightarrow Z \rightarrow Y$ is an equivalent to a serial connection shown in Figs. 2.2(a) and 2.2(b), and also has the same conditional independence relation as the divergence connection: $P(Y|X, Z) = P(Y|Z)$, which is equivalent to $P(X, Y|Z) = P(X|Z)P(Y|Z)$ (see Appendix B for the proof). Two variables X and Y, and their common effect Z have the same graph shown in Fig. 2.2(d) and have the same marginal independence relation: $P(X, Y) = P(X)P(Y)$, which indicates $\mathit{Ind}(X;Y)$. Therefore, causal patterns are associated with conditional independencies.

Based on this theory, Rebane and Pearl [1987] were acutely aware that finding these statistical patterns can engender the causal discovery from observational data, and they explored causal discovery research domains. This philosophical foundation is followed for causal discovery throughout this dissertation.

When trying to associate the statistical patterns with causal relations as discussed in this section, we might wonder about some possible contradiction between the temporal information implied by the directed edges and the actual ones. See Pearl [2000] for the discussion.

2.4 Parameter Learning

In the research domain of machine learning, numerous approaches are often classified into two categories. Probabilistic graphical models, which include Bayesian network models, are in generative models. Generative models are those that describe how the observed data can be generated via its structures and parameters. Therefore, in BNs, we want to estimate parameters that represent the underlying probability distributions in limited observational data.

Presuming that we are provided a finite training dataset $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N\}$ comprising *i.i.d.* (independent and identically distributed) samples from a distribution P, then the sample size can be represented as $|\mathcal{D}| = N$. We might not need to capture the whole limited data $\mathcal{D}$ exactly because of of the goal of avoiding overfitting to samples.

[1] Reichenbach derived those relations for binary random variables.

In this section, we present the essences of learning the parameters of BNs, where we assume that the data are complete, i.e., have no missing values, and that variables are all discrete throughout the thesis. We assume also that the structure of BNs has been provided or learned, which is denoted as $\mathbb{G}$. For learning parameters of BNs, two methods are often used: the maximum likelihood (ML) approach and Bayesian methods. First, we describe the ML approach and then, the more robust Bayesian approach.

2.4.1 Maximum Likelihood Estimation

The maximum likelihood (ML) estimation method is widely used in machine learning, which is based on the maximum likelihood principle. This principle states that we should select the model that generates the most observed data.

Definition 2.14 *The likelihood function $L(\mathcal{D}\,|\,\boldsymbol{\theta})$ is the probability of the i.i.d. instances of $\mathcal{D}$ given the parameter set $\boldsymbol{\theta}$:*

$$L(\mathcal{D}\,|\,\boldsymbol{\theta}) = P(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{m=1}^{N} P(\boldsymbol{d}_m\,|\,\boldsymbol{\theta}) \tag{2.9}$$

where $P(\boldsymbol{d_m}\,|\,\boldsymbol{\theta})$ is the probability of the m-th instance under the parameter set. The log-likelihood function is

$$LL(\mathcal{D}\,|\,\boldsymbol{\theta}) = \sum_{m=1}^{N} \log P(\boldsymbol{d}_m\,|\,\boldsymbol{\theta}). \tag{2.10}$$

Based on the ML approach, we should choose the parameters $\hat{\boldsymbol{\theta}}$ that maximize the log-likelihood function of $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\; LL(\mathcal{D}\,|\,\boldsymbol{\theta}). \tag{2.11}$$

For a BN, presuming a set of parameters $\boldsymbol{\theta} : \theta_{ijk}, 1 \leq i \leq n,\ 1 \leq j \leq q_i,\ 1 \leq k \leq r_i$ where i is an index of X, j is an index of parent nodes' configurations ($j \leq q_i$), r_i signifies the number of states of the X_i, and k stands for X_i's state ($k \leq r_i$). Then the likelihood can be expressed as [Cooper and Herskovits, 1992]

$$P(\mathcal{D}\,|\,\boldsymbol{\theta}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \tag{2.12}$$

where N_{ijk} represents the number of cases in the dataset in which $X_i = x_i^k$, given the condition that $Pa(X_i) = \boldsymbol{\pi}_i^j$ [Heckerman, 1995, revised June 1996].

Applying Lagrangian multipliers to the log likelihood function with multipliers to constrain the parameters to a normalized probability distribution, the constraints are as follows:

$$0 \leq \theta_{ijk} \leq 1, \qquad k = 1, 2, \ldots, r_i \tag{2.13}$$

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1 \tag{2.14}$$

for any $1 \leq i \leq n$, $1 \leq j \leq q_i$. We drop off the indices i and j because of global and local independencies of the parameters on BNs [Spiegelhalter and Lauritzen, 1990]. Let Lagrangian L as

$$L = \prod_{k=1}^{r} \theta_k^{N_k} + \lambda \left(1 - \sum_{k=1}^{r} \theta_k \right). \tag{2.15}$$

Using partial derivative $\partial L / \partial \theta_k$ and letting it equal zero yields

$$N_k \prod_{k=1}^{r} \theta_k = \lambda \theta_k \tag{2.16}$$

for any $1 \leq k \leq r$. Summing both sides with respect to k and applying the constraint eq. (2.14) to it,

$$\lambda = \prod_{k=1}^{r} \theta_k \sum_{k=1}^{r} N_k. \tag{2.17}$$

Therefore, the ML estimator is obtained as

$$\theta_k = \frac{N_k}{\sum_{k'=1}^{r} N_{k'}}. \tag{2.18}$$

The parameters of BNs with ML estimation are obtained by restoring the indices i and j and denoting $\hat{\theta}_{ijk}$ as

$$\hat{\theta}_{ijk} := P(x_i^k \mid \boldsymbol{\pi}_i^j) = \frac{N_{ijk}}{\sum_{k'=1}^{r_i} N_{ijk'}}. \tag{2.19}$$

The parameter tables are designated with respect to the indices i, j, and k as the conditional probability tables (CPTs).

2.4.2 Bayesian Estimation

Next, we describe the Bayesian estimation approach. When using Bayesian statistics on discrete random variables, the estimators are often obtained using a posterior mean or by maximizing a posterior, where the Dirichlet distributions and their hyperparameters (smoothing parameters) set $\boldsymbol{\alpha}$ are usually used. According to Bayesian statistics [Gelman et al., 2004], a posterior probability density function $\rho(\boldsymbol{\theta}|d)$ given data d is expressed as follows from Bayes' theorem: $\rho(\boldsymbol{\theta}|d) \propto \rho(\boldsymbol{\theta}) f(d|\boldsymbol{\theta})$, where $f(d|\boldsymbol{\theta})$ is the likelihood function and $\rho(\boldsymbol{\theta})$ is a prior distribution. In discrete variables, the likelihood function is the multinomial distribution function. Prior and posterior functions are both written as Dirichlet distributions when $\rho(\boldsymbol{\theta})$ and $\rho(\boldsymbol{\theta}|d)$ are both represented as *natural conjugate family distributions* of the likelihood function. The Dirichlet distribution function with parameters $\alpha_1, \alpha_2, \dots, \alpha_r$, where $\alpha_1, \alpha_2, \dots, \alpha_r > 0$, is

$$\rho(\theta_1, \theta_2, \dots, \theta_r) = \frac{\Gamma(\sum_{k=1}^{r} \alpha_k)}{\prod_{k=1}^{r} \Gamma(\alpha_k)} \prod_{k=1}^{r} \theta_k^{\alpha_k - 1}, \tag{2.20}$$

where $0 \leq \theta_k \leq 1$, $\sum_{k=1}^{r} \theta_k = 1$, and $\Gamma(x)$ is the Gamma function, which is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{2.21}$$

The posterior is also the Dirichlet density, which is

$$\rho(\theta_1, \theta_2, \dots, \theta_r \,|\, d) \propto \prod_{k=1}^{r} \theta_k^{\alpha_k + N_k - 1}, \tag{2.22}$$

where N_k is the total number of occurrences in a state k. In the description provided above, $\boldsymbol{\alpha}$ are the Dirichlet hyperparameters. For instance, we demonstrate the Beta distribution function, which is the special cases of the Dirichlet distribution function for a binary variable, as following:

$$\rho(\theta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1}, \tag{2.23}$$

where $0 \leq \theta \leq 1$, $\alpha_1 > 0$, $\alpha_2 > 0$, $\alpha = \alpha_1 + \alpha_2$. We refer to the beta density function as $Beta(\theta; \alpha_1, \alpha 2)$. Figure 2.4 presents some examples of the Beta distribution functions with various parameters, where $Beta(\theta; 1, 1)$ represents the uniform density function, In fact, $Beta(\theta; 0.5, 0.5)$ is equivalent to the Jeffrey's prior distributions for binomial distributions. It is noteworthy that the Beta distribution functions with $\alpha_1 < 1$ and $\alpha_2 < 1$ are not probability density functions because those diverge at both ends and then do not satisfy the integral condition as seen in Fig. 2.4. In general, the Dirichlet distributions with $\alpha_k < 1$ for all k in eq. (2.20) are also not probability density functions.

Those functions are applied to conditional probabilities for expression in BNs. Then the BN parameters are obtained by taking the posterior mean of the Dirichlet distributions as (for details, see Neapolitan [2004]):

$$\theta_{ijk} := P(x_i^k \,|\, \boldsymbol{\pi}_i^j) = \frac{\alpha_{ijk} + N_{ijk}}{\sum_{k'=1}^{r_i} (\alpha_{ijk'} + N_{ijk'})}. \tag{2.24}$$

As expressed in eq. (2.24), it is clear that the hyperparameters $\boldsymbol{\alpha}$ play a role in avoiding overfitting because $\boldsymbol{\alpha}$ can contribute to some extent when the data size is not

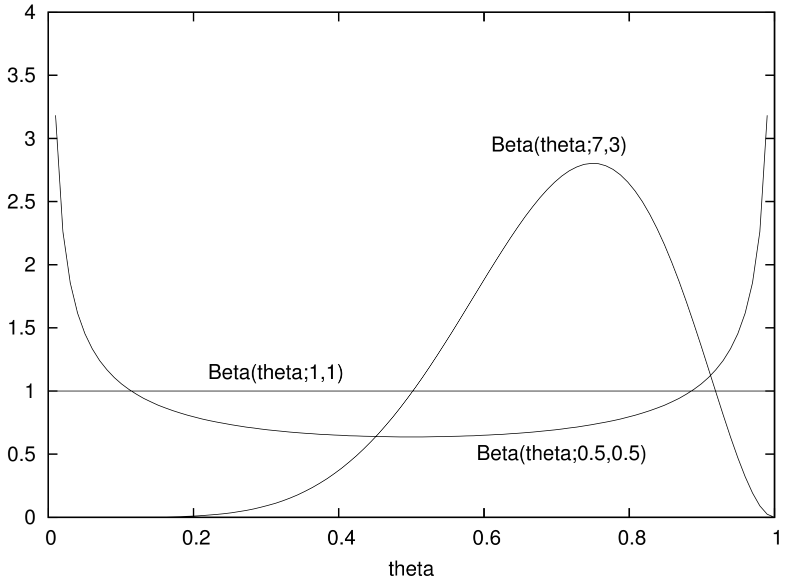


Figure 2.4: Some Beta distribution functions, which are special cases of the Dirichlet distribution functions for a binary variable

large. However, because of the lack of intelligible meaning of $\boldsymbol{\alpha}$ in a case without prior knowledge about the data, $\boldsymbol{\alpha}$ is often assigned manually to various values including $\alpha_{ijk} = \alpha = 1$, which represents uniform prior distributions [Cooper and Herskovits, 1992] and $\alpha_{ijk} = \alpha = 0.5$ [Clarke and Barron, 1994; Suzuki, 1996].

For learning BNs, Silander et al. [2007] reported that the optimal values of hyperparameters are highly sensitive to each dataset. It has not been found within Bayesian framework how to decide the optimal values using principled methods. For that reason, it has persisted as a critical issue.

This dissertation provides an alternative learning approach to the Bayesian approach, which has the similar effect of avoiding overfitting and which is not sensitive for selecting entailed hyperparameters.

2.5 Structure Learning

In this section, we describe the structure learning methods for BNs, where two main approaches exist: constraint-based and score & search based, and their hybrid approach also exists.

Learning BN structures means searching the best encoding for some criterion. If we search it for whole space of structures, then it has been shown that the size of the space $f(n)$ grows more than exponentially with the number of nodes [Jensen and Nielsen, 2007]:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)!\,n!} \, 2^{\,i(n-i)} f(n-1). \tag{2.25}$$

Therefore, to find the best BN (or DAG pattern) by considering all DAG patterns as computationally infeasible when the number of nodes is not small. In fact, Chickering et al. [2004] shows that learning BNs is an NP-hard problem for widely acceptable conditions. In addition, the consumed time grows in general with training samples because the dataset is scanned for each calling to calculate a score or a statistic.

2.5.1 Constraint-Based Structure Learning

Herein, we state a basic concept of the constraint-based structure learning methods using the notation presented above. The approach uses concepts and properties such as conditional independencies, d-separations and faithfulness conditions, which are described in the preceding sections.

Reportedly, Wermuth and Lauritzen [1983] began the use of conditional independence tests for constructing graphical models. The constraint-based structure learning approaches are often related to causal discovery because the important procedure of the approaches is finding *v-structures with marginal independencies*, which can be considered as causal patterns of the common effects based on the theory of Reichenbach [1956] described in 2.3.5. In fact, Rebane and Pearl [1987] specifically devoted attention to the adjacent triplets of three possible types, that is, serial, divergent, and convergent triplets, which Reichenbach [1956] suggested that those are important causal patterns. Additionally, they took particular note of the fact that the only convergent v-structures are distinguishable from other types by statistical tests. Then they developed a causal network recovery algorithm that uses this property and orients edges within the constraints of DAG as far as possible. From this work, the causal discovery tasks are launched. However, their algorithm assumed that the DAGs representing causation are limited to poly tree networks, which cannot allow any loop in the associated undirected graphs. Then Verma and Pearl [1990] and Spirtes et al. [1990] developed basic algorithms for recovering general DAG patterns in the context of causal inference, which is called the Inductive Causation (IC) algorithm and the SGS algorithm for each[2].

We describe the basic algorithm following Pearl [2000] as follows. For a BN that satisfies the faithfulness condition, the basic concepts are the following:

- Search for a set $\boldsymbol{Z}$ for each pair of variables X and Y in $\boldsymbol{V}$ such that $Ind\,(X;Y|\boldsymbol{Z})$ holds in P. Therefore, X and Y are conditionally independent given a set $\boldsymbol{Z}$ in P. Construct an undirected graph such that nodes X and Y are connected with an undirected edge if and only if no set $\boldsymbol{Z}$ can be found.

- For each pair of nonadjacent variables X and Y with a common neighbor W, check if $W \in \boldsymbol{Z}$. If not, then add arrowheads pointing at W (i.e. $X \to W \leftarrow Y$), the type of which is called a v-structure.

- Orient as many of the undirected edges as possible subject to two conditions: (i) the orientation should not create a new v-structure; and (ii) the orientation should not create a directed cycle graph.

Orientation Rules

In the last step of the basic algorithm presented above, Verma and Pearl [1992] provided the orientation rules after recognizing v-structures, as shown in Fig. 2.5 as follows:

Theorem 2.7 (Verma and Pearl [1992]) *Presuming that we have a hidden DAG structure and we have a partially oriented DAG $\mathbb{G}$ that is assumed to have Markov and faithfulness conditions and is oriented by recognizing v-structures. The following three rules can also facilitate edge orientation:*

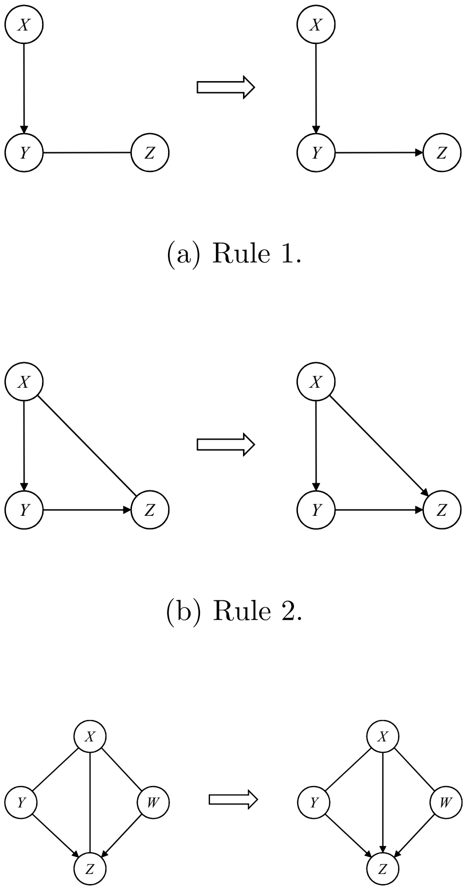
- *Rule1: In $\mathbb{G}$ for three nodes X, Y and Z, if a directed edge exists from X to Y, if an undirected edge links Y and Z, and if neither a directed or undirected edge exists between X and Z, then orient the edge from Y to Z.*

- *Rule2: In $\mathbb{G}$ for three nodes X, Y and Z, if two directed edges link X to Y and link Y to Z, and if an undirected edge exists between X and Z, then orient the edge from X to Z.*

- *Rule3: In $\mathbb{G}$ for four nodes X, Y, Z and W, if directed edges exist from Y and W to Z and if X are adjacent to Y, Z and W with undirected edges, then orient the edge from X to Z.*

[2] In another context, Fung and Crawford [1990] also developed an algorithm for constructing an undirected network structure based on conditional independences.

The proof is presented in Appendix A. The completeness of the rules was proved by Meek [1995a].



(a) Rule 1.

(b) Rule 2.

(c) Rule 3.

Figure 2.5: Orientation rules for DAG patterns.

PC Algorithm

After developing the basic algorithm (SGS), Spirtes and Glymour [1991] provided more sophisticated and efficient algorithm, called the *PC algorithm*, in which the conditional independence tests are conducted in order of the size of conditioning sets $|\boldsymbol{Z}|$, starting from the empty set. It has been a representative constraint-based algorithm because the order makes the tests finish in polynomial time. The algorithm is also used later. The PC, which constructs partial DAGs (PDAGs) that represent the Markov equivalent models, efficiently finds the conditional independences; that is the following algorithm:

```
1. Assume a non-negative integer $m = 0$.
2. Let $\mathbb{G}$ be a complete undirected graph.
3. Repeat:
  (a) For all pairs of variables $(X, Y)$, check $Ind(X, Y|\boldsymbol{Z})$ for all subsets $\boldsymbol{Z}$
      such that $|\boldsymbol{Z}| = m$ and $\boldsymbol{Z} \subset Adj(X)$ or $\boldsymbol{Z} \subset Adj(Y)$.
      If there exists a $\boldsymbol{Z}$ such as $Ind(X, Y|\boldsymbol{Z})$,
      then remove the edge $X - Y$ from $\mathbb{G}$, and add $\boldsymbol{Z}$ to $SepSet(XY)$.
  (b) Set $m = m + 1$.
      Until no variable has more than $m$ adjacencies,
      or a stopping condition is satisfied.
4. Orientation rules are performed.
5. Return the partially directed acyclic graph $\mathbb{G}$.
```

Therein, $|\boldsymbol{X}|$ denotes the size of members in $\boldsymbol{X}$; $Adj(X)$ is a set of adjacent nodes to X. The orientation rules [Verma and Pearl, 1992], described in step 4 of the algorithm, are as follows:

```
4-1. If $U \notin SepSet(XY)$, orient $X - U - Y$ as $X \rightarrow U \leftarrow Y$ (v-structure)
     for each uncoupled set of $X$ and $Y$ such as $X - U - Y$.
4-2. Repeat this step while more edges can be oriented.
4-2-1. Orient $U - Y$ as $U \rightarrow Y$ for each uncoupled set of $X$ and $Y$ such as
       $X \rightarrow U - Y$.
4-2-2. Orient $X - Y$ as $X \rightarrow Y$ for each set of $X$ and $Y$ such that a path exists
       from $X$ to $Y$.
4-2-3. Orient $U - W$ as $U \rightarrow W$ for each uncoupled set of $X$ and $Y$ such as
       $X - U - Y$, $X \rightarrow W$, $Y \rightarrow W$, and $U - W$.
```

The computational complexity of the algorithm is provided, which we express as $W(n)$ for the number of conditional independence tests required in the algorithm. Let n be the number of nodes in $\boldsymbol{V}$ and let k be the maximum size of adjacent nodes in the produced DAG pattern $\mathbb{G}p$. Then n choices exist for the value of X in the first for loop. Once X is chosen, then $n - 1$ choices exist for Y. For given values of X, Y, and i, we must check at most ${}_{n-2}C_i$ subsets for d-separating X and Y, where we designate combination number taking m from n as ${}_nC_m$. The value of i, which are considered, are at most 0–k.

Therefore, we have the bound as follows:

$$W(n) \leq n(n-1)\sum_{i=0}^{k} {}_{n-2}C_i \leq \frac{n^2(k+1)(n-2)^k}{k!}. \tag{2.26}$$

The algorithm is reasonably efficient if the DAG pattern is sparse (i.e. no node is adjacent to numerous other nodes).

Furthermore, the correctness of the SGS and the PC is proved.

Theorem 2.8 (Spirtes et al. [2000]) *If the input data to the SGS or PC algorithm is from a joint probability distribution P of the random variables in some set $\boldsymbol{V}$ and a DAG $\mathbb{G} = (\boldsymbol{V}, E)$ and $(\mathbb{G}, P)$ satisfies the faithfulness condition, and if the conditional independence relations are correctly detected, then each output is a pattern that represents the faithful Markov equivalent class of $\mathbb{G}$.*

The proof is presented in Appendix A.

Classical hypothesis testing such as χ^2 and G^2, for checking conditional independence, has been used frequently within BN learning algorithm under a faithfulness assumption [Spirtes et al., 2000; Tsamardinos et al., 2006]. It is also used herein later. Statistics such as χ^2 or G^2 are expressed here as S^2. If S^2 can be approximated to a χ^2 distribution with degrees of freedom *df*: χ^2_{df} , and if $S^2 < \chi^2_{\alpha,df}$, where $\chi^2_{\alpha,df}$ is a threshold value such that $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$, in which α is a fixed confidence level, then we do not reject the null hypothesis of (conditional) independence between two selected variables given selected conditional sets; otherwise we reject it. The validity of approximation of statistics such as χ^2 or G^2 is proved in asymptotic regions [Kullback, 1968]. However, it is not justified for a small sample size. Spirtes et al. [2000] used it in their PC algorithm, as a criterion for the validity: the algorithm does not perform an independence test if the sample size is less than 10 times the number of different possible joint patterns of the two variables and conditional sets, which means that the variables are assumed to be conditionally dependent. This impracticality is a weak point of the constraint-based learning methods of BNs because learning BNs often must process insufficient data.

Although few studies were done after the PC, another representative constraint-based algorithm is that called *Three Phase Dependency Algorithm* (TPDA) [Cheng et al., 1997, 2002], which is the algorithm that starts from constructing the Chow–Liu maximum spanning tree [Chow and Liu, 1968] networks. The name came from the fact that the algorithm has three phases (Chow–Liu, thickening, and thinning phases). Ramsey et al. [2006] improved PC algorithm by introducing ambiguity in deciding orientation of edges, which is related to our research because of their identical problem consciousness (see also Chapter 4).

This approach was soon developed by Verma and Pearl [1990] and Spirtes [1991] for detecting latent common causal variables. A research group at Carnegie Mellon University has developed latent variable models for over 10 years and then they extended DAG models to Partial Ancestral Graph (PAG) and Markov Ancestral Graph or Mixed Ancestral Graph (MAG) [Richardson and Spirtes, 2002; Ali et al., 2005; Zhang and Spirtes, 2005]. Their descriptions are omitted because we do not deal with the latent variable models in this thesis.

2.5.2 Score and Search Based Structure Learning

The other major approach is called the *score and search* approach. In general, this approach searches the best score structure by adding, removing, and reversing edges. One might say that the first trial in this approach was a Chow–Liu maximum spanning tree algorithm, where the score used mutual information existed between variables X and Y [Cover and Thomas, 2006]:

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}. \tag{2.27}$$

However, their algorithm generates undirected tree networks. Herskovits and Cooper [1990] used the maximum entropy principle[3] with Bayesian smoothing, which is apparently resembles our method only in using entropy (see, Chapter 4). In 1991, Cooper and Herskovits [1991] developed a greedy search algorithm called *K2* and used the Bayesian posterior function, which has dominated this approach to date. For discrete variables, they obtained a posterior probability distribution using Dirichlet distributions as

$$P(\mathcal{D} \,|\, \mathbb{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{2.28}$$

where they used uniform prior distributions (i.e. hyperparameter $\alpha_{ijk} = \alpha = 1$). It is called the Dirichlet prior score metric (DPSM). Buntine [1991] also assumed the Dirichlet prior and introduced a metric called BDeu, which is a special case of BDe (Bayesian Dirichlet equivalence) metric Heckerman et al. [1995] proposed. Their proposal is based on the likelihood equivalence assumption and they showed that the Dirichlet prior with the constant sum of the hyperparameters for a variable is a sufficient condition to satisfy the assumption. This hyperparameter is called the equivalent sample size (ESS).

Actually, Suzuki [1993] first applied the correct minimum description length (MDL) [Rissanen, 1978] principle to learning BNs, and proved that it is approximated from the

[3] In fact, they used scores as minimum entorpies by mistake, such as maximum likelihood, see chapter 5

DPSM with $\alpha = 0.5$. Other score metrics such as the Bayesian information criterion (BIC) [Schwarz, 1978] and Akaike information criterion (AIC) [Akaike, 1974] are also often used in approaches. For example we described the MDL score for BNs as

$$\begin{aligned}\text{Score[MDL]} &:= \log P(\mathcal{D} \mid \mathbb{G}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} - \frac{1}{2} \sum_{i=1}^{n} q_i (r_i - 1) \log N, \end{aligned} \tag{2.29}$$

where $N := \sum_{j=1}^{q_i} N_{ij}$ and q_i denotes the degree of freedom in the parent set of node i. It is noteworthy that BIC is approximately the same formula as that shown in eq. (2.29) and that AIC is the form of MDL replaced $\log N$ with 2.

One disadvantage in this approach is that it is not computationally efficient because it must search a large number of combinations of nodes and their parent nodes to obtain the best score structures. The K2 algorithm proposed by Cooper and Herskovits [1991] used a greedy approach. Many studies since then have adopted the greedy search. Suzuki [1996] first proposed a practical exact search algorithm using the branch and bound algorithm. Tian [2000] improved Suzuki's algorithm to be more computationally efficient. During the decade, many other scores and search methods have been proposed. For use with a few dozen nodes, some researchers proposed exact search methods [Koivisto and Sood, 2004; Silander and Myllymaki, 2006]. Chickering [2002] proposed Greedy Equivalent Search (GES) algorithm, for which he proved correctness in a large sample limit. Friedman et al. [1999] and Moore and Wong [2003] proposed fast and accurate search algorithms, respectively called Sparse Candidate (SC) and Optimal Reinsertion (OR) algorithms. Ueno [2008] recently proposed an optimal method of ESS in Dirichlet prior distributions using the empirical Bayesian approach. Furthermore, Tsamardinos et al. [2006] showed that the score and search methods, which have been advancing considerably as described above, are superior to constraint-based methods in their large scale experiments using PC, TPDA, GES, SC, OR, and their algorithms.

One advantage of score and search method is the ability of treating missing data, for which Friedman [1998] proposed a structural EM algorithm in which the golden standard EM algorithm [Dempster et al., 1977] is applied for learning BNs.

One disadvantage of these methods is treating latent variable models, especially latent common causes. However, Elidan et al. [2001] proposed a discovering latent variables method from perspective of network cardinality, which can find a succinct model as more statistical predictive models than causal models.

A current major interest in this approach is finding the optimal hyperparameters (ESS) of the prior distributions [Steck, 2008] because high sensitivity for those was found to be obvious in the optimal structure learning [Silander et al., 2007]. The result seems to explain the mystery of poor performance of learning BNs using MDL score, which Allen and Greiner [2000] reported: As described above, Suzuki showed that the MDL score for BNs is derived from the DPSM with a constant Dirichlet hyperparameter while the result of Silander et al. [2007] showed that the optimal Dirichlet hyperparameter is not a constant value, but is needed for each dataset. The author's problem consciousness is focused on the issue in this dissertation.

2.5.3 Hybrid Structure Learning

Singh and Valtorta [1993] first developed a hybrid algorithm both of constraint-based (PC) and score-search (K2) methods. Subsequently, Spirtes and Meek [1995] combined their PC algorithm and a greedy Bayesian pattern search (GBPS) algorithm for finding the highest score, which accomplished good results. Recently Tsamardinos et al. [2006] proposed an algorithm that they called the Max–Min Hill Climbing (MMHC) search algorithm, which is well-balanced for speed and accuracy. These studies showed the effective a strategy by which the constraint-based algorithms are used for obtaining initial patterns rapidly, which reduces the search spaces for accurate and slow score and search algorithms. They are used last.

2.5.4 Differences between the approaches

Next, differences between the constraint-based and score-search-based structure learning are considered. Constraint-based approaches have emphasized the graphical properties of independencies based on d-separations, while the score-search-based approaches on statistical properties based on the Bayesian estimations. The DPSM score in eq. (2.28) can be decomposed into the local score of a variable and its parent variables as

$$\text{score}(\mathcal{D}, X_i, PA(X_i)) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{2.30}$$

which means that the score-search approaches, in general, are intended to find the best combinations of a node and its parent nodes in view of statistical models. Because of its relativity, it seems difficult to have clues for finding latent common causal variables. However, the constraint-based approaches attend to find absolutely conditional independence relations between variables, which can find v-structures that have important roles of causality. Therefore, the two approaches should probably be used in different purposes: the score-search approach is suitable for finding statistical predictive models, and the constraint-based approach is suitable for finding causal models because of Reichenbach's philosophical foundation that associates causality with statistical patterns [Reichenbach, 1956; Pearl, 2000], although the difference vanishes when sufficient data are available under the assumption of no latent variables. In that sense, it does not seem suitable to use score–search methods or hybrid methods for finding gene regulatory networks, where the main purpose should be finding causal networks.

Additionally, it would be quite different to estimate statistical amounts such as scores or G^2 (or χ^2) statistics. The former is estimated using Bayesian method— the latter by classical statistics. Therefore, the overfitting issue is expected to reduce the learning ability of the constraint-based approaches, and should then be resolved to yield high performance using constraint-based approaches. However, the Bayesian method entails the difficulty of high sensitivity for the value of hyperparameters and of selecting their optimal values. A new approach is proposed for this issue in Chapter 4 after some preparations in Chapter 3.

Chapter 3

Parameter Learning using the MFE Principle

The salient goal of this dissertation is providing a novel framework of learning Bayesian networks for both parameters and structures that is effective for insufficient data. In this chapter, we first propose a new learning methodology using the minimum free energy (MFE) principle, which originates from thermodynamics and which has a relation to the second law of thermodynamics. The concept “Data Temperature” is introduced for using the MFE principle effectively, which is a key idea of the thesis. The framework described in this chapter plays a significant role in the progress of structure learning because parameter estimation is also needed for dealing with probability distributions, even for learning structures. This research was presented earlier by the author [Isozaki et al., 2008, 2009].

3.1 Introduction

As described in Chapter 2, maximum likelihood (ML) estimation is often used for estimating parameters of BNs, which means conditional probabilities in this thesis. However, when training data are few, the estimated parameters of BNs with ML are likely to fall into overfitting of the data. Bayesian methods, which involve prior distributions, are effective to avoid this problem. For expressing the prior distributions of discrete variables, the Dirichlet distribution function is usually used [Gelman et al., 2004]. This prior has hyperparameters, which can be interpreted as prior imaginary instances (we designate it as α for consistency in this dissertation). Some studies have used hyperparameters such as $\alpha = 1$ (meaning uniform prior distributions) [Cooper and Herskovits, 1992] or $\alpha = 0.5$ [Clarke and Barron, 1994; Suzuki, 1996]) when no prior knowledge exists.

However, it remains controversial to decide hyperparameters of prior distributions in theoretical perspectives, which are related to noninformative priors [Gelman et al., 2004; Robert, 2007]. From a practical perspective, Yang and Chang [2002] evaluated various hyperparameters in some simulation experiments for learning BNs and reported that $\alpha = 10$ is best. Therefore, it seems difficult to find optimal α consistently from both theoretical and practical perspectives. Additionally, it has been clarified as a critical issue to decide optimal hyperparameters because Silander et al. [2007] recently reported that optimal hyperparameters are quite sensitive for each dataset.

Another approach to avoid overfitting to data is incorporation of proper entropy into estimators of the parameters. One effective idea for treating entropy is using the principle of minimum (Helmholtz) free energy (MFE), which has its roots in statistical thermal physics [Kittel and Kroemer, 1980; Callen, 1985; Tazaki, in Japanese, 2000]. The free energy F, if described in the manner of physics, consists of (internal) energy U, entropy H, and (inverted) temperature β. In fact, β balances the contributions of U and H to the free energy.

In recent years, the MFE principle and similar concepts have been used in widely various areas of machine learning and data mining. Nevertheless, to our knowledge, the meaning of temperature has not been established yet. Consequently, β is treated in various ways at this stage: annealing parameters [Pereira et al., 1993; Ueda and Nakano, 1995], fixed parameters [Basu et al., 1998; Watanabe, 2001; LeCun and Huang, 2005; Yedidia et al., 2005], or optimizable parameters in each dataset [Hofmann, 1999]. The pre-fixed method apparently has a poor foundation and the optimization method using held-out data is not efficient in computational costs and is ineffective for very small data size. Consequently, no universal or robust value of β has been reported to date.

Differently from the approach described above, we take *model-based approaches* for β. For that purpose, a hyperparameter is introduced for β. Using this approach, we intend to explore robust estimation methods against the hyperparameter for BNs. As described in this dissertation, a meaning of β in the MFE principle is proposed by combining a physical quantity with a statistical quantity, and an explicit model of β as a result of our interpretation of the role of β. We assessed our method with respect to accuracies and robustness in relation to classification tasks using real world data.

This chapter presents an alternative method for estimating the proper parameters of BNs. The method uses the principle of minimum (Helmholtz) free energies (MFEs). The free energies with finite temperature are minimized for estimating proper parameters of BNs instead of maximizing likelihood or taking the expectation values of the Bayesian posterior distributions. For presenting the approach, entropy, energy, and temperature must be properly defined. Temperature is particularly regarded as important:

it determines the degree of contribution of entropy to the free energy.

3.2 Free energies

The (Helmholtz) free energies were introduced originally into the field of thermodynamics. The energies are defined such that a maximum thermodynamical work is the difference between values of free energies in two distinct states [Tazaki, in Japanese, 2000], where the maximum work is obtained by an isothermal quasistatic operation from a closed system under the condition of a constant temperature. Therefore, the free energy can be regarded as an amount, in a constant temperature, corresponding to a potential energy in dynamics (e.g., gravitational and electro-magnetic potential energy). In light of this meaning, the free energy is viewed as an amount that is extracted freely from a thermodynamical system.

In thermodynamics, the (Helmholtz) free energy F of a system is defined using internal energy U, entropy H, and the (inverted) temperature β_0 (= 1/temperature) as

$$F := U - \frac{H}{\beta_0}, \tag{3.1}$$

where (inverted) temperature β_0, which is a parameter, balances contributions of U and H to F. According to the principle of MFE, given some temperature β_0, the stable state of the system is realized to minimize F [Callen, 1985].

3.3 Minimum free energy principle on probability distributions

The principle of MFE is used for parameter learning. We denote random variables as X, which are assumed to be discrete variables, and internal states as x. For a definition of entropy terms, Shannon entropy [Cover and Thomas, 2006] is adopted, using probability distributions $P(X)$ as

$$H(X) := -\sum_{x} P(X = x) \log P(X = x). \tag{3.2}$$

The system variable is assumed as a probability distribution. We define U as the Kullback–Leibler (KL) divergence [Cover and Thomas, 2006], which represents the similarity or distortion between the hidden distribution and the distribution estimated using the ML method because we will incorporate the ML principle under some conditions.

Consequently, the internal energy is defined as

$$\begin{aligned} U(X) &:= D(P(X)\,||\,\hat{P}(X)) \\ &= \sum_x P(X=x)\log\frac{P(X=x)}{\hat{P}(X=x)}, \end{aligned} \tag{3.3}$$

where $\hat{P}(X)$ signifies the probability distribution estimated using the ML method and $P(X)$ stands for the hidden probability distribution. It is noteworthy that estimations obtained by minimizing the KL divergence in eq. (3.3) are equivalent to ML estimations because, for general distributions P and Q, $D(P||Q) \geq 0$ and $D(P||P) = 0$.

The hidden probability distribution parameterized by β_0 is the solution of the minimizing F with a constraint as $\sum_{X=x} P(X=x) = 1$. Therefore, it is solved using Lagrangian multipliers. The Lagrangian L is expressed as

$$\begin{aligned} L &= F + \lambda(\sum_x P(x) - 1) \\ &= \frac{1+\beta_0}{\beta_0}\sum_x P(x)\log P(x) - \sum_x P(x)\log\hat{P}(x) \\ &\quad +\lambda(\sum_x P(x) - 1) \\ &= \frac{1}{\beta}\sum_x P(x)\log P(x) - \sum_x P(x)\log\hat{P}(x) \\ &\quad +\lambda(\sum_x P(x) - 1), \end{aligned} \tag{3.4}$$

where λ is the Lagrange multiplier and we define a parameter β for later convenience, transformed from the β_0, as

$$\beta := \frac{\beta_0}{\beta_0 + 1}. \tag{3.5}$$

In relation to that expression, if $\beta_0 \to 0$, then $\beta \to 0$ (high temperature limit); if $\beta_0 \to \infty$, then $\beta \to 1$ (low temperature limit). We designate the β temperature later. Then the solution is derived from the partial derivative: $\partial L/\partial P(x) = 0$. Therefore, the estimated parameter $P_\beta(X)$ is expressed in the form of Boltzmann's law [Kittel and Kroemer, 1980], which is well known in statistical physics as

$$P_\beta(X=x) = \frac{\exp(-\beta(-\log\hat{P}(X=x)))}{\sum_{x'}\exp(-\beta(-\log\hat{P}(X=x')))}. \tag{3.6}$$

Practically the equivalent form is used as

$$P_\beta(X=x) = \frac{\hat{P}^\beta(X=x)}{\sum_{x'}\hat{P}^\beta(X=x')}, \tag{3.7}$$

where $\hat{P}$ represents the relative frequency: the ML estimator.

It is straightforward to extend the method described above to cases of multivariate systems by proper indexing for joint states. Therefore, Boltzmann's law, corresponding to eq. (3.6), becomes

$$P_\beta(\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(-\beta(-\log \hat{P}(\boldsymbol{X} = \boldsymbol{x})))}{\sum_{\boldsymbol{x}'} \exp(-\beta(-\log \hat{P}(\boldsymbol{X} = \boldsymbol{x}')))}, \tag{3.8}$$

where $\boldsymbol{X}$ is denoted as a multivariate set, and $\boldsymbol{x}$ is a joint state of $\boldsymbol{X}$.

These formulas (as eq. (3.6) or (3.7)) have been reported elsewhere, in works of Hofmann [1999] and Ueda and Nakano [1995]. However, the combination of the explicit definitions of U such as eq. (3.3) and the transformation in eq. (3.5) have not been reported in the literature. They can more easily lead to intuitive comprehension of the role of minimizing the free energy and temperature β in information science. Therefore, we can proceed to modeling β.

3.4 Introducing "Data Temperature"

From the definitions of U, H, and F provided above, it is apparent that parameter estimators by MFE tend to be ML estimators at low temperature (large β) and tend to be dominated by the entropy at high temperature (small β). On the other hand, from the perspective of data science, we hope to realize ML-like estimators for large samples and avoid overfitting for small samples. Therefore, temperature is related to the number of samples as follows. *Large* sample size corresponds to *low* temperature. *Small* sample size corresponds to *high* temperature. In other words, probabilistic fluctuation, which is large for small data size, is regarded as thermal fluctuation, which is large for high temperature, and vice versa in our approach. This concept is designated as the "Data Temperature". Then, we assume the following statement:

Assumption 3.1 ("Data Temperature") *"Data temperature" β is defined as $0 < \beta < 1$ and as a monotone increasing function of available data size.*

Based on this assumption of the relation described between data size and β, we can express β explicitly as some monotone function of the number of samples, which enables us to leverage the "Data Temperature" concept effectively. Although the exact mode of measuring β is left open, some clues for modeling β exist. First, β approaches 1 such that estimated probabilities approach those by ML when the data size is large, whereas β approaches 0 such that estimators are uniform for internal states when the data size is small. The larger the data size N is, the smaller the difference coefficient of β for N is. However, the smaller N is, the larger the difference coefficient. For that reason, a reasonable model would be a convex monotone increasing function that fulfills the boundary conditions described above. Next, the necessary data size is apparently dependent on the degrees of freedom of the random variables X. In other words, the more degrees of freedom the random variables have, the larger the data size would be needed to regard the estimators as near-ML estimators. Then, γ and N_c are introduced for separating effects of X's degrees of freedom from the β. Furthermore, γ is a function of the degrees of freedom, and N_c is a decoupling constant, which is introduced as a hyperparameter for β, and which is expected to play some role other than that related to γ.

Then, we create a model of β as a simple monotone function of data size N, $\tilde{N}$, γ, and N_c, as shown in the following:

$$\beta := 1 - \exp\left(-\frac{\tilde{N}}{N_c}\right), \tag{3.9}$$

$$\tilde{N} := \frac{N}{\gamma}, \tag{3.10}$$

where we adopt an *exponential decay* function that often appears in natural science. $\tilde{N}$ denotes averaged sample size per (effective) degree of freedom (γ), and plays an significant role in statistical science such as statistical model selection. Therefore, this model is a very simple one under the assumption of the exponential function with a parameter. Three examples of the proposed function are portrayed in Fig. 3.1, which are the cases in which γ is assumed to be 1 for simplicity and $N_c = 1, 2, 5$, where we can recognize that N_c denotes the decay rate of β.

According to the description given above, the function γ must necessarily be decided. The simplest form of γ is one's own degrees of freedom,

$$\gamma := |X| - 1, \tag{3.11}$$

where $|X|$ is denoted as a number of states of a random variable X. It is designated as the "linear-state model". However, we consider that this model might be an approximate model under the limit of uniform distributions over the internal states. In practice, fewer data are necessary than in the uniform distributions because data distributions have some bias. For that reason, we consider another model of γ that is denoted with *effective* degrees of freedom, which can be expressed, in light of the explanation given above, as the following:

$$\gamma := \log(|X|). \tag{3.12}$$

This form of γ is an approximate expression of the effective degrees of freedom. The expression in eq. (3.12) is denoted as a "log-state model". These parameter learning

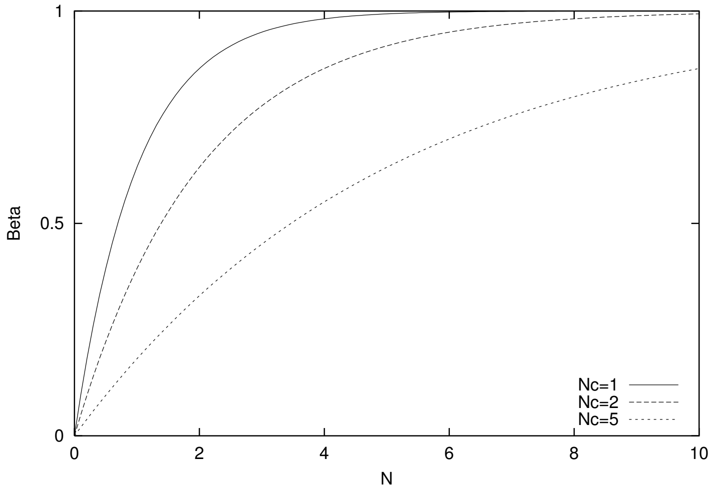


Figure 3.1: Examples of the proposed exponential function. $\gamma = 1$ and $N_c = 1, 2, 5$.

methods are called MFE with explicit β (MFE–EB) methods.

The relation between the temperature and data size can provide a perspective to unify the maximum likelihood (ML) and the maximum entropy (ME) principles under the MFE principle with varying data size because the eq. (3.6) has the same form of the ME principle because β can be regarded as an associated constraint condition.

It is important to refer to the relation between KL divergence and the MFE principle. The MFE principle can be regarded as an extension of minimizing KL divergences by defining a *tempered KL divergence* denoted as $D_\beta(P \,||\, Q)$, which is defined as

$$D_\beta(P(X) \,||\, Q(X)) := \sum_x P(x) \log \frac{P(x)^{1/\beta}}{Q(x)}. \tag{3.13}$$

Therefore, a free energy F can be expressed as a distribution $P(x)$, which should be estimated, and a probability function estimated by ML, which is designated by $\hat{P}(x)$ as follows:

$$F = D_\beta(P(X) \,||\, \hat{P}(X)) = \sum_x P(x) \log \frac{P(x)^{1/\beta}}{\hat{P}(x)}. \tag{3.14}$$

Consequently, adopting the MFE principle for statistical estimation, the preferred distributions have added extra entropies to the ML distributions according to "Data Temperature" (available data size) under non-zero and finite temperature: $0 < \beta < 1$, where, if $\beta \to 1$, then the *tempered KL divergence* converges to the KL divergence.

In closing this subsection, we can comment on the meaning of using the MFE principle in information sciences. In analyzing data, the free energy can be regarded similarly with the view used for thermodynamical systems: as an amount that is extracted freely from a data system under a given data size (temperature). This property is apparently very much preferred for inference, learning, and estimation of various kinds under a finite available data size because we wish to obtain maximum effective information from limited exploitable data.

3.4.1 Estimating conditional probabilities

In a BN that has discrete variables, conditional probability tables are often assumed to be independent in each conditioning event [Spiegelhalter and Lauritzen, 1990]. Using this local independent assumption, we naturally extend the form of β to local forms, which we attach to each node and configuration of its parent set. Consequently, in BNs, the free energy is defined in each node and configuration. Therefore, more detailed control of entropy is possible in conditional probabilities than in multivariate joint probabilities.

In fact, N_{ij} is defined as $N_{ij} := \sum_{k'} N_{ijk'}$ if the same indices i, j, k and notation N_{ijk} described in Section 2 are used. Furthermore, β_{ij} is definable in an exponential function as

$$\beta_{ij} = 1 - \exp\left(-\frac{N_{ij}}{\gamma_i N_c}\right), \tag{3.15}$$

where the “linear-state model” can be adopted as

$$\gamma_i := |X_i| - 1, \tag{3.16}$$

or the “log-state model”, as

$$\gamma_i := \log(|X_i|). \tag{3.17}$$

Finally, the parameters of BNs, θ_{ijk}, are expressed as the following.

$$\theta_{ijk} = \frac{\exp(-\beta_{ij}\ |MLL_{ijk}\ |)}{\sum_{k'} \exp(-\beta_{ij}\ |MLL_{ijk'}\ |)} = \frac{\hat{\theta}^{\beta}_{ijk}}{\sum_{k'} \hat{\theta}^{\beta}_{ijk'}} \tag{3.18}$$

Therein, MLL_{ijk} is defined as an expression using ML estimators $\hat{\theta}_{ijk}$: $MLL_{ijk} = \log \hat{\theta}_{ijk} \leq 0$.

3.5 Relation to Dirichlet hyperparameters

The MFE–EB method is an alternative to the Bayesian method. The two methods share some mutual relations. Here, the Bayesian Dirichlet hyperparameters can be derived from the MFE principle given some temperature by using eq. (2.24) and eq. (3.7) as follows.

$$\frac{\alpha_k + N_k}{\sum_{k'=1}^{r}(\alpha_{k'} + N_{k'})} = \frac{N_k^{\beta}}{\sum_{k'=1}^{r} N_{k'}^{\beta}} \tag{3.19}$$

Therefrom, we omit the indices i, j for simplicity. Consequently, we obtain a transformation formula between Bayesian Dirichlet hyperparameters and a temperature as the following compact representation of r-dimensional simultaneous equations:

$$\alpha_k = \frac{N_k^{\beta}(\alpha + N) - N_k N(\beta)}{N(\beta)}, \tag{3.20}$$

where we designate $\alpha := \sum_{k=1}^{r} \alpha_k$, $N := \sum_{k=1}^{r} N_k$, and $N(\beta) := \sum_{k=1}^{r} N_k^{\beta}$, and which can be generally solved using numerical methods. Therefore, despite a lack of prior knowledge, the obtained Dirichlet hyperparameters are dependent on the internal state, which is denoted as k, when we adopt the estimator that is expressed in the Boltzmann formula parameterized with temperature. This formula asserts that the MFE principle gives Dirichlet hyperparameters that are dependent on the internal state and available data under a given temperature. In other words, this formula indicates to us that the MFE principle gives Dirichlet prior distributions that vary with available data if some temperature is given. Additionally, N_c can be regarded in "Data Temperature" model as a *hyper-hyperparameter* for estimating the parameters of BNs. Consequently, a deeper hierarchical structure of parameters of BNs is assumed in the "Data Temperature" assumption, although it has not been introduced into Dirichlet hyperparameters when no prior knowledge is available.

3.6 Experiments

The experiments described in this section are undertaken to investigate whether the MFE–EB method can avoid overfitting in practice to the same degree that the Bayesian method can when the available data are few. Furthermore, the robustness of our method is examined against values of hyperparameters because it is important in practical use, especially when it is difficult to search optimal hyperparameters. For this purpose, UCI repository data [Newman et al., 1998] and the classification accuracy of Bayesian network classifiers (BNCs) are used along with a pre-trained network structure to evaluate the parameter estimation accuracy because the classification accuracy depends on parameter estimation accuracy in such situations.

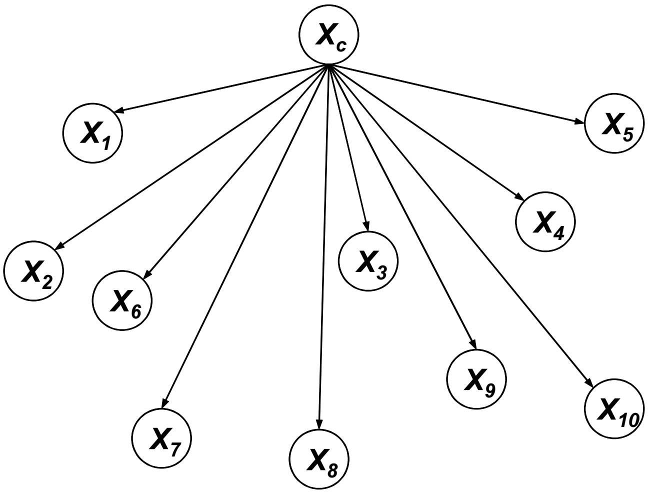


Figure 3.2: Example of a Näive Bayes classifier.

3.6.1 Bayesian network classifiers

In fact, BNCs are restricted models of BNs, which are often used for classification tasks. The BNCs have one class variable in addition to other variables, and predict the class label with given information related to the other attributes. The most famous models among BNCs are Näive Bayes classifiers [Langley et al., 1992], which are assumed to be conditionally independent of each attribute X_i and X_j given class label X_c, denoted as $X_i \perp\!\!\!\perp X_j \,|\, X_c$. A sample of Näive Bayes classifiers is presented in Fig. 3.2, where X_c denotes a class node and X_i $(i = 1, 2, \ldots, 10)$ are other attributes. Friedman et al. [1997] observed that, in many benchmark datasets, unrestricted BNs underperform Näive Bayes classifiers, Näive Bayes type models are used for evaluation of our method. This work adopted generally augmented Näive Bayes classifiers (GANs), which have an unrestricted network in attributes, except for class variables [Cheng and Greiner, 1999; Friedman et al., 1997]. Networks of this type are useful because they reportedly achieve the highest accuracies in many datasets among NBs, BNs, and GANs [Cheng and Greiner, 1999]. A network of this type is shown in Fig. 3.3 as an example, where X_c denotes a class node and X_i $(i = 1, 2, \ldots, 10)$ are other attributes. Furthermore, an unrestricted DAG structure is introduced among the attributes $(\{X_1, X_2, \ldots, X_{10}\})$.

For executing structural learning of GANs, the PC algorithm [Spirtes et al., 2000] was used, which is described in Chapter 2 for structure learning of BNs, and which usually uses hypothesis tests or mutual information tests for identifying conditional independence relations among variables (see Chapter 2). The PC algorithm was modified for application to BNCs according to the methodologies of Cheng and Greiner [1999] as follows (we designate whole variables $\boldsymbol{V}$, a set of variables $\boldsymbol{Z}$ and a variable X or Y):

- replacing mutual information between attributes $X, Y \in \boldsymbol{V}$: $I(X;Y)$ with a conditional mutual information $I(X;Y|X_c)$, where the mutual information is defined as [Cover and Thomas, 2006]

  $$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}, \tag{3.21}$$

  and the conditional mutual information as [Cover and Thomas, 2006]:

  $$I(X;Y|\boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} P(x,y,\boldsymbol{z}) \log \frac{P(x,y \mid \boldsymbol{z})}{P(x \mid \boldsymbol{z})P(y \mid \boldsymbol{z})}. \tag{3.22}$$

- replacing every conditional mutual information test $I(X;Y|\boldsymbol{Z})$ with $I(X;Y|\{\boldsymbol{Z}, X_c\})$, where $\boldsymbol{Z} \subset \boldsymbol{V} \setminus X_c$,
- and adding the class node X_c as a parent of every other attribute,

where the conditional independence of X and Y given subsets $\boldsymbol{Z}$ is measured using the conditional mutual information [Cheng et al., 2002] in our experiments. Actually, X and Y are conditionally independent given that the condition set $\boldsymbol{Z}$ if $I(X;Y \mid \boldsymbol{Z})$ is smaller than a certain threshold value $\epsilon > 0$, as Cheng and Greiner [1999] did for BNC. The threshold values are decided respectively for each dataset by limiting each network to a maximum of five parents per variable, which we consider is suitable for constructing predictive models such as BNCs. In addition, the graphs produced by the PC algorithm are partially directed acyclic graphs (PDAGs). Therefore, we oriented the undirected edges to avoid cyclic graphs and to reduce the number of parameters.

3.6.2 Evaluation using UCI data

Seven datasets were selected from the UCI machine learning repository. These datasets were formatted and discretized by Greiner[1] using the Fayyad and Irani method [Fayyad and Irani, 1993], except for the Car and Nursery datasets, which were pre-discretized. Datasets with numerous cases were selected to train the structure of GANs as precisely as possible because the effect of parameter estimation is expected to be separate from that of the correctness of structural inference. The structures might be designed by prior knowledge but the parameters are not pre-estimated in real applications such as user modeling (e.g., [Isozaki et al., 2005]). A brief description of the datasets is presented in Table 3.1[2].

[1] available from his site: http://www.cs.ualberta.ca/ greiner/ELR/

[2] Car dataset was split into 1000 training samples and 500 test samples using random selection.

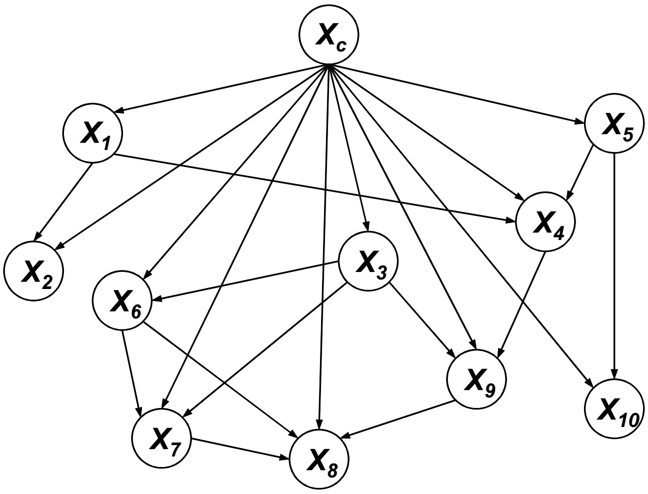


Figure 3.3: Example of a generally augmented Näive Bayes classifier (GAN).

Table 3.1: Description of datasets used for experiments.

	Dataset	Attributes	Classes	Instances	
				Train	Test
1	Car	6	4	1000	500
2	Chess	36	2	2130	1066
3	Letter	16	26	15000	5000
4	Nursery	8	5	8640	4320
5	Satimage	36	6	4435	2000
6	Segment	19	7	1540	770
7	Shuttle-small	9	7	3866	1934

First, structure learning was conducted using the PC algorithm. Next, parameter learning was conducted using the ML with eq. (2.19), Bayesian with eq. (2.24), and MFE–EB with eq. (3.15), eq. (3.18) and eq. (3.16) or eq. (3.17) processes. In contrast to structure learning, small samples that had been selected randomly from each dataset were used. Those sample sizes are selected to present large deviations of accuracy in some Bayesian hyperparameters from that in ML under the limit sample size of 100 (Letter, 1000; Chess/Nursery/Satimage, 250; Car/Segment/Shuttle-small, 100). The Dirichlet hyperparameters are the same, $\alpha_{ijk} = \alpha$, because of the lack of prior knowledge about datasets. Their accuracy was compared to that of ML to confirm the effectiveness of the Bayes and the MFE–EB. To avoid zero probability, which generates a contradiction when testing data have evidence that has not emerged in training data, a small positive

Table 3.2: Accuracies [%] of respective methods. We denote that *Bayes (α)* means that the value of a hyperparameter equals α in this table.

Dataset	ML	Bayes (0.5)	Bayes (1.0)	Bayes (10.0)	MFE–EB (lin)	MFE–EB (log)
Car	65.8	74.4	73.8	70.2	74.0	74.0
Chess	86.0	88.6	88.9	86.1	86.3	86.0
Letter	52.4	62.2	60.6	52.0	59.8	60.6
Nursery	66.2	75.3	76.5	77.5	76.7	76.7
Satimage	67.1	64.2	60.5	52.7	76.6	76.8
Segment	81.4	80.2	79.2	69.7	82.4	82.4
Shuttle-small	85.8	97.9	97.9	86.6	97.2	97.4
Ave.$\pm$ σ	72.1$\pm$11.7	77.6$\pm$11.8	76.8$\pm$12.7	70.7$\pm$13.2	79.0$\pm$10.7	79.1$\pm$10.5

number (0.0001) was added to all conditional frequencies.

For this study, we adopted $\alpha = 0.5, 1, 10$ because $\alpha = 1$ is the famous Laplace method; it means a uniform distribution [Cooper and Herskovits, 1992], and $\alpha = 0.5$ and $\alpha = 10$ are recommended from theoretical [Suzuki, 1996] and practical [Yang and Chang, 2002] perspectives. On the other hand, we examined some values for N_c because no knowledge exists for them. Accuracies of BNCs are presented with parameters estimated using ML, the Bayesian, and maximum values of the MFE–EB ("linear-state" and "log-state"), as portrayed in Table 3.2. The MFE–EB methods have effects of avoiding overfitting to small data size as the Bayesian does, because of the control of entropy according to the available data size. Moreover, regarding comparison with the Bayesian with recommended hyperparameters and with MFE–EB, the latter is superior with respect to maximum values of accuracy and variances. It seems to leverage likelihood and entropy more effectively than the Bayesian–Dirichlet method.

Next, we evaluate the robustness of the MFE–EB against various values of the hyperparameter N_c. Figs. 3.4 and 3.5 show the accuracies against the various values of N_c, which are both shown in differences of accuracies from that by ML. Actually, MFE–EB apparently shows good performance in common $N_c \sim 1$ for a linear model, and in common $N_c \sim 2$ for a log model in every dataset. For a more precise description, Table 3.3 presents the effective range over which the classification accuracy is greater than 95% of the maximum accuracy for respective datasets, where "*" signifies minimum or maximum values of hyperparameters in the range of the experiments. In both the linear-state and log-state models, the range in which good performance is shown has some overlap among all datasets. It can be said that the hyperparameter N_c has good common ranges of $(1 \leq N_c \leq 1.5)$ in the linear-state MFE–EB, and ranges of $(2 \leq N_c \leq 4)$ in the log-state MFE–EB. Therefore, it can be said that MFE–EB is not sensitive for selecting values of the entailed hyperparameter. These results are expected to result from the adopted functional form of β, where β approaches 1 rapidly with an increase in the

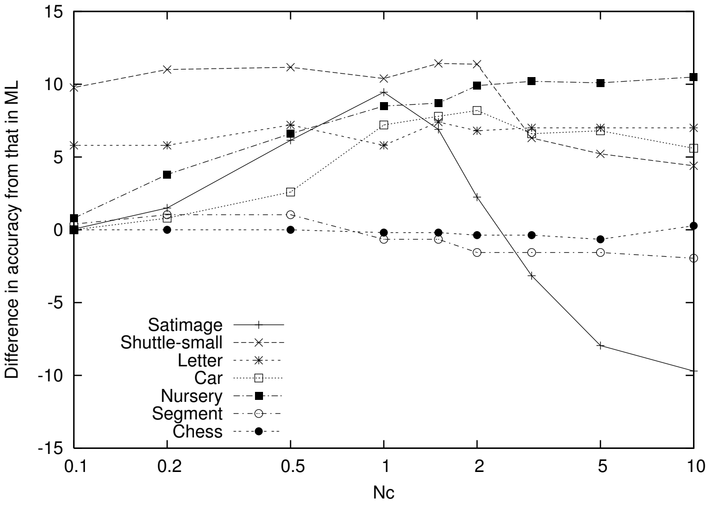


Figure 3.4: MFE–EB ("linear-state") estimation: differences in accuracy from ML [%].

Table 3.3: Effective ranges of hyperparameters.

Dataset	Linear-MFE (N_c)		Log-MFE (N_c)	
	min	max	min	max
Car	1.0	10*	2.0	10*
Chess	0.1*	10*	0.5*	10*
Letter	0.1*	10*	0.5*	10*
Nursery	1.0	10*	2.0	10*
Satimage	0.5	1.5	2.0	5.0
Segment	0.1*	10*	0.5*	10*
Shuttle-small	0.1*	2.0	0.5*	4.0

number of samples. Furthermore, the MFE–EB might be expected to have universal ranges of hyperparameters. Moreover, our MFE–EB method is apparently attractive in the sense that there is room for improvement of the function of β.

3.7 Discussion

The ML, the Bayesian, and the MFE–EB are all called generative methods, although discriminative methods have recently received significant attention in parameter learning of BNCs [Greiner and Zhou, 2002; Shen et al., 2003; Grossman and Domingos, 2004; Jing et al., 2005]. Their approaches are aimed at improving the classification accuracies

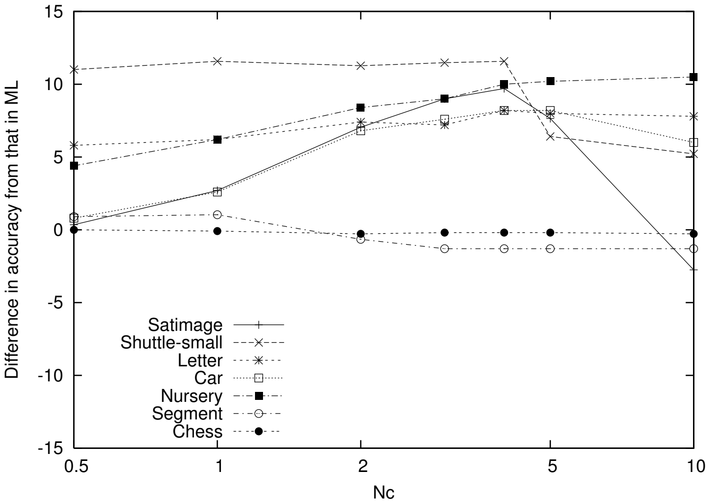


Figure 3.5: MFE–EB (“log-state”) estimation: differences in accuracy from ML [%].

of BNCs given some restricted structures. They are not intended to estimate hidden true probability distributions correctly. However, their studies suggest some insights about both the structure and parameter-learning of BNCs. Jing et al. [2005] found that, when the structure is incorrect, their discriminative methods outperform their generative counterparts. Shen et al. [2003] showed that, the better the structures are, the smaller the advantage of their discriminative method over the ML in classification tasks. Their results imply that the parameters are trained to compensate incompleteness of structure in BNs (BNCs) in discriminative methods. Therefore, we consider that the hyperparameters in the generative methods can have equivalent effects. Table 3.2 shows that the accuracies trained using the Bayesian method, are apparently slightly superior to those using the MFE–EB methods for some datasets. As might be expected, the opposite is true for some other datasets. We consider that the results are attributable to the incompleteness of structure in BNCs. In such situations, the Bayes estimators can compensate for incompleteness because α contributes to the parameters to some degree, even in situations with not a few data. However, the MFE–EB estimators less compensate because they are closer to ML estimators than the Bayes because of the functional form of β, which we assumed. Therefore, the MFE–EB method might be less effective in multivariate systems that are not properly inferred for their structures between variables, although the method is shown in this chapter to be extended theoretically to the case of discrete joint probability estimations.

It is worthwhile to discuss the possibility of improving the MFE–EB method. The change of accuracy over the hyperparameters presents similar behaviors derived using the two models of γ as shown in Fig. 3.4 and 3.5. In the Nursery dataset for example, values of N_c in both models, for which the accuracies are high, are larger than those in the other datasets. In contrast, in the Satimage dataset, both are smaller than those in the other datasets. These results imply that the optimal ranges of values in hyperparameters depend on the dataset properties. Therefore, it is apparently possible to improve MFE–EB by incorporating those properties of each dataset into the function of β.

3.8 Summary

A new parameter learning method of BNs is explored because Bayesian–Dirichlet methods, which are broadly accepted methods, have high sensitivity for selecting their hyperparameters and difficulty in deciding the optimal values. We propose an alternative method based on the principle of minimum free energy (MFE), which is well known in thermodynamics and statistical dynamics.

Our main conceptual contribution is the proposition of a "Data Temperature" assumption, which is generated by combining thermal fluctuation with probabilistic fluctuation. Our explicit model of the "Data Temperature" is assumed to have monotonic functions according to the available data size. The approach enables treatment of the two major principles of maximum likelihood and maximum entropy in a unified manner in the MFE principle with varying data size. In addition, this approach is an attempt at extracting maximum information under given finite samples by translating the concepts in thermodynamics of extracting maximum works under given finite temperature, using the MFE principle. This consciousness is preferable for inference, learning, estimation, and mining of various kinds because researchers in those domains wish to obtain maximum effective information from limited exploitable data.

Our method is superior to Bayesian–Dirichlet methods with recommended Dirichlet hyperparameters, although our explicit model of temperature is not sophisticated. Furthermore, it is not sensitive in classification accuracy for a choice of hyperparameters, unlike Bayesian–Dirichlet, which is attractive for practical use. Consequently, our method provides an effective tool for use as a parameter estimation method, especially for a small data size or for sparse data.

Chapter 4

Constraint-Based Structure Learning using MFE Principle

The main results on the dissertation in this chapter are presented here. As described in Chapter 1, we attempted to investigate the causal discovery methods effective in practical view. Constraint-based approaches for causal discovery were adopted because the approaches need not attach partial order in nodes because of high computational efficiency and having Reichenbach's theoretical foundation that associates causal patterns with conditional independencies, as described in Chapter 2. However, for improving the accuracy in practical sense, we can no more assume the validity of statistical testing, which other studies for constraint-based learning often assume [Spirtes et al., 2000; Cooper, 1997; Tsamardinos et al., 2006]. In fact, this learning approach often suffers from overfitting because of insufficient samples: the tests are based on an ML estimation approach, which is based on the *frequentism*. We consider that to be one reason why the approaches are inferior to score and search methods [Tsamardinos et al., 2006]. Therefore, in this chapter, we propose a new conditional independence testing method that is designed to be especially effective for small data size, and which is designed to be connected asymptotically with classical hypothesis testing. We use our methodology developed in the previous chapter, and unify learning methods of parameters and structures in a manner of MFE principle under a "Data Temperature" assumption. The outcomes are partially presented in an earlier paper [Isozaki and Ueno, 2009].

4.1 Introduction

4.1.1 Why is the MFE Principle Needed?

Many studies of learning BNs have often used mutual information (e.g., [Friedman et al., 1997]) for measuring dependence, which often means *minimizing entropy*, as described below. For asymptotic regions, in which sufficiently large samples are available, the guiding principle in statistics is the maximum likelihood (ML) principle [Lehmann, 1986]. Friedman et al. [1997] derived that, for a BN $\mathbb{G}$, given a dataset $\mathcal{D}$ with N data size for n random variables, maximizing the log likelihood $LL(\mathcal{D} \,|\, \mathbb{G})$ is equivalent to maximizing empirical mutual information between a node and its parent nodes (represented as Π_i for a node X_i):

$$LL(\mathcal{D} \,|\, \mathbb{G}) = N(\sum_{i=1}^{n} \hat{I}(X_i; \Pi_i) - \sum_{i=1}^{n} \hat{H}(X_i)),$$

where $\hat{I}$ and $\hat{H}$ denotes empirical mutual information and Shannon entropy [Cover and Thomas, 2006], and the second term of the right-hand-side of the equation has nothing to do with the learning structure. Therefore, from the definition of mutual information, it is readily derived that

$$LL(\mathcal{D} \,|\, \mathbb{G}) = -N \sum_{i=1}^{n} \hat{H}(X_i \,|\, \Pi_i) = -N \, \hat{H}(X_1, \ldots, X_n) \ . \tag{4.1}$$

The last equation is derived from the definition of BNs described in eq. (2.8). This equation shows that maximizing the log likelihood is equivalent to minimizing the entropy of BNs. This equation also implies that maximizing the log likelihood for constructing the DAG structures engenders *complete DAG* because the following inequality is justified [Cover and Thomas, 2006]:

$$-N \sum_{i} \hat{H}(X_i \,|\, \Pi_i) \geq -N \sum_{i} \hat{H}(X_i), \tag{4.2}$$

because

$$0 \leq H(X \,|\, Y) \leq H(X). \tag{4.3}$$

In contrast, when we obtain insufficient data, it is reasonable to use the maximum entropy (ME) principle [Cover and Thomas, 2006], which states that the most preferred probabilistic model should maximize its entropies under some constraint related to available data. Consequently, with no constraint, maximizing entropies of BNs engenders the DAG with *no edges*, which means that a BN is a collection of complete independent distributions: $P(\boldsymbol{X}) = \prod_i P(X_i)$.

A tradeoff exists between maximum likelihood and maximum entropy for obtaining the valid structures, which is similar to the case of parameter learning as described in Chapter 3. In the asymptotic region, the ML principle is expected to be dominant; in an insufficient sample region, the ME principle is expected to be dominant. Therefore, setting of a problem of how to decide the tradeoff between the ML and ME principles can be done according to an arbitrarily given sample size. The situation is seen as a metaphor of thermodynamics even here. The tradeoff between minimizing internal energy and maximizing entropy in thermodynamics apparently corresponds to the tradeoff between maximizing likelihood and entropy in statistics; and temperature can be regarded as a parameter that brings harmony of the two amounts. We can deal with these amounts in a free energy, and their tradeoff in the minimum free energy (MFE) principle.

4.1.2 Bayesian Approaches and MFE Principle

During recent decades, many researchers investigated Bayesian methods, which can avoid overfitting derived from using the ML with insufficient data, and which can be regarded as the same problem setting described in this dissertation. For example, Dash and Druzdzel [2003] proposed a robust conditional independence testing procedure using pseudo-Bayesian–Dirichlet smoothing. However, the Bayesian method presents the difficulty of deciding optimal hyperparameters simultaneously in both theoretical (related to noninformative priors) [Gelman et al., 2004; Robert, 2007] and practical [Yang and Chang, 2002] perspectives, when no prior knowledge exists, which is also described in Chapter 3. Furthermore, learned structures are highly sensitive to selection of the hyperparameters [Silander et al., 2007]. Although Steck [2008] proposed a solution of their optimal values in BDeu score metric, the method is applicable for data that are not small, and has inconsistency because the method is derived by AIC [Akaike, 1974], which is a different score metric from the BDeu for which his method is proposed. Therefore, we propose a new structure learning methodology for avoid the issue in Bayesian–Dirichlet approaches by developing our parameter learning method that is not sensitive for selecting hyperparameters, as described in Chapter 3.

In the use of MFE principles for statistical science, temperature is an unknown parameter in the MFE principle, which is in the similar situation of Bayesian–Dirichlet hyperparameters. However, we presented a model of inverted temperature that is a monotonic increasing function of available data size, as the “Data Temperature” assumption introduced in the preceding chapter. This approach can also be useful in structure learning BNs for estimating optimal entropies of the network structure. To realize this, remaining problems are how to define amounts corresponding to energies, entropies, and temperatures for constraint-based structure learning.

4.2 Representation of Free Energy in Probabilistic Models

Different from usual applications of the MFE principle in data science, we start with a description of the free energy definitely as a function of internal energy, entropy, and temperature to recognize important properties of temperature and use effectively free energies clearly. Fortunately, entropy was introduced into information theory by Shannon. It has since become a fundamental concept of computer science and statistical science [Cover and Thomas, 2006]. Therefore, we define the entropy of a random variable X as Shannon entropy. The entropy is intended to avoid overfitting for small samples. Kullback–Leibler (KL) divergence is adopted between two probabilistic distributions, which are an empirical distribution and the optimal distribution in view point of MFE. Here, the "Data Temperature" assumption is followed, which makes the MFE principle express a harmony between the ML and ME principle according to the available data size: *temperature is defined as a monotonic function of the available data size such that temperature $\beta_0 \to \infty$ if data size $N \to \infty$, and $\beta_0 \to 0$ if $N \to 0$.*

4.3 An MFE Representation of Hypothesis Testing on BNs

Conditional independence tests are represented using the MFE principle for constraint-based learning BNs. To do so, as in the usual manner [Spirtes et al., 2000; Tsamardinos et al., 2006], we represent the null hypothesis as conditional independent relations, and the opposite hypothesis as conditional dependent relations between two variables X and Y given conditional sets $\boldsymbol{Z}$.

The internal energies are defined for each hypothesis. First, we represent the internal energy U such that the relative entropy (KL divergence) between the graphs expressing the null hypothesis (expressed as $\mathbb{H}_1$, corresponding distributions as $\hat{P}_1$) and the true graphs (corresponding as P_0), where $\hat{P}_1$ of the null hypothesis is defined as a distribution estimated using the ML method. Therefore, an internal energy U_1 can be defined which expresses the null hypothesis such as

$$\begin{aligned} U_1(X,Y,\boldsymbol{Z}) &:= -D(\,\hat{P}_1(X,Y,\boldsymbol{Z})\,||\,P_0(X,Y,\boldsymbol{Z})\,) \\ &= \sum_{x,y,\boldsymbol{z}} \hat{P}(x,y,\boldsymbol{z}) \log \frac{P(x,y\,|\,\boldsymbol{z})}{\hat{P}(x\,|\,\boldsymbol{z})\hat{P}(y\,|\,\boldsymbol{z})}\ , \end{aligned} \tag{4.4}$$

where $\hat{P}$ is a maximum likelihood distribution and P is a distribution that will be estimated using the MFE principle with a "Data Temperature" model. In turn, the internal energy U_2 expresses a part of the opposite hypothesis (denoted as $\mathbb{H}_2$), which expresses a dependent relation as

$$\begin{aligned} U_2(X, Y, \boldsymbol{Z}) &:= -D(\, \hat{P}_2(X, Y, \boldsymbol{Z}) \,||\, P_0(X, Y, \boldsymbol{Z})\,) \\ &= \sum_{x,y,\boldsymbol{z}} \hat{P}(x, y, \boldsymbol{z}) \log \frac{P(x, y \,|\, \boldsymbol{z})}{\hat{P}(x, y \,|\, \boldsymbol{z})} \ . \end{aligned} \tag{4.5}$$

The definitions of conditional independences described in Chpter 2 are used for representing each internal energy.

In the next step, the entropy term is defined with respect to each hypothesis. Probability distributions that constitute the entropy are estimated under given available samples. The entropy of the null hypothesis is described as

$$H_1(X, Y, \boldsymbol{Z}) := -\sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log(P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z}) P(\boldsymbol{z})) \ . \tag{4.6}$$

The other entropy, that of the opposite hypothesis, is

$$H_2(X, Y, \boldsymbol{Z}) := -\sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log(P(x, y \,|\, \boldsymbol{z}) P(\boldsymbol{z})) \ . \tag{4.7}$$

Now we are almost prepared to express the free energy of each hypothesis. The temperature in each hypothesis are regarded (β_1 and β_2) as a *global temperature* over related variables. According to the "Data Temperature" assumption, $\beta_1 = \beta_2 = \beta_0$, which means the same sample size. We can describe the hypotheses $\mathbb{H}_1$ and $\mathbb{H}_2$ as *free energies* F_1 and F_2 as

$$F_1 = U_1 - \frac{1}{\beta_0} H_1, \tag{4.8}$$

$$F_2 = U_2 - \frac{1}{\beta_0} H_2. \tag{4.9}$$

Therefore, the difference of the free energy of each hypothesis is expressed as

$$F_1(X, Y, \boldsymbol{Z}) - F_2(X, Y, \boldsymbol{Z}) = \hat{I}(X; Y | \boldsymbol{Z}) - \frac{1}{\beta_0} I(X; Y | \boldsymbol{Z}) \ , \tag{4.10}$$

where

$$\hat{I}(X; Y | \boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} \hat{P}(x, y, \boldsymbol{z}) \log \frac{\hat{P}(x, y \,|\, \boldsymbol{z})}{\hat{P}(x \,|\, \boldsymbol{z}) \hat{P}(y \,|\, \boldsymbol{z})} \ , \tag{4.11}$$

and

$$I(X; Y | \boldsymbol{Z}) = \sum_{x,y,\boldsymbol{z}} P(x, y, \boldsymbol{z}) \log \frac{P(x, y \,|\, \boldsymbol{z})}{P(x \,|\, \boldsymbol{z}) P(y \,|\, \boldsymbol{z})} \ . \tag{4.12}$$

According to the notation used in Chapter 3, we define the parameter β as

$$\beta := \frac{\beta_0}{\beta_0 + 1} \ , \tag{4.13}$$

where if $\beta_0 \to 0$, then $\beta \to 0$ (high temperature limit); if $\beta_0 \to \infty$, then $\beta \to 1$ (low temperature limit).

For estimating the non-empirical conditional mutual information $I(X, Y|\boldsymbol{Z})$ as described above, the MFE-EB method denoted in Chapter 3 for parameter learning is used, for which a different definition of internal energies U is needed. Let $P(\boldsymbol{X})$ and $\hat{P}(\boldsymbol{X})$ respectively represent probability distributions of joint random variables $\boldsymbol{X}$ to be estimated from the MFE principle and ML principle. Internal energies $U(\boldsymbol{X})$ are defined for parameter learning as

$$U(\boldsymbol{X}) = D(\, P(\boldsymbol{X}) \,||\, \hat{P}(\boldsymbol{X})\,) = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{\hat{P}(\boldsymbol{x})} \; . \tag{4.14}$$

Following the previous chapter, the estimated probability $P_\beta(\boldsymbol{x})$ is expressed in Boltzmann's formula, reproduced as below:

$$P_\beta(\boldsymbol{x}) = \frac{\exp(-\beta(-\log \hat{P}(\boldsymbol{x})))}{\sum_{\boldsymbol{x'}} \exp(-\beta(-\log \hat{P}(\boldsymbol{x'})))} = \frac{[\hat{P}(\boldsymbol{x})]^\beta}{\sum_{\boldsymbol{x'}} [\hat{P}(\boldsymbol{x'})]^\beta} \tag{4.15}$$

Therein, $\hat{P}$ is a relative frequency: the ML estimator.

Finally, we obtain the condition of conditional independence (CI), which we call MFE based CI condition as

$$\hat{I}(X; Y|\boldsymbol{Z}) < \frac{1-\beta}{\beta} I_\beta(X; Y|\boldsymbol{Z}) \; , \tag{4.16}$$

where I_β is defined as

$$\begin{aligned} I_\beta(X; Y|\boldsymbol{Z}) &= \sum_{x,y,\boldsymbol{z}} P_\beta(x, y, \boldsymbol{z}) \log \frac{P_\beta(x, y \,|\, \boldsymbol{z})}{P_\beta(x \,|\, \boldsymbol{z}) P_\beta(y \,|\, \boldsymbol{z})} \\ &= \sum_{x,y,\boldsymbol{z}} P_\beta(x, y, \boldsymbol{z}) \log \frac{P_\beta(x, y, \boldsymbol{z}) P_\beta(\boldsymbol{z})}{P_\beta(x, \boldsymbol{z}) P_\beta(y, \boldsymbol{z})} \; . \end{aligned} \tag{4.17}$$

Therein, β only plays the role of a symbolic index; it does not represent a sole parameter. In each estimator, β must be calculated using the explicit model of "Data Temperature." Therefore, β in (4.17) represents *local temperature*. In (4.16), the left-hand-side corresponds to the likelihood term, which is dominant for a large data size (large β), and the right-hand-side corresponds to the entropy term, which is dominant for a small data size (small β). We designate g_β^2 and represent the MFE based CI condition with it as

$$g_\beta^2 = \hat{I}(X; Y|\boldsymbol{Z}) - \frac{1-\beta}{\beta} I_\beta(X; Y|\boldsymbol{Z}) < 0 \; . \tag{4.18}$$

This is useful for combination with the classical hypothesis tests.

4.4 “Data Temperature” Model

In searching for the values of β, a simple model of temperature is used; it is proposed as a function of data size N described in the preceding chapter. The model function of β is defined as

$$\begin{aligned}\beta &:= 1 - \exp\left(-\frac{N}{\gamma N_c}\right) \ , \\ \gamma &:= |\boldsymbol{X}| - 1 \ ,\end{aligned} \tag{4.19}$$

where γ is defined as the degrees of freedom of related random variables $\boldsymbol{X}$, and where N_c is a decoupling constant, which can be regarded as a hyperparameter for β. We use N_c as only one *common* hyperparameter in learning of both parameter and structure. This explicit model shows good performance and robustness against selected hyperparameters N_c in classification tasks using Bayesian network classifiers with structure learning, as described in the previous chapter.

4.5 Asymptotic Theoretical Analysis

The proposed method is hoped to provide consistency with the classical hypothesis test for an asymptotic region because it is theoretically justified. However, the conditional independence conditions using the inequality (4.16) cannot be used straightforwardly for large data sizes because $g^2_\beta \geq 0$ always for sufficiently large data size because $\hat{I}(X;Y|\boldsymbol{Z}) \geq 0$ and $[(1-\beta)/\beta]\, I_\beta(X;Y|\boldsymbol{Z}) \to 0$ as β goes to 1 (as N becomes sufficiently large), which means that our method would produce an overly dense graph for sufficiently large data size. In such regions, the effect of enlarging the entropy term has vanished and the likelihood term has become dominant. However, different from parameter learning, hypothesis testing for BNs means that extra edges should be removed even for a large sample size, based on *Occam's razor* [Pearl, 2000]. This connecting problem is solved as described below.

For a large sample size region, we wish to use the G^2 statistic for conditional independence testing, which is often used [Spirtes et al., 2000; Tsamardinos et al., 2006]. The G^2 test is used to identify $Ind\,(X, Y|\boldsymbol{Z})$, by which the null hypothesis of conditional independence is represented. Let $N_{xy\boldsymbol{z}}$ represent the number of times in the data where $X = x, Y = y$ and $\boldsymbol{Z} = \boldsymbol{z}$. We define $N_{x\boldsymbol{z}}, N_{y\boldsymbol{z}}$, and $N_{\boldsymbol{z}}$ similarly. Consequently, the G^2 statistic is defined as follows:

$$G^2 = 2 \sum_{x,y,\boldsymbol{z}} N_{xy\boldsymbol{z}} \log \frac{N_{xy\boldsymbol{z}} N_{\boldsymbol{z}}}{N_{x\boldsymbol{z}} N_{y\boldsymbol{z}}} \ . \tag{4.20}$$

The degrees of freedom *df* are defined as

$$df = (|X| - 1)(|Y| - 1) \prod_{Z \in \boldsymbol{Z}} |Z| \ , \tag{4.21}$$

where we designate $|X|$ as the number of states in X. It is noteworthy that the G^2 statistics have a relation with the empirical mutual information with data size N [Kullback, 1968] as

$$G^2 = 2N \, \hat{I}(X;Y|\boldsymbol{Z}) \ . \tag{4.22}$$

The statistic is proven to be approximated asymptotically to a χ^2 distribution with degrees of freedom *df* [Kullback, 1968]. Therefore, in a large sample size region, we should set the condition in which the null hypothesis (i.e. conditional independence) is not rejected, as

$$G^2 < \chi^2_{\alpha,df} \ , \tag{4.23}$$

where α is a significance level such as 0.05, and where *df* are the degrees of freedom, as defined in (4.21).

It is worthy to note that if we need a threshold with information theoretical approach as Cheng et al. [2002] did, then we can decide the threshold incorporating degrees of freedom and available sample size using the G^2 statistics, as pointed out as a problem related to using mutual information tests and arbitrary thresholds [Tsamardinos et al., 2006].

Here, we intend to connect the classical condition with the MFE based CI condition represented by eq. (4.16). A formal correspondence amount G^2_β to G^2 is defined using (4.18) and (4.22) as

$$\begin{aligned} G^2_\beta &:= 2N g^2_\beta \\ &= G^2 - 2N \, \frac{1-\beta}{\beta} \, I_\beta(X;Y|\boldsymbol{Z}) \ . \end{aligned} \tag{4.24}$$

We can recognize that G^2_β converges to G^2 if β converges to 1 faster than $O(N)$ in an asymptotic region. Furthermore, a stronger convergence characteristic can be proven as follows:

Theorem 4.1 *If $N \to \infty$, then G^2_β converges to the G^2 statistics.*

Proof. *In the asymptotic region, β approaches* 1 *because of the "Data Temperature" assumption; for that reason, $I_\beta(X;Y|\boldsymbol{Z})$ approaches $\hat{I}(X;Y|\boldsymbol{Z})$. Then, G^2_β is described* *as*

$$\begin{aligned} G^2_\beta &\to 2N\hat{I}(X;Y|\boldsymbol{Z}) - 2N\frac{1-\beta}{\beta}\hat{I}(X;Y|\boldsymbol{Z}) \\ &= 2N\hat{I}(X;Y|\boldsymbol{Z})\left(1 - \frac{1-\beta}{\beta}\right) \\ &\to 2N\hat{I}(X;Y|\boldsymbol{Z}) = G^2. \end{aligned}$$

□

Then, the MFE and the classical condition can be treated in the unified treatment because a condition $G^2_\beta < \chi^2_{\alpha,df}$ can include the MFE based CI condition (4.18) and the classical CI condition (4.23). Even when the data size is small and $G^2_\beta \geq 0$, the classical hypothesis tests can be conducted because our method was shown to generate pseudo-samples similarly to the Bayesian methods, as described in eq. (3.20). Consequently, conditional independence tests can be conducted on variables X and Y given $\boldsymbol{Z}$ using the MFE principle and G^2 tests, as described below.

- If $G^2_\beta < 0$ because of the MFE principle, then we set $X \perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$ (conditional independence),
- else if $0 \leq G^2_\beta < \chi^2_{\alpha,df}$, because of the classical test, then we set $X \perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$,
- else, we set $X \not\!\perp\!\!\!\perp Y \,|\, \boldsymbol{Z}$ (conditional dependence).

We designate this conditional independence method as MFE–CI.

4.6 Experiments

Next the performance of our approach is demonstrated compared with traditional statistical testing methods. Some experiments of learning BNs are done using the PC algorithm [Spirtes et al., 2000], which is a well known benchmark algorithm of constraint-based methods, embedding conditional independence tests or classical independence tests using χ^2 distributions with fixed significant level $\alpha = 0.05$ for each hypothesis test of conditional independence. The PC algorithm was implemented as described in section 2.5.1 for embedding the MFE–CI method using C++ programming language.

The PC algorithm is performed under the *faithfulness assumption* described in Chapter 2. Consequently, the algorithm can infer correct graph structures by finding conditional independence for probability distributions. However, if the assumption is violated, even though the true graph means $Ind\,(X, Y|\boldsymbol{Z})$ for X and Y and a conditional set $\boldsymbol{Z}$, then the algorithm might find another false conditional set $\boldsymbol{Z'}$ for the test between X

and Y, and then add $\boldsymbol{Z'}$ to *SepSet(XY)*, which denotes a separator set between X and Y, in PC algorithm described in section 2.5.1. This false detection has *no influence* on removing the edge between X and Y correctly, which means that Adjacency Faithfulness [Ramsey et al., 2006] is satisfied but that Orientation Faithfulness [Ramsey et al., 2006] is violated. However, the algorithm decides the wrong direction of edges using the orientation rules described in section 2.5.1. In this situation, finding correctly conditional sets strongly influences the directionality of edges in BNs. When the conditional sets $|\boldsymbol{Z}|$ are numerous, the number of CI tests is intractably large because of a combinatorial explosion. Therefore, we did not perform CI tests and assume *conditional dependence* when $|\boldsymbol{Z}| \geq 5$. A value of the hyperparameter N_c was selected for β in (4.19) as 2.0, which shows good performance in preliminary experiments.

4.6.1 Simulation Studies

Settings

We conducted the simulation study with various quantities of variables: {10, 20, 40, 80}, where each variable has all four possible states, and with networks of two types, i.e. the sparser and denser graphs, where sparser cases have the same number of edges as variables; the denser cases have twice. For each such graph, a random structure network was constructed with conditional probability tables (CPTs) of five types that were set by random numbers. The available sample size varies in a range of {500, 1000, 2500, 5000, and 10000}. The performance criteria were set as counting *added edges*, *removed edges*, and *reversed edges*. Counting added edges expresses the consequence where two variables X and Y are not adjacent in original BNs, but where an edge exists between them in reconstructed BNs. On the other hand, counting removed edges indicates the opposite. Counting reversed edges means that if $X \to Y$ in the original, then $Y \to X$ in the output.

Results

Table 4.1 presents results for sample sizes of 500 and 1000. Table 4.2 shows those for 2500 and 5000. Table 4.3 shows those for 10000. The values in the tables are averaged values of simulations for five random sets of CPTs. We designate the PC algorithm with a standard G^2 test as *Std-PC* or *Std*, and the PC embedded with the MFE–CI as *MFE-PC*.

These tables show that the counted quantities of extra added edges were very small, even for a small sample size such as 500 and even for denser structures. In contrast, quantities of removed edges are large to some degree in both Std-PC and MFE-PC. The

MFE-PC removed true edges more than Std-PC to a certain degree. Reversed checks revealed many errors in Std-PC. These characteristics were noticeable in large and dense networks. These results are discussed later. A key for understanding the results is apparently the *faithfulness condition*.

The MFE-PC shows great effectiveness for deciding the direction of edges. It might be unfair, however, to conclude that because the MFE-PC removed more edges than Std-PC. Therefore, we defined *reversed ratio* as (number of reversed edges)/((true number of edges) – (number of removed edges)). Results of *reversed ratio* for denser networks are portrayed in Fig. 4.1, 4.2 and 4.3. where the horizontal axis expresses the true number of edges, the vertical axis expresses the *reversed ratio*, and G2 and MFE respectively signify Std-PC and MFE-PC. In addition, Figs. 4.1, 4.2 and 4.3 show that the algorithm found wrong v-structures, which are marked as *V-err* in the figures; and those results resemble that of the reversed ratio, which is discussed later. These figures show that the MFE-PC outperforms Std-PC in deciding the direction of edges, especially for denser networks, even using samples such as 5000, which are not regarded as small samples in general.

Discussion

Discussion of these comparative results demands some reference to the validity of evaluation using *added edges* and *removed edges* in this simulation. In fact, MFE-PC performs CI tests in more cases than Std-PC, which does the test only for sufficiently large data size. For example, even for data size $N = 5000$, Std-PC was unable to perform CI tests for $|\boldsymbol{Z}| \geq 3$ in this simulation, which implies that Std-PC might sometimes correctly happen to maintain some existing edges. Short of undoing the tests for such frequent cases, there is not so great a difference in the number of errors for added edges between MFE-PC and Std-PC. This result suggests that Std-PC detects wrong separator sets, and then the simulation data were likely to be regarded as violationg faithfulness condition. In other words, unfortunately, these data were under threat of violating the condition that $Ind(X;Y|\boldsymbol{Z}) \Rightarrow Dsep^{\mathbb{G}}(X;Y|\boldsymbol{Z})$. This situation was also reported by Ramsey et al. [2006] for linear Gaussian models of DAGs where they also used simulation data. The fact complicates recognition of the differece for true power of the test between Std-PC and MFE-PC. Therefore, we must consider a greater deal of the ratio of reversed edges than the counts of added and removed edges for this simulation. For the results shown in Figs. 4.1, 4.2 and 4.3, it is necessary to emphasize that MFE-PC more correctly decided the direction of edges than Std-PC. This fact means that Std-PC was likely to detect conditional independence for invalid conditional sets $\boldsymbol{Z}$ more than MFE-PC. In fact, Std-PC generated more wrong v-structures, divergence connections such as $X \leftarrow U \rightarrow Y$, and serial connections such as $X \rightarrow U \rightarrow Y$ in the edge orientation algorithms. This result influences on the correctness for the results of edge orientations in step 4 of the PC algorithm, which are described in section 2.5.1. *V-errs* of Figs. 4.1, 4.2, and 4.3 definitely show the mistakes, in the wrong v-structure counts on MFE-PC and Std-PC, appearantly show that the reversed edges are mainly attributable to the incorrectly detecting colliders, which result from the following situations: for triplet $\{X, Y, W\}$, Std-PC incorrectly detected $Ind\,(X;Y|\boldsymbol{Z}')$ for $W \in \boldsymbol{Z}'$ while MFE-PC correctly detected $Ind\,(X;Y|\boldsymbol{Z})$ for $W \notin \boldsymbol{Z}$ (see, Chapter 2). For example, assuming a DAG that consists of four nodes $\{X, Y, U, W\}$, if $Ind\,(X;Y|\{U, W\})$, then the graph appears as shown in Fig. 4.4(a), which shows an undirected graph as a Markov equivalent class. If the algorithm wrongly detected $Ind\,(X;Y|\,U)$ for the DAG, then the constructed network are presented in Fig. 4.4(b). The reason is that a node W is a collider if wrong $Sepset(X, Y)$ has only U while true $Sepset(X, Y)$ has U and W, which means that both U and W cannot be colliders.

Table 4.1: Results for the simulation using data sizes of 500 and 1000

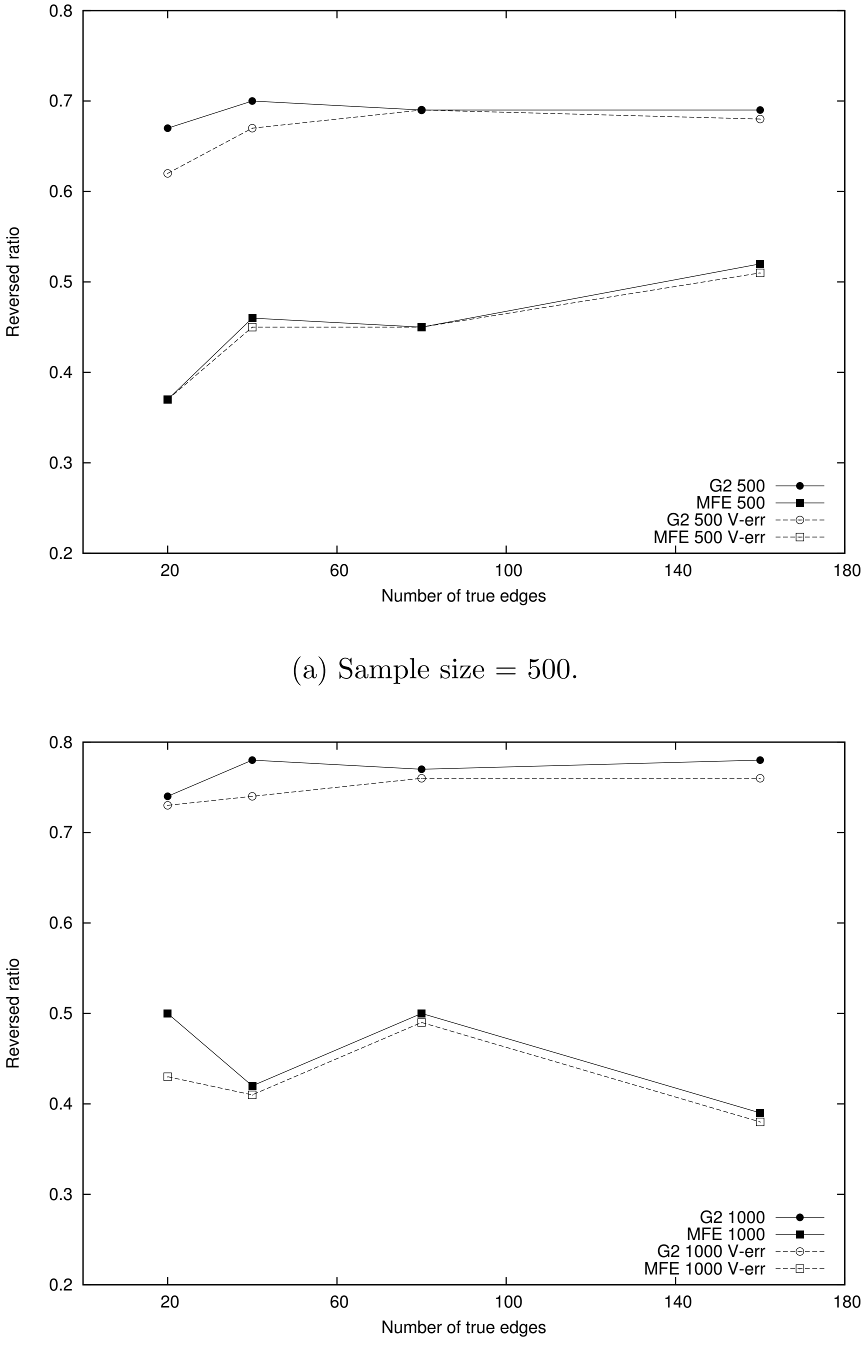
		500				1000			
		Sparser		Denser		Sparser		Denser	
Type	Nodes	Std	MFE	Std	MFE	Std	MFE	Std	MFE
Added	10	0	0	0.6	0.2	0.4	0	1.8	0
	20	0	0	0.4	0.4	0.2	0	1.2	0
	40	0.2	0	0.4	0	0.6	0.4	0.8	0
	80	0.6	0.4	1.4	0.8	0.4	0.2	2.2	0.2
Removed	10	3.0	3.4	9.0	14.0	2.4	2.8	3.6	11.6
	20	4.2	7.6	19.0	26.2	3.0	5.8	11.4	22.0
	40	11.2	15.8	41.6	53.6	6.4	12.0	25.0	46.6
	80	21.4	32.0	81.2	109	11.0	21.2	48.6	93.8
Reversed	10	1.8	1.4	7.4	2.2	0.8	0.8	12.2	4.2
	20	5.4	2.0	14.6	6.4	5.2	2.6	22.2	7.6
	40	10.6	5.2	26.6	11.8	10.2	4.0	42.6	16.8
	80	18.8	10.6	54.2	26.6	21.0	11.2	86.4	26.0

Table 4.2: Results for simulations using data sizes of 2500 and 5000

		2500				5000			
		Sparser		Denser		Sparser		Denser	
Type	Nodes	Std	MFE	Std	MFE	Std	MFE	Std	MFE
Added	10	0	0	0	0	0	0	1.2	0.2
	20	0	0	0.2	0	0	0	0.4	0.0
	40	0	0	0	0	0.4	0.4	0.0	0.0
	80	0.2	0.2	0	0	0.4	0.4	0.0	0.0
Removed	10	1.8	1.8	4.4	8.6	1.0	1.0	1.8	6.4
	20	2.0	2.8	12.8	18.6	0.4	1.4	6.2	14.6
	40	6.0	7.2	26.0	37.2	2.6	4.6	16.0	30.0
	80	9.0	12.0	54.4	73.8	4.6	7.2	24.2	59.2
Reversed	10	0.6	0.6	9.2	5.2	0.6	0.6	7.2	4.4
	20	2.8	2.4	13.2	7.6	2.6	2.4	20.8	9.6
	40	6.6	6.4	31.8	19.4	5.0	3.8	36.6	20.8
	80	11.6	10.8	56.6	38.8	8.4	7.8	78.0	38.0

Table 4.3: Results for simulations using data size of 10000

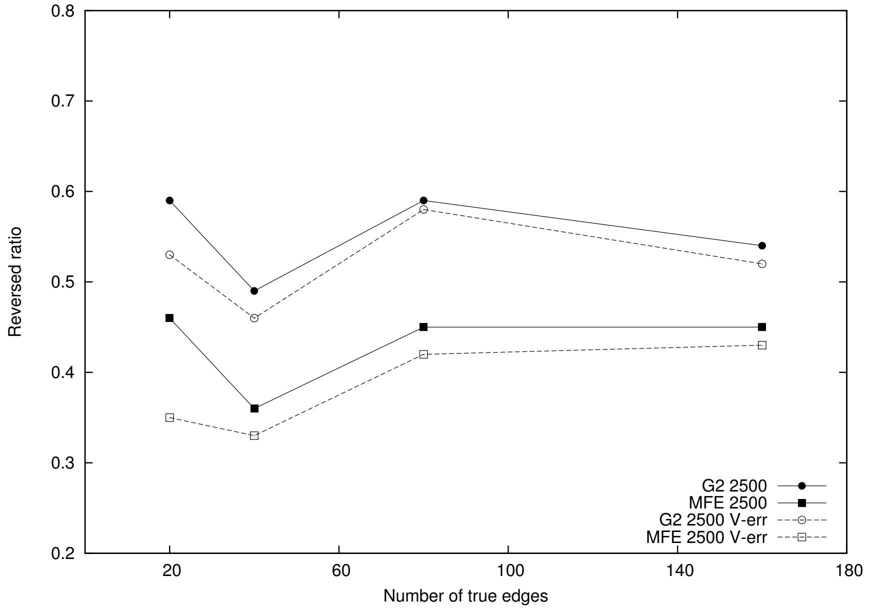
		10000			
		Sparser		Denser	
Type	Nodes	Std	MFE	Std	MFE
Added	10	0	0	1.2	0.2
	20	0	0	0.4	0.0
	40	0.4	0.4	0.0	0.0
	80	0.4	0.4	0.0	0.0
Removed	10	1.0	1.0	1.8	6.4
	20	0.4	1.4	6.2	14.6
	40	2.6	4.6	16.0	30.0
	80	4.6	7.2	24.2	59.2
Reversed	10	0.6	0.6	7.2	4.4
	20	2.6	2.4	20.8	9.6
	40	5.0	3.8	36.6	20.8
	80	8.4	7.8	78.0	38.0



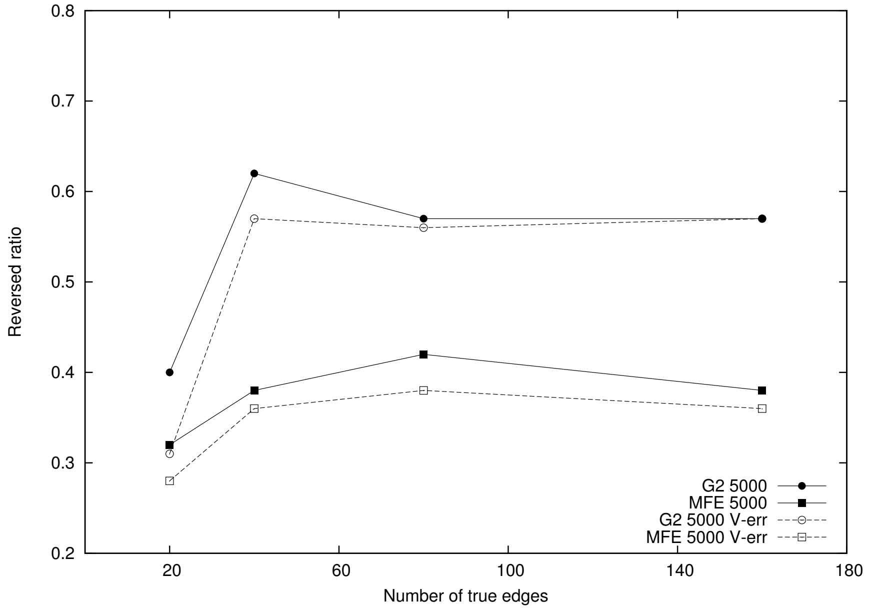
(a) Sample size = 500.

(b) Sample size = 1000.

Figure 4.1: Ratio of reversed edges in the resultant graphs with denser BNs from the use of a standard PC and PC embedded with the MFE–CI method: (a) Sample size = 500 and (b) 1000.

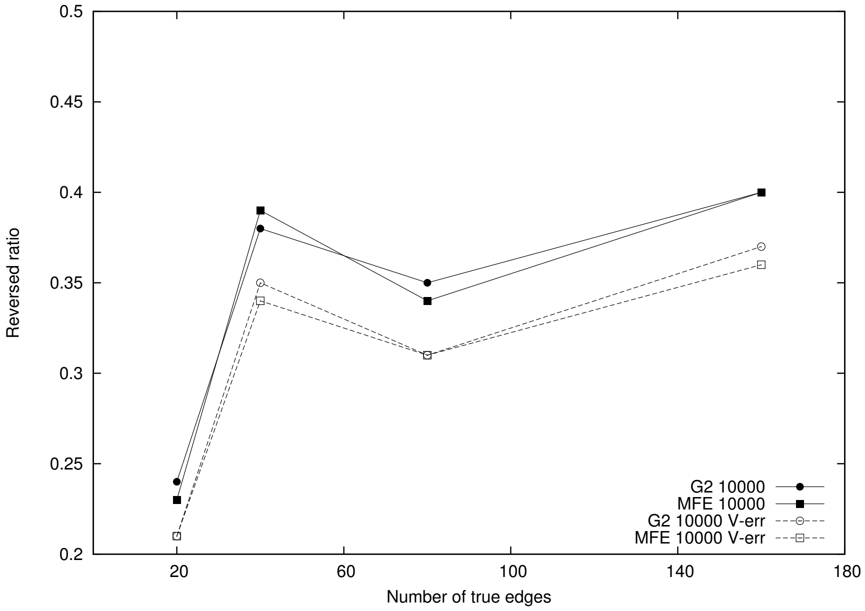


(c) Sample size = 2500.



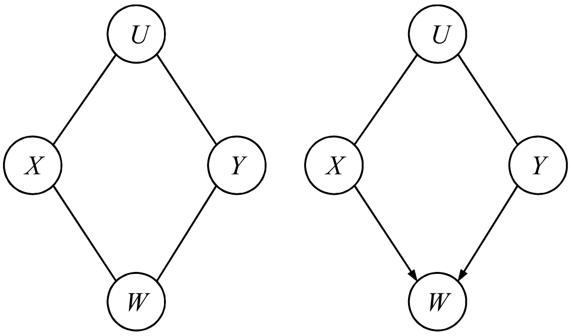
(d) Sample size = 5000.

Figure 4.2: Ratio of reversed edges in resultant graphs with denser BNs from the use of a standard PC and PC embedded with the MFE–CI method: (c) Sample size = 2500 and (d) 5000.



(e) Sample size = 10000.

Figure 4.3: Ratio of reversed edges in the resultant graphs with denser BNs from the use of a standard PC and PC embedded with the MFE–CI method: (e) Sample size = 10000.



(a) True Graph. (b) False Graph.

Figure 4.4: True graph (a) represents $Ind\,(X;Y|\{U,W\})$ while a false graph (b) represents $Ind\,(X;Y|\,U)$, which generates a wrong collider W.

4.6.2 Real World Datasets

Settings

In addition to simulation experiments, we will show other experiments using real-world datasets. Training cases are sampled from the probability distributions of known networks, and ask the Std-PC and MFE-PC to reconstruct the original network structures from the data. We selected five Bayesian networks from datasets, which consist of definite DAG structures and discrete conditional probability distributions, typically used for reconstructing BN structures: Alarm (Fig. 4.5), Insurance (Fig. 4.6), Barley (Fig. 4.7), Mildew (Fig. 4.8) and Hailfinder (Fig. 4.9). The Alarm network [Beinlich et al., 1989] was constructed by medical experts for monitoring patients in intensive care wards. The Insurance network [Binder et al., 1997] is used for evaluating car insurance risks. The Mildew [Jensen and Jensen, 1996] is a preliminary model for deciding on the amount of fungicides to be used against attack of mildew in wheat. The Barley network [Kristensen and Rasmussen, 2002] is a model of barley crops yield. The Hailfinder network [Abramson et al., 1996] is a normative system that forecasts severe summer hail in northeastern Colorado. Information of the networks used in the thesis is presented in Table 4.4. In the table, *Num.vars*, *Num.edges* and *Num.params* represent the number of variables, edges and all parameters for each. *Max In/Out degree* signifies the maximum degree of incoming edges and outgoing edges at a node. *Domain range* represents the ranges of internal states of variables. The Barley and Mildew have larger networks than Alarm and Insurance, as viewed from the scale in the parametric space size. Learning performances were examined for various training sample sizes:{250, 500, 1000, 1500, 2000, 5000, 10000}. For the experiments, we sampled 10 distinct datasets from CPTs for each number of training data; the values shown herein are averaged ones for 10 datasets.

Table 4.4: Real-world Bayesian networks used in the experiments.

Network	Num. vars	Num. edges	Num. params	Max In/Out- degree	Domain range
Alarm	37	46	509	4 / 5	2-4
Insurance	27	52	1008	3 / 7	2-5
Hailfinder	56	66	2656	4 / 16	2-11
Barley	48	84	114005	4 / 5	2-67
Mildew	35	46	540150	3 / 3	3-100

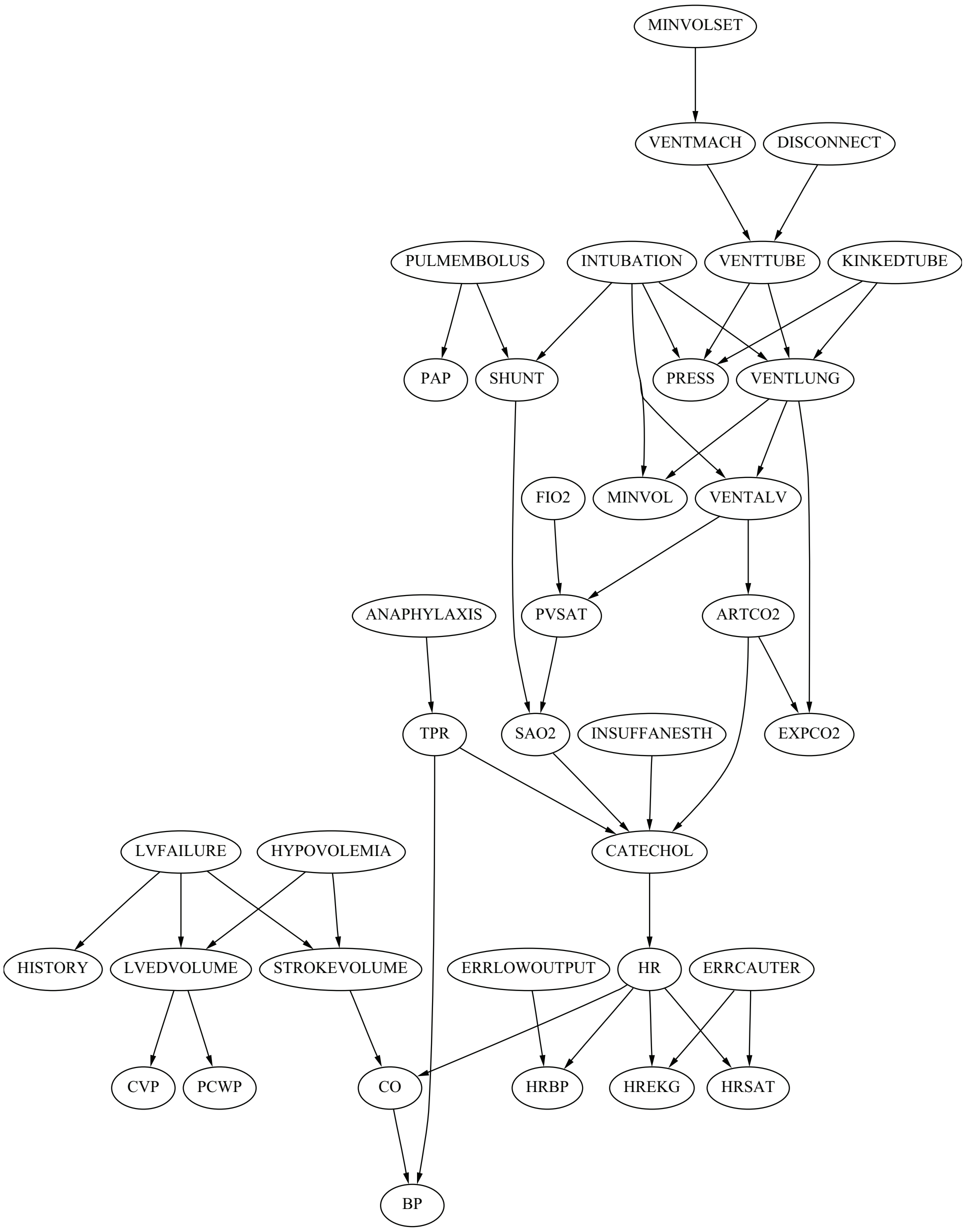


Figure 4.5: Alarm network

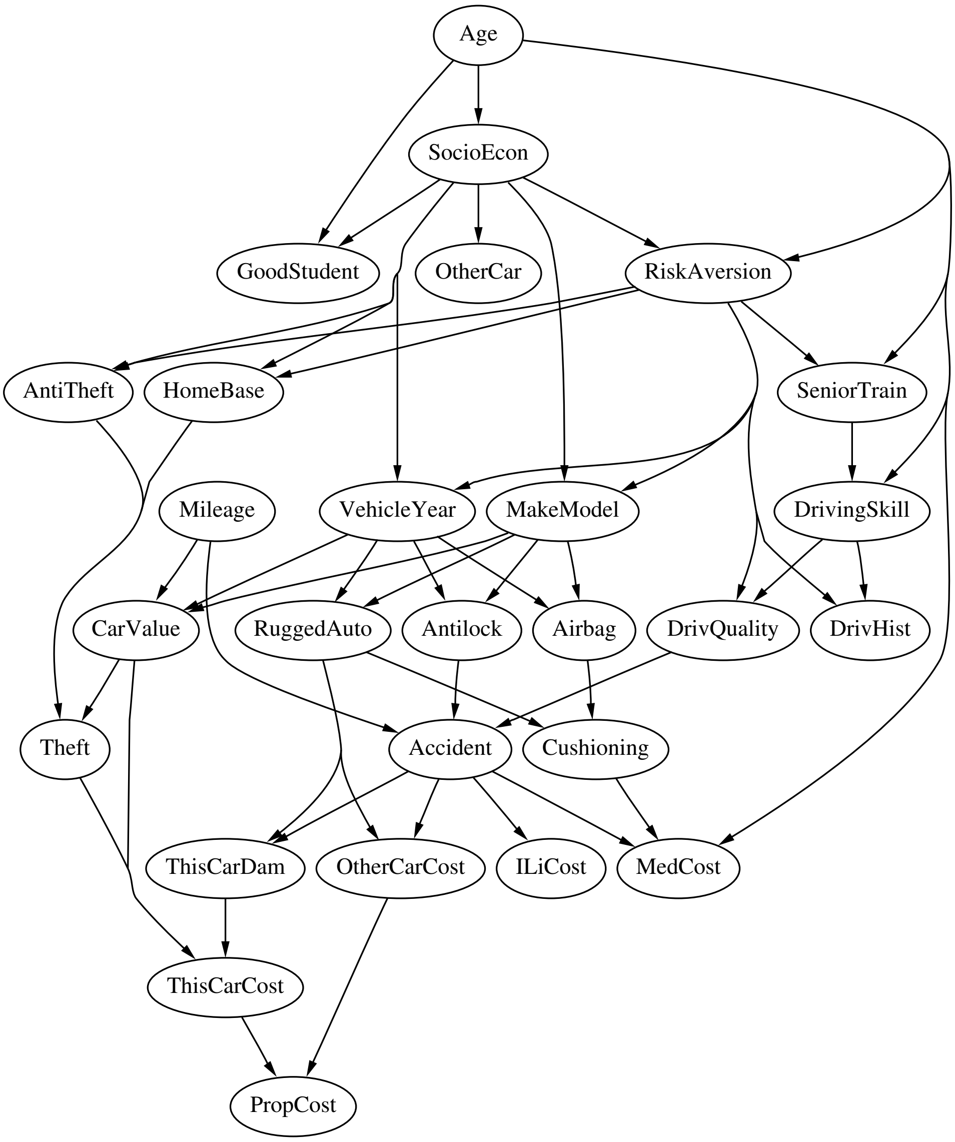


Figure 4.6: Insurance network

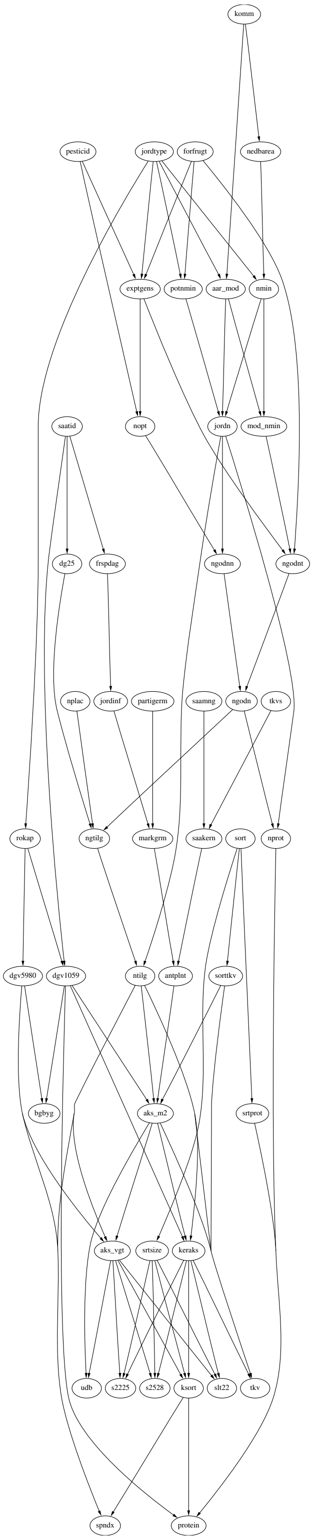


Figure 4.7: Barley network

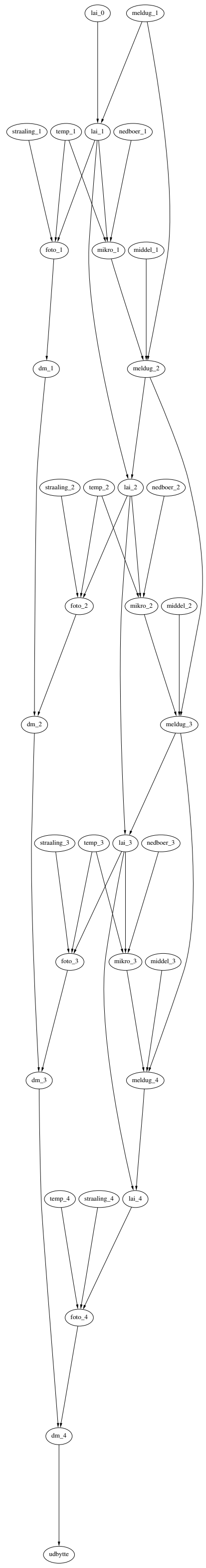


Figure 4.8: Mildew network

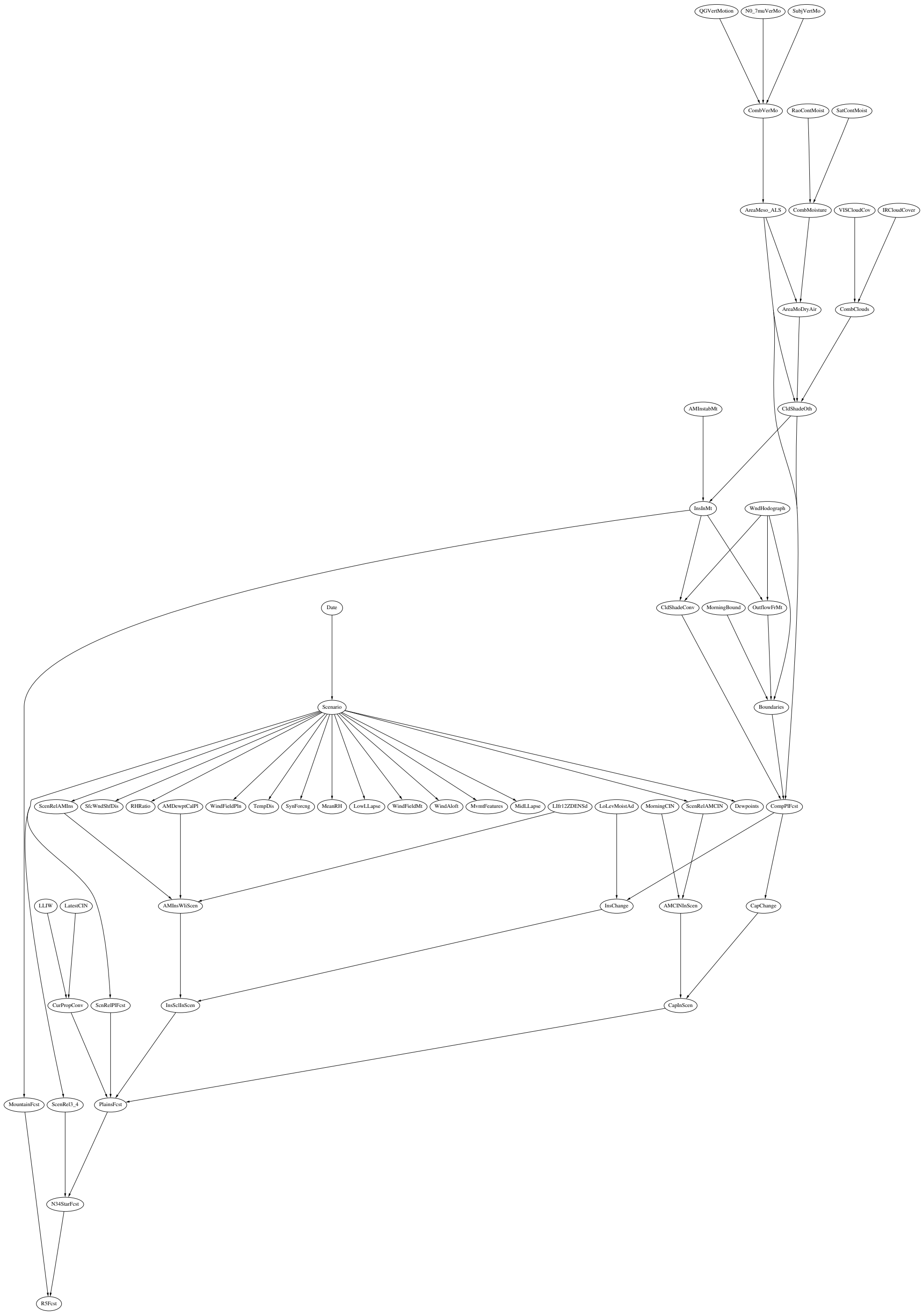


Figure 4.9: Hailfinder network

Results

The obtained results present some degree of difference from simulation data. In the real-world data, there also exist add–remove errors, although they are only slightly recognized in the simulation results. Two-tailed t-tests were used to test the differences between MFE-PC and Std-PC with 0.05 of significant level. First, we show the reversed edge ratio on each network in Fig. 4.10, 4.11, 4.12, 4.13 and 4.14, where the vertical axes denote the sum of numbers of extra edges and missing edges, and the horizontal axes denote number of training samples and, for each column, designate Std-PC for *G2* and MFE-PC for *MFE*. Here, MFE-PC is superior for

- $N = |\mathcal{D}| \leq 1000$ in the Alarm network,
- $N \leq 500$ in the Insurance network,
- $N \leq 2000$ in the Hailfinder network,
- $N \leq 10000$ in the Barley network, and
- $N \leq 10000$ in the Mildew network,

and there was no statistical difference between two methods in the other sample sizes. The results show clearly that MFE-PC is superior to Std-PC for small data size for Alarm and Insurance, which we intend to be, and for Hailfinder, Barley and Mildew superior for from small to medium data size.

The results of the add–remove errors are shown next, differences of which are scarcely found for the simulation study. The results are shown in Figs. 4.15, 4.16, 4.17, 4.18, and 4.19. The range in which MFE-PC has advantages is over Std-PC, as in the following:

- $N = 250$ for Alarm,
- $N = 250$ for Insurance,[1]
- $N \leq 2000$ for Hailfinder,
- $N \leq 5000$ for Barley, and
- $500 \leq N \leq 10000$ for Mildew,

and no statistical difference was observed between Std-PC and MFE-PC in the other sample sizes. Therefore, from the add–remove errors combined to the reversed edge ratio, we can conclude that MFE-PC shows superiority to Std-PC for small data size

[1] For only $N = 500$, Std-PC is superior to MFE-PC.

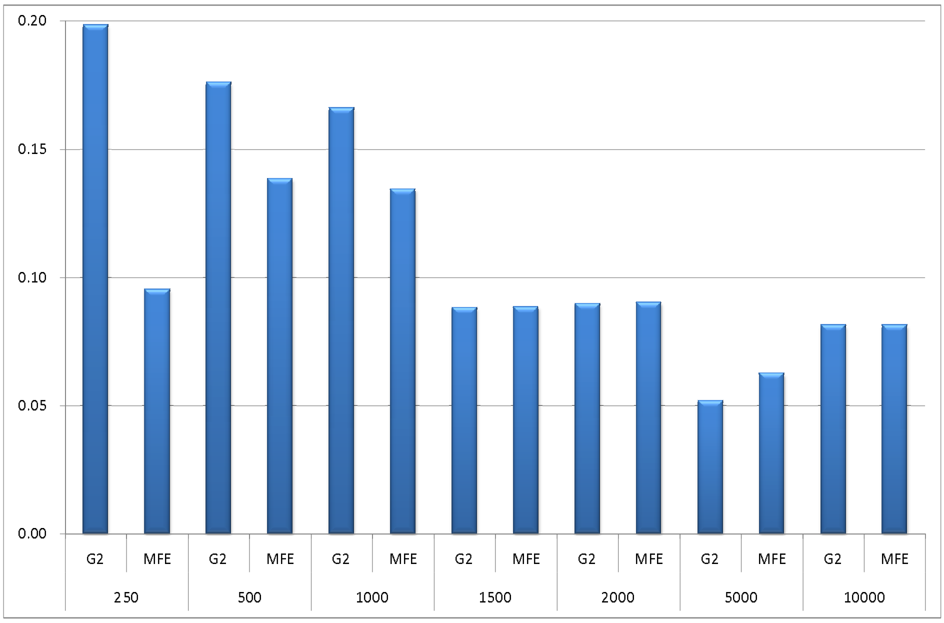


Figure 4.10: Reversed edge ratios in the Alarm network

in general and even for medium data size in large networks in view of parametric space size.

For 250 sample sizes of both the Barley and Mildew networks, no difference was found between Std-PC and MFE-PC in add–remove errors. The Std-PC did not conduct the CI test for the size of conditioning set $|\boldsymbol{Z}| \geq 1$, whereas MFE-PC decided all these as independent. Therefore, too small samples exist to conduct the CI tests appropriately for the sample size.

Barley and Mildew are very large scale networks, as shown from the perspective of number of parameters in Table 4.4, and Mildew, as expected, seems to need very large training samples over 10000. However, Barley does not apparently need such very large data for deciding existence of edges in relation to its parametric size.

According to results of both the reversed ratio and add–remove errors, the difference in the former type of error is seen for wider sample ranges between MFE-PC and Std-PC. This result implies that it is more difficult to detect correct separator sets (*SepSet* in PC algorithm, see section 2.5.1) than to detect independencies.

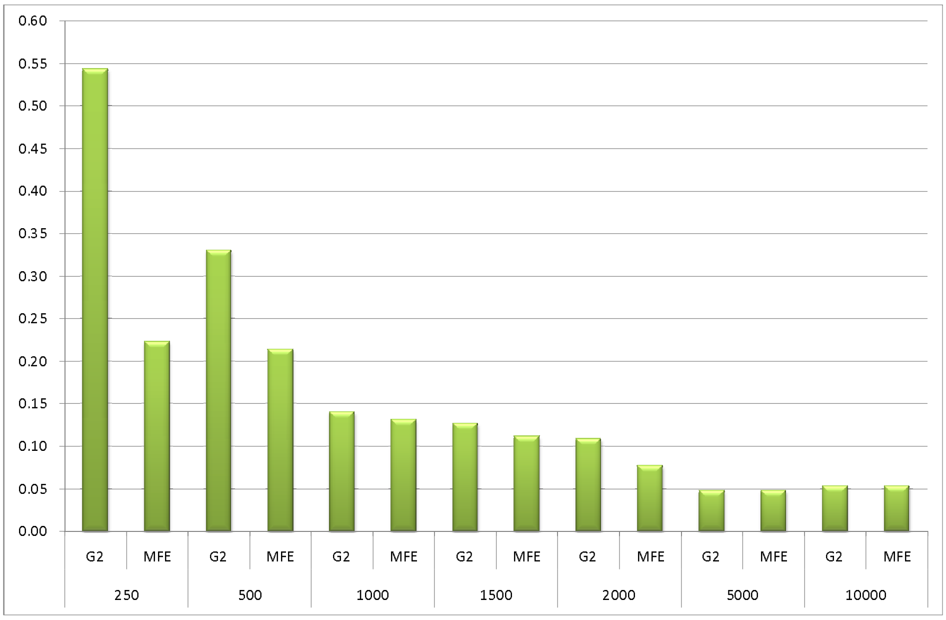


Figure 4.11: Reversed edge ratios in the Insurance network

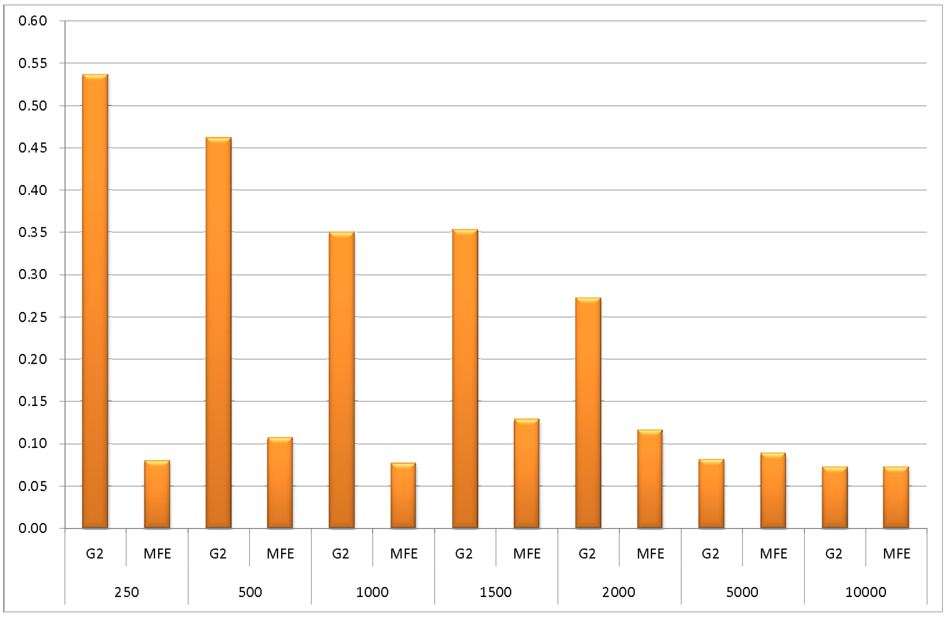


Figure 4.12: Reversed edge ratios in the Hailfinder network

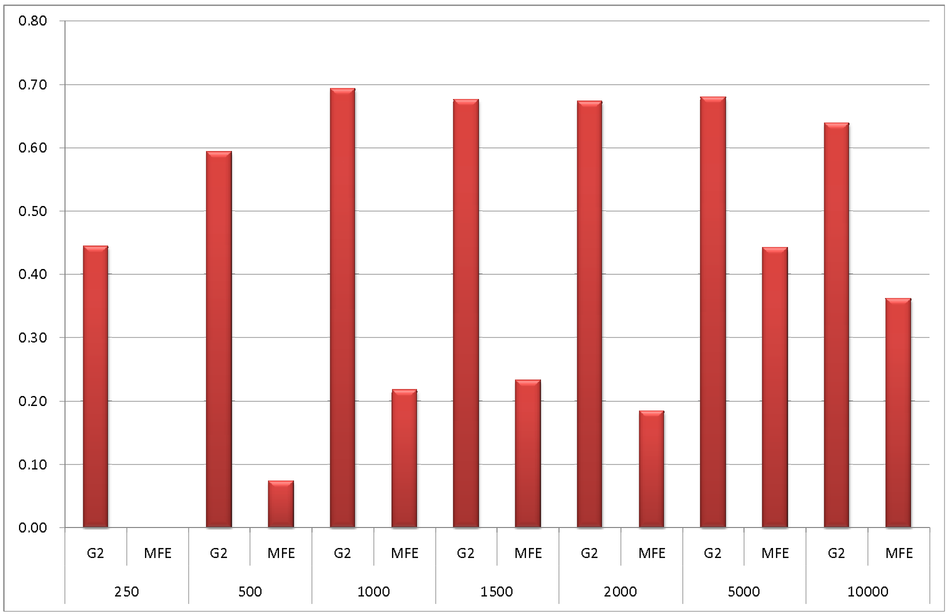


Figure 4.13: Reversed edge ratios in the Barley network

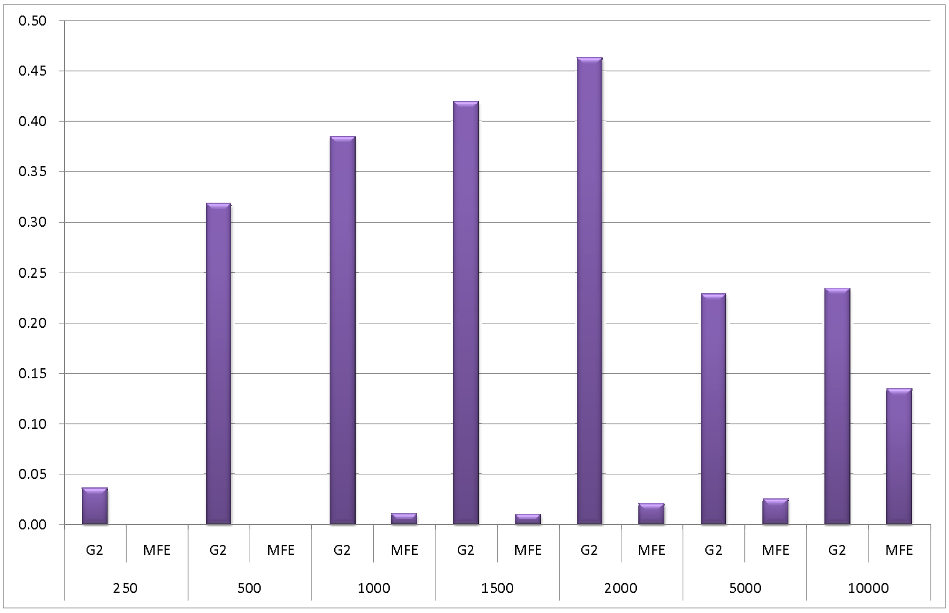


Figure 4.14: Reversed edge ratios in the Mildew network

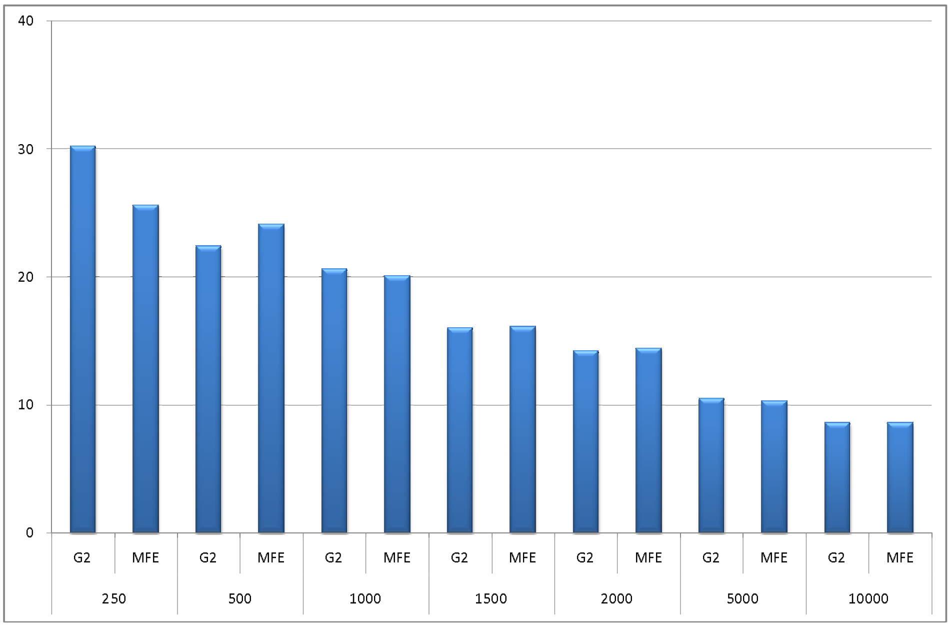


Figure 4.15: Number of add–remove errors in the Alarm network

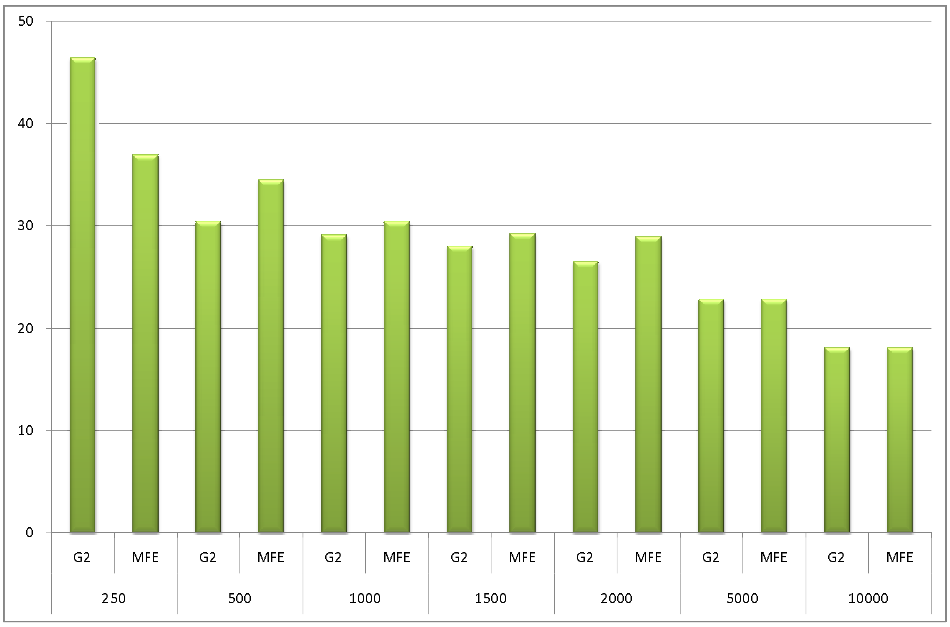


Figure 4.16: Number of add–remove errors in the Insurance network

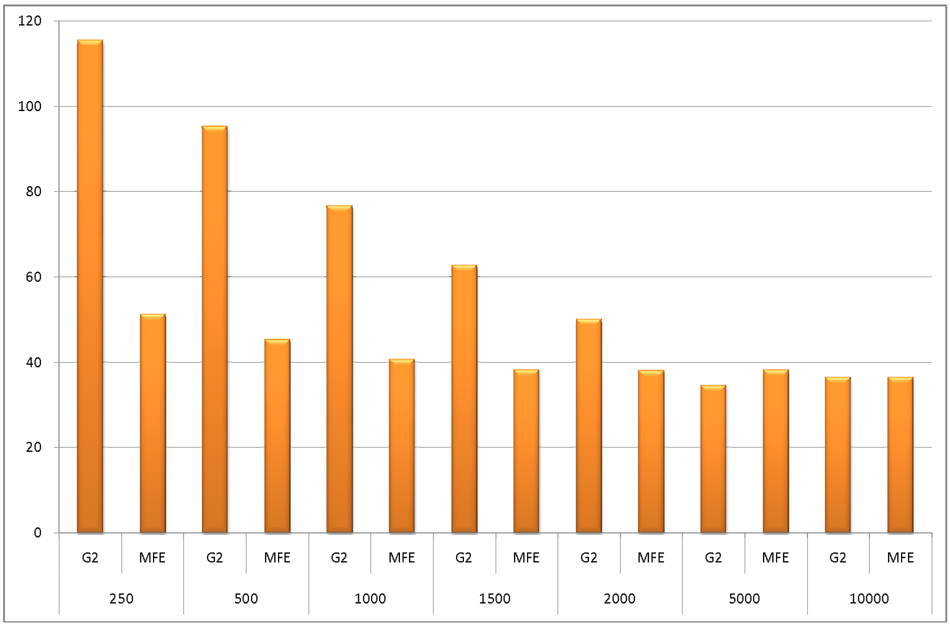


Figure 4.17: Number of add–remove errors in the Hailfinder network

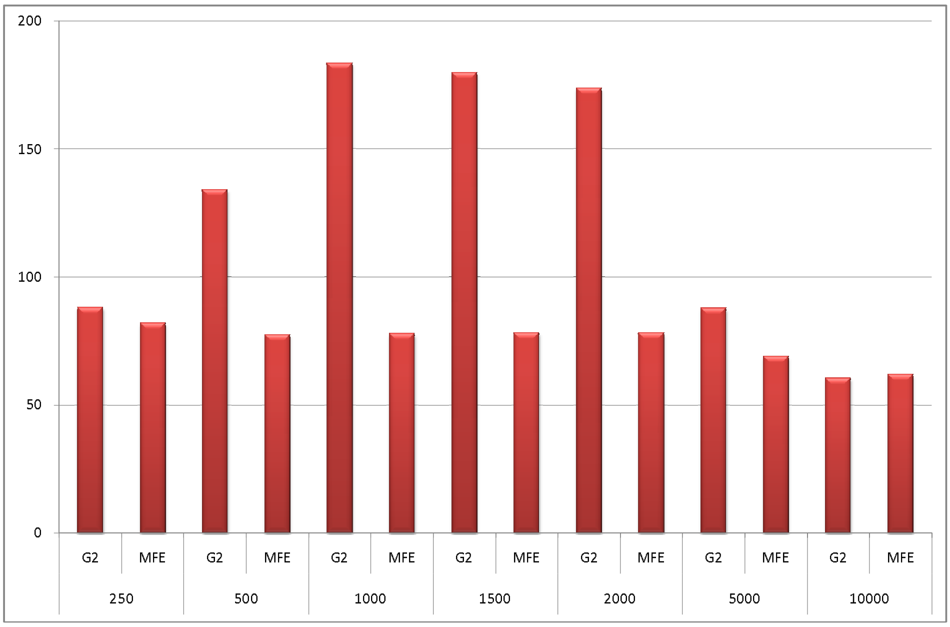


Figure 4.18: Number of add–remove errors in the Barley network

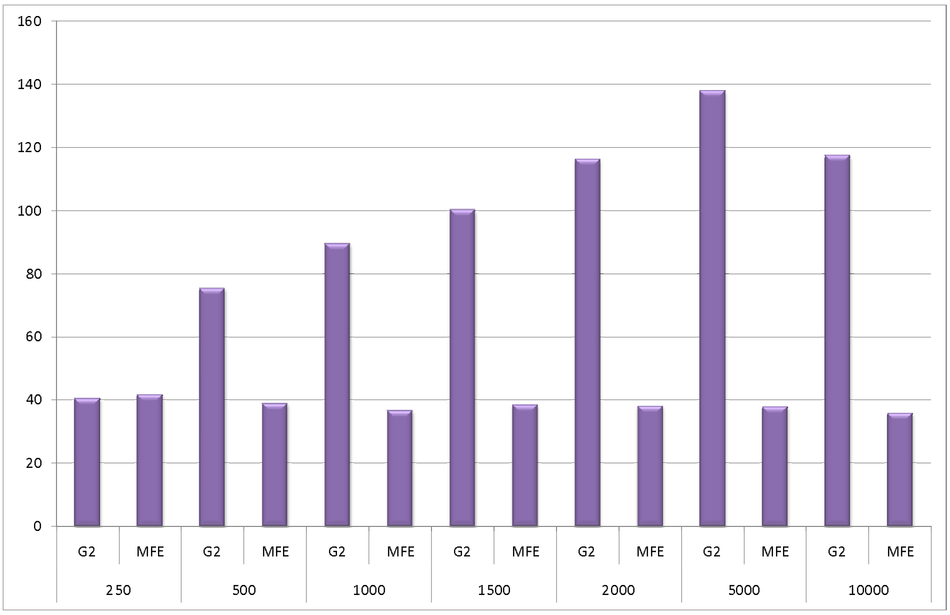


Figure 4.19: Number of add–remove errors in the Mildew network

4.6.3 General Discussion of Experiments

Next, we discuss the results for both the simulation and real-world datasets. Although a tendency to regard the faithfulness condition as violated was noted in the simulation studies, such was not the case for the real-world datasets. The reason is probably the randomness of the generation process of simulation data, although real data have a definite bias or tendency and are not random in (conditional) probabilities. Although Ramsey et al. [2006] also reported violation of the faithfulness condition, where they found 40% violation of the condition, their study used simulation data only. Therefore, that awful situation for the violation probably occurs only rarely when using real data.

In summary, MFE–CI method is robust for the data that have the tendency of nearly violating the faithfulness condition. This property is preferred for causal discovery, in which existing edges are expected to represent definite direct dependence between variables, and where direction has important meaning.

4.7 Related Work

Some studies have used the MFE principle, as described in the previous chapter. Herskovits and Cooper [1990] first proposed a score for unrestricted network structure in the score and search approaches, where they claimed the score was based on the Maximum Entropy principle. However, in fact, they used the Maximum Likelihood principle with Bayesian smoothing, as shown by their use of an inverted minus sign of entropy. Subsequently, they proposed the Bayesian–Dirichlet score metric described in Chapter 2. In addition, Beal [2003] attempted to produce learning DAG structures that contain latent variables using variational free energies and variational Bayes EM (VBEM) methods [Jordan et al., 1999] in score and search procedures, which aimed at the rapid construction of DAGs, although learning DAGs with latent variables is generally expected to be time-consuming in marginalization processes. It was then reported that VBEM outperforms standard BIC [Schwarz, 1978] and CS [Cheeseman and Stutz, 1996] scores. More recently, Watanabe et al. [2009] analyzed the upper bound of variational free energy of biparticle Bayesian networks. These studies do not consider the role of temperature. Consequently, their studies assumed equivalently constant temperature, i.e., constant data size or a large sample limit if our “Date Temperature” assumption or similar concept is needed for learning with free energy.

Some researchers improved the constraint-based approach along with the PC algorithm. For example, Steck and Tresp [1999] proposed a necessary path condition (NPC), which is a simple but important improvement of the PC algorithm in both theoretical and practical views. The NPC improves the PC search step, where the NPC, when testing conditional independences on X and Y, detects only nodes on a currently existing path between X and Y, which reduces the search spaces of conditioning sets and errors of detecting false separator sets. Somewhat later, Cooper [1997] proposed a constraint-based method, which he called the Local Causal Discovery (LCD) algorithm. In fact, LCD is a specialization of the PC and Fast Causal Inference (FCI) [Spirtes et al., 1999] algorithms that use background knowledge related to an non-causal variable, which is a poorer language than those. However, LCD is simpler to implement and it is more computationally efficient than those, even for worst cases. Ramsey, Spirtes and Zhang introduced decomposition of the faithfulness condition into two parts—Adjacency-Faithfulness and Orientation-Faithfulness—and proposed a new algorithm that is a variation of the PC algorithm [Ramsey et al., 2006; Zhang, 2006]. Their problem consciousness was similar to ours: they attacked the problem of improving the accuracy of constraint-based algorithm. However, the approach explained herein differed from theirs. We attempt to solve the accuracy problem in constraint-based learning, which we consider results from the shortage of sample data, whereas they attempted to reduce mistakes of detecting the separator sets by introducing ambiguity into their PC algorithm, which they called the conservative PC (CPC) algorithm. Therefore, the approach explained herein is probably a more aggressive attempt to solve the problem.

4.8 Summary

For constraint-based learning Bayesian networks, which are used for causal discovery and which have a weak point of overfitting for insufficient samples in conditional independence (CI) testing, we proposed a method for its improvement. To do this, the minimum free energy (MFE) principle was used with the "Data Temperature" assumption. As a result, a new CI condition was derived, which can be used with a broad range of sample sizes. This CI method incorporates the maximum entropy and maximum likelihood principles and converges to the classical hypothesis tests in asymptotic regions. Through this work, we provide a unified framework of learning parameters and structures of BNs using MFE principle because we already proposed a parameter learning method of BNs in the previous chapter.

Results presented herein demonstrated the effectiveness of this novel method by embedding it in the well known PC algorithm. The results show that our method correctly identified the direction of the edges, at least in some simulation studies, better than the standard tests did, which is expected to be effective for causal discovery where the orientation of edges is significant. Furthermore, for five real-world datasets, our method shows better performance and identification of the direction for detection of true edges for small samples.

Chapter 5

Conclusions and Future Projects

5.1 Conclusions

In this dissertation, we have pursued the challenge of proposing a new learning methodology for a probabilistic graphical model—Bayesian networks (BNs) that deal with multivariate probability distributions in combination with directed acyclic graphs—in which we have taken particular note of potential discovery of causality from observational data. In attempting to consolidate effective causal discovering tools using BNs, we decomposed the causal discovery problems into two parts: (1) estimating parameters that have a strong influence on causal discovery because BNs entail probability distributions and statistical inference of the distributions; and (2) learning structures that represent causal modeling qualitatively. Practical effectiveness has been emphasized. Therefore, we sought to improve the weakness of accuracy that is dependent on overfitting to an insufficiently large amount of data. Because of the existence of this problem, the author wondered how the most effective information can be derived from available finite data. The Bayesian approach, which has attracted many researchers in broad domains of statistical science, apparently has difficulty at consistently deciding the optimal values of the entailed hyperparameter, under the condition of no prior knowledge, from both theoretical views (associated to noninformative priors) and practical views. Particularly, from recent studies of the latter views, because accuracies of learning are found to be highly sensitive to the values, it has been a critical issue that principled methods have not been found to decide the optimal values within a Bayesian framework. Consequently, some other principles are sought to let us obtain the maximum effective information, even from insufficient data.

This new approach comes from noticing that the Bayesian priors play a role of enlarging entropy depending on the available data size by adopting uniform hyperparameters when no prior knowledge is available. One might wonder whether any principle exists to decide optimal entropies of target probability distributions for learning, according to the training data size. Is there any principle-based method that generates a similar effect to that of Bayesian methods? Thermodynamics has provided attractive frameworks for which the minimum free energy (MFE) principle maintains a balance between minimum (internal) energy and maximum entropy. It seems to be a metaphor of conflict between maximum likelihood and maximum entropy in statistical science. Therefore, description of the learning methodology was addressed using the MFE principle. We sought to provide a unified framework of learning BNs including parameters and structures; then we obtained the results described in Chapter 3 and 4.

In Chapter 3, we first introduced and defined the free energy, internal energy, entropy so that the free energy represents an objective function with respect to the parameter learning of BNs. At that time, we proposed the "Data Temperature" assumption for making minimizing the free energies represent both the maximum likelihood and the maximum entropy principles with weights according to available data size, and providing a meaning of free energy, internal energy and temperature. Next, we proposed a simple "Data Temperature" model for leveraging the role of temperature with computational efficiency. The method showed superiority to the Bayesian Dirichlet parameter learning method with some recommended hyperparameters, and showed low sensitivity against selection of hyperparameters of our "Data Temperature" model. Furthermore, this new method has the advantage of not presenting difficulty of the inconsistency for selecting hyperparameters in theoretical and practical views as the Bayesian Dirichlet method has.

Chapter 4 presented an attempt to improve constraint-based structure learning BNs using the method developed in the previous chapter. Internal energy was defined for representing the classical hypothesis testing that is usually used in constraint-based structure learning of BNs. Then we derived a conditional independence condition for constraint-based learning attributable to the MFE principle, and proved that our method connects naturally to the classical G^2 statistics in asymptotic region, which means that our formation can be regarded as an extension of the classical statistics to that including the maximum entropy principle explicitly. Subsequently, it is shown that our method is effective for structure learning with small data size in simulation studies and experiments using real-world datasets.

From these studies, the main contributions of the dissertation to this field can be regarded as follows:

- We introduce an assumption of "Data Temperature" enable MFE principle to address both the maximal likelihood and the maximum entropy principles in a unified manner, and assign meaning to temperature for use of free energy in statistical sciences.

- An actually useful learning method is provided, which is effective even for a small sample size, which Bayesian network learning often suffers from.
- A unified learning methodology of parameters and structures of BNs is defined; it presents advantages over the standard information criterion such as AIC, BIC, and MDL, none which is applicable for parameter learning though those have similar effect on structure learning BNs.
- The methodology presented herein has advantages over the Bayesian–Dirichlet method, which presents controversial problems in theoretical aspects related to noninformative priors for no prior knowledge and which has difficulty finding optimal hyperparameters, despite the fact that learning accuracies are highly sensitive to the values.

5.2 Future Works

The framework of the MFE principle with the "Data Temperature" assumption will provide many further studies that we will attempt to undertake. Some studies are planned in addition to investigating and improving the present simple model of "Data Temperature", as explained below.

We intend to compare the MFE-EB method embedded in a PC algorithm with state-of-the-art structure learning algorithms such as Sparse Candidate [Friedman et al., 1999], Optimal Reinsertion [Moore and Wong, 2003], Greedy Equivalent Search [Chickering, 2002], TPDA [Cheng et al., 2002], and MMHC [Tsamardinos et al., 2006] in view of learning from insufficient data. As described in section 2.5.4, the learning performance of BNs using score-search methods is dependent on the selected hyperparameters, which means that the overfitting issue is important in learning BNs. Therefore, the issue is expected to be a main cause of underperformance of typical constraint-based methods against score-search methods. We alleviated it in the work described in the thesis, then the comparison seems to be an attractive research.

We intend to develop a method based on another main approach—a score-search approach combined with our MFE framework. Cowell [2001] claimed that conditional independence tests based and the score-search-based methods engender identical models of BNs under some conditions: (1) constructing BNs in a predictive sense, (2) assuming preferable to simple models, (3) assuming partial node ordering. However, we suspect that these conditions might hide the keys to understanding their mutual differences, which are to be investigated using our method.

Spirtes et al. [1995] extended their PC algorithm to deal with latent common causes and selection bias, which they call the FCI algorithm. Further studies must investigate the effectiveness of our method in the task of detecting latent variables using the FCI algorithm.

In this dissertation, we decomposed the poor accuracy issue of constraint-based algorithms into a pure shortage of data (theoretical aspects) and algorithmic aspects, and attacked the former using the MFE principle. We will address the latter: algorithmic aspects, for improving accuracy of causal discovery, including the latent variable models.

Additionally, the possibilities of justification of our manner of definitions of internal energies must be investigated. Especially, attention is devoted herein to the views of information theory and information geometry [Amari and Nagaoka, 2000].

One reason that Bayesian methods are broadly accepted in machine-learning domains is the fact that they can incorporate prior background knowledge, which avoids overfitting to training data. Prior knowledge is intended to be incorporated into our framework. Additionally, as presented in this dissertation, our method generates equivalent imaginary samples to those of Bayesian methods. Therefore the relation between Bayesian methods and our framework should be investigated further.

Many researchers have tackled learning problems using variational free energy method. However, no approach incorporating temperature is known. Consequently, it might be valuable to investigate whether this approach is effective for the method.

Appendix A

Some Lemmas and Proofs of the Theorems in Chapter 2

In this appendix, we describe the Lemmas related to the theorems in Chapter 2 and the proofs of the Theorems.

A.1 Theorem 2.1

The theorem 2.1 is provable from the following three lemmas.

Lemma A.1 (Verma and Pearl [1988]) *Presume P as a probability distribution for $\boldsymbol{V}$ and $\mathbb{G}$ be a DAG. Here, $(\mathbb{G}, P)$ satisfies the Markov condition if and only if, for every three mutually disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$, whenever $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent in P given $\boldsymbol{Z}$. That is, $(\mathbb{G}, P)$ satisfies the Markov condition if and only if*

$$Dsep^{\mathbb{G}}(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}) \Longrightarrow Ind(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}), \tag{A.1}$$

Is there any other conditional independence not required by d-separation? The answer is No, as proven by the following two lemmas. The definition 2.9 for it is provided in Chapter 2; then the lemmas are stated.

Lemma A.2 *Any conditional independence entailed by a DAG, based on the Markov condition, is equivalent to a conditional independence among disjoint sets of random variables.*

Proof. *The proof is shown in Neapolitan [2004].*

Lemma A.3 (Geiger and Pearl [1988b]) *Let $\mathbb{G} = (\boldsymbol{V}, E)$ be a DAG, and $\mathcal{P}$ be the set of all probability distributions P such that $(\mathbb{G}, P)$ satisfies the Markov condition. For every three mutually disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{V}$,*

$$Ind(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \text{ for all } P \in \mathcal{P} \Rightarrow Dsep^{\mathbb{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}).$$

A.2 Theorem 2.3

The following three lemmas are necessary for the proof of theorem 2.3.

Lemma A.4 (Verma and Pearl [1990]) *Let $\mathbb{G}$ be a DAG and X and $Y \in \boldsymbol{V}$. Then X and Y are adjacent in $\mathbb{G}$ if and only if they are not d-separated by any set in $\mathbb{G}$.*

Corollary A.1 (Verma and Pearl [1990]) *Let $\mathbb{G}$ be a DAG and X and $Y \in \boldsymbol{V}$. Then, if X and Y are d-separated by some set, they are d-separated either by the set consisting of the parents of X or the set consisting of the parents of Y.*

Lemma A.5 (Verma and Pearl [1990]) *Presuming that we have a DAG $\mathbb{G} = (\boldsymbol{V}, E)$ and an unlooped connection $X - Z - Y$, then the following are equivalent:*

- *$X - Z - Y$ is a v-structure.*
- *There exists a set not containing Z by which X and Y are d-separated.*
- *All sets containing Z do not d-separate X and Y.*

Lemma A.6 (Verma and Pearl [1990]) *If $\mathbb{G}_1$ and $\mathbb{G}_2$ are Markov equivalent, then X and Y are adjacent in $\mathbb{G}_1$ if and only if they are adjacent in $\mathbb{G}_2$: Markov equivalent DAGs have the same links (edges without direction).*

The following lemmas related to the patterns of Markov equivalent classes are derived from the corresponding lemmas for DAG

Lemma A.7 *Let $\mathbb{G}p$ be a DAG and X and Y be nodes in $\mathbb{G}p$. Then X and Y are adjacent in $\mathbb{G}p$ if and only if they are not d-separated by some set in $\mathbb{G}p$.*

The proof follows from Lemma A.4.

Lemma A.8 *Presuming a DAG pattern $\mathbb{G}p$ and an unlooped connection $X - Z - Y$, then the following are equivalent:*

- *$X - Z - Y$ is a v-structure.*
- *There exists a set not containing Z which d-separates X and Y.*
- *All sets containing Z do not d-separate X and Y.*

The proof follows from Lemma A.5.

A.3 Theorem 2.6

Here, we describe the proof of the theorem 2.6.

Proof. *Presuming that $\mathbb{G}p$ is the DAG pattern faithful to P. Then, because of Theorem 2.5, all and only the independencies in P are identified by d-separation in $\mathbb{G}p$, which are the d-separations in any DAG $\mathbb{G}$ in the equivalence class represented by $\mathbb{G}p$. Therefore, Condition (1) follows from Lemma A.4, and Condition (2) follows from Lemma* **??***.*

Arguing in the other direction, presume that Conditions (1) and (2) hold for $\mathbb{G}p$ and P. We have assumed P admits a faithful DAG representation. Therefore, there is some DAG pattern $\mathbb{G}p$' faithful to P. By what was just proven, we know that Conditions (1) and (2) also hold for $\mathbb{G}p$' and P. However, this means that any DAG $\mathbb{G}$ in the Markov equivalence class represented by $\mathbb{G}p$ must have the same links and same set of v-structures as any DAG $\mathbb{G}$' in the Markov equivalence class represented by $\mathbb{G}p$'. Theorem 2.3 therefore says $\mathbb{G}$ and $\mathbb{G}$' are in the same Markov equivalence class, which means that $\mathbb{G}p = \mathbb{G}p$'. □

A.4 Theorem 2.7

The proof is described as following:

Proof.

- *Rule1: Y becomes an unshielded collider if a directed edge from Z to Y exists. However unshielded colliders must have been recognized because of the assumption of the theorem. Therefore, Y should not be an unshielded collider and we should orient an edge from Y to Z.*

- *Rule2: If a directed edge from Z to X exists, then X, Y, and Z form a cyclic closed path, which fact contradicts the assumption that we have a hidden DAG structure. Therefore, the edge should be oriented from X to Z.*

- *Rule3: If a directed edge from Z to X exists, then for avoiding cyclic closed paths, we should orient edges from X to W and from X to Y, which both generate new v-structures that must have been recognized before applying the rules. Therefore, we should orient edges from X to Z.*

□

A.5 Theorem 2.8

The proof requires the following three lemmas.

Lemma A.9 (Spirtes et al. [2000]) *If the set of conditional independencies admits a faithful DAG representation, then the algorithms create a link between X and Y if and only if a link exists between X and Y in the DAG pattern $\mathbb{G}p$ containing the corresponding d-separations in this set.*

Proof. *The algorithms generate a link if and only if X and Y are not d-separated by any subset of $\boldsymbol{V}$, which is the case if and only if X and Y are adjacent in $\mathbb{G}p$ because of Lemma A.7.* □

Lemma A.10 (Spirtes et al. [2000]) *If the set of conditional independencies admit a faithful DAG representation, then any directed edge generated by the algorithms is a directed edge in the DAG pattern containing the corresponding d-separations in this set.*

Proof. *In step 4-1 of the algorithms, the fact that such edges must be directed as follows from Lemma A.8. In step 4-2 of the algorithms, the fact that such edges must be directed follows from Theorem 2.7.* □

Lemma A.11 (Meek [1995a]) *If the set of conditional independencies admits a faithful DAG representation, then all the directed edges, in the DAG pattern containing the corresponding d-separations, are directed by the algorithms.*

Proof. *Meek [1995a] proved the lemma.* □

Then the proof of the Theorem 2.8 follows from the preceding three lemmas.

Appendix B

An Example of Unfaithful DAG

In this appendix, we describe an example of distribution that is unfaithful to a DAG depicted in Neapolitan [2004].

Presuming that we have a DAG $\mathbb{G}$ with conditional probabilities in Fig. B.1, then, from the figure, one finds that $Dsep^{\mathbb{G}}(X; Z \,|\, Y)$.

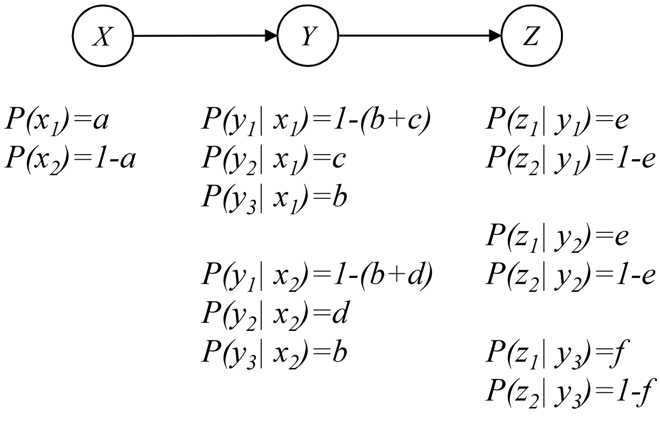


Figure B.1: Example of Bayesian network with assignments of conditional probability distribution unfaithful to the DAG.

However, it can be unfaithful to the DAG as shown by calculating the conditional probabilities. First, we derive the other expression of conditional independencies denoted in eq. 2.3. For an assumption $P(y, z) > 0$, equivalent transformations can be performed as follows.

$$\begin{aligned} & P(x,y\,|\,\boldsymbol{z}) = P(x\,|\,\boldsymbol{z})P(y\,|\,\boldsymbol{z}) \\ \Longleftrightarrow\quad & P(x,y\,|\,\boldsymbol{z})P(\boldsymbol{z}) = P(x\,|\,\boldsymbol{z})P(y\,|\,\boldsymbol{z})P(\boldsymbol{z}) \\ \Longleftrightarrow\quad & P(x,y,\boldsymbol{z}) = P(x\,|\,\boldsymbol{z})P(y,\boldsymbol{z}) \\ \Longleftrightarrow\quad & P(x\,|\,y,\boldsymbol{z})P(y,\boldsymbol{z}) = P(x\,|\,\boldsymbol{z})P(y,\boldsymbol{z}) \\ \Longleftrightarrow\quad & P(x\,|\,y,\boldsymbol{z}) = P(x\,|\,\boldsymbol{z}). \end{aligned} \tag{B.1}$$

Therefore, one can alternatively note that

$$Ind\,(X;Y|\boldsymbol{Z}),$$

if $\forall x, y, \boldsymbol{z}$ where $P(y,\boldsymbol{z}) > 0$,

$$P(x\,|\,y,\boldsymbol{z}) = P(x\,|\,\boldsymbol{z}).$$

Additionally and similarly, it is notable that

$$Ind\,(X;Z),$$

if $\forall x, z$ where $P(z) > 0$,

$$P(x\,|\,z) = P(x). \tag{B.2}$$

Therefore, whether the equation B.2 is consistent or not is investigated. If it consistent, then $Dsep^{\mathbb{G}}\,(X;Z\,|\,Y)$ and $Ind\,(X;Z)$, which shows unfaithful distribution to the DAG in Fig. B.1.

$$\begin{aligned} P(z_1\,|\,x_1) &= \sum_y P(z_1\,|\,y,x_1)P(y\,|\,x_1) \\ &= \sum_y P(z_1\,|\,y)P(y\,|\,x_1) \\ &= P(z_1\,|\,y_1)P(y_1\,|\,x_1) + P(z_1\,|\,y_2)P(y_2\,|\,x_1) + P(z_1\,|\,y_3)P(y_3\,|\,x_1) \\ &= e - be + bf. \end{aligned}$$

On the other hand,

$$\begin{aligned} P(z_1) &= \sum_{x,y} P(z_1\,|\,y)P(y\,|\,x)P(x) \\ &= P(z_1\,|\,y_1)P(y_1\,|\,x_1)P(x_1) + P(z_1\,|\,Y_2)P(y_2\,|\,x_1)P(x_1) \\ &\quad +P(z_1\,|\,Y_3)P(y_3\,|\,x_1)P(x_1) + P(z_1\,|\,y_1)P(y_1\,|\,x_2)P(x_2) \\ &\quad +P(z_1\,|\,Y_2)P(y_2\,|\,x_2)P(x_2) + P(z_1\,|\,Y_3)P(y_3\,|\,x_2)P(x_2) \\ &= e - be + bf \\ &= P(z_1\,|\,x_1). \end{aligned} \tag{B.3}$$

Similarly, it is shown that $P(z_2 \,|\, x_1)$, $P(z_1 \,|\, x_2)$ and $P(z_2 \,|\, x_2)$ present equal corresponding marginal probabilities.

Therefore, the assignments of conditional probabilities show unfaithfulness to the DAG.

Bibliography

B. Abramson, J. Brown, and R. L. Winkler. Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12:57–71, 1996.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, AC-19:716–723, 1974.

R. Ali, T. Richardson, P. Spirtes, and J. Zhang. Towards characterizing markov equivalence classes for directed acyclic graph models with latent variables. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 10–17, 2005.

T. Allen and R. Greiner. A model selection criteria for learning belief nets: An empirical study. In *Proc. of International Conference on Machine Learning (ICML-00)*, pages 1047–1054, 2000.

S. Amari and H. Nagaoka. *Method of Information Geometry.* Oxford University Press, New York, NY, 2000.

A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85:549–559, 1998.

M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference.* Ph.D. thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.

I. Beinlich, H. Suermondt, R. Chavez, and G. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proc. of European Conference on Artificial Intelligence in Medicine (AIME-89)*, pages 247–256, 1989.

J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

W. Buntine. Theory refinement on bayesian networks. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 52–61, 1991.

H. B. Callen. *Thermodynamics and An Introduction to Thermostatistics.* John Wiley & Sons, Hoboken, NJ, second edition, 1985.

P. Cheeseman and J. Stutz. Bayesian classification (Autoclass): Theory and results. In U. M. Fayyad, G. Piatesky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press, Menlo Park, CA, 1996.

J. Cheng, D. A. Bell, and W. Liu. An algorithm for Bayesian belief networks construction from data. In *International Workshop on Artificial Intelligence and Statistics (AISTATS-97)*, pages 101–108, 1997.

J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 101–108, 1999.

J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1-2):43–90, 2002.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 87–98, 1995.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Information Theory*, 14:462–467, 1968.

B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1:203–224, 1997.

G. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 86–94, 1991.

G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

G. F. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. In C. Glymour and G. F. Cooper, editors, *Computation, Causation, and Discovery*, pages 3–62. AAAI/MIT Press, Cambridge, MA, 1999.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Hoboken, NJ, second edition, 2006.

R. G. Cowell. Conditions under which conditional independence and scoring methods lead to identical selection of Bayesian network models. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 91–97, 2001.

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York, NY, 1999.

D. Dash and M. J. Druzdzel. Robust independence testing for constraint-based learning of causal structure. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 167–174, 2003.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems*, pages 479–485, 2001.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 1022–1027, 1993.

N. Friedman. The Bayesian structural EM algorithm. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 129–138, 1998.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

N. Friedman, I. Nachman, and D. Peér. Learning Bayesian network structure from massive datasets: The “sparse candidate” algorithm. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 206–215, 1999.

R. M. Fung and S. L. Crawford. Constructor: A system for the induction of probabilistic models. In *Proc. of National Conference on Artificial Intelligence (AAAI-90)*, pages 762–769, 1990.

D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence and qualitative independence. Technical Report R-97, UCLA, Cognitive Systems Laboratory, 1988a.

D. Geiger and J. Pearl. On the logic of causal models. In *Proc. of Workshop on Uncertainty in Artificial Intelligence (UAI-88)*, pages 3–14, 1988b.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. In *Proc. of National Conference on Artificial Intelligence (AAAI-02)*, pages 167–173, 2002.

D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proc. of International Conference on Machine Learning (ICML-04)*, pages 361–368, 2004.

D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995, revised June 1996.

D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

E. Herskovits and G. Cooper. Kutató: An entropy-driven system for construction of probabilistic expert systems from databases. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 117–125, 1990.

T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 289–296, 1999.

T. Isozaki, K. Horiuchi, and H. Kashimura. A new e-mail agent architecture based on semi-supervised Bayesian networks. In *Proc. of International Conference on Computational Intelligence for Modelling, Control and Automation (CIMCA-05)*, volume 1, pages 739–744, 2005.

T. Isozaki, N. Kato, and M. Ueno. Minimum free energies with “data temperature” for parameter learning of Bayesian networks. In *Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI-08)*, pages 371–378, 2008.

T. Isozaki, N. Kato, and M. Ueno. “Data temperature” in minimum free energies for parameter learning of Bayesian networks. *International Journal on Artificial Intelligence Tools*, 18(5):653–671, 2009.

T. Isozaki and M. Ueno. Minimum free energy principle for constraint-based learning Bayesian networks. In *Proc. of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009), Part1, LNAI5781*, pages 612–627, 2009.

A. Jensen and F. Jensen. Midas–an influence diagram for management of mildew in winter wheat. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 349–356, 1996.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs.* Springer-Verlag, New York, NY, 2007.

Y. Jing, V. Pavlović, and J. M. Rehg. Efficient discriminative learning Bayesian network classifier via boosted augmented naive Bayes. In *Proc. of International Conference on Machine Learning (ICML-05)*, pages 369–376, 2005.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. MIT Press, Cambridge MT, 1999.

J. H. Kim and J. Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *Proc. of International Joint Conference on Artificial Intelligence*, pages 190–193, 1983.

C. Kittel and H. Kroemer. *Thermal Physics.* W. H. Freeman, San Francisco, CA, 1980.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

D. Koller and N. Friedman. *Probabilistic Graphical Models.* MIT Press, Cambridge, MA, 2009.

K. Kristensen and I. A. Rasmussen. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33:197–217, 2002.

S. Kullback. *Information Theory and Statistics.* Dover Publications, Mineola, NY, 1968.

P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proc. of National Conference on Artificial Intelligence (AAAI-92)*, pages 223–228, 1992.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B*, 50(2):157–224, 1988.

Y. LeCun and F. J. Huang. Loss functions for discriminative training of energy-based models. In *Proc. of International Workshop on Artificial Intelligence and Statistics (AISTATS-05)*, pages 206–213, 2005.

E. L. Lehmann. *Testing Statistical Hypotheses.* John Wiley & Sons, second edition, 1986.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, 1995a.

C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–418, 1995b.

A. Moore and W. Wong. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In *Proc. of International Conference on Machine Learning (ICML-03)*, pages 552–559, 2003.

R. E. Neapolitan. *Learning Bayesian Networks.* Prentice Hall, Upper Saddle River, NJ, 2004.

D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

J. Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proc. of Cognitive Science Society*, pages 329–334, 1985.

J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–688, 1995.

J. Pearl. *Causality, models, reasoning, and inference.* Cambridge University Press, New York, NY, 2000.

J. Pearl, D. Geiger, and T. Verma. The logic of influence diagrams. *Kybernetica*, 25(2): 33–44, 1989.

F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. of Annual Meeting on Association for Computational Linguistics (ACL-93)*, pages 183–190, 1993.

J. Ramsey, P. Spirtes, and J. Zhang. Adjacency-faithfulness and conservative causal inference. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 401–408, 2006.

G. Rebane and J. Pearl. The recovery of causal poly-trees from statistical data. In *Workshop on Uncertainty in Artificial Intelligence (UAI-87)*, pages 222–228, 1987.

H. Reichenbach. *The Direction of Time.* Dover Publications, Mineola, NY, 1956. Republication of the work published by University of California Press, Berkely.

T. Richardson and P. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30: 962–1030, 2002.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* Springer-Verlag, New York, NY, second edition, 2007.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

B. Shen, X. Su, R. Greiner, P. Musilek, and C. Cheng. Discriminative parameter learning of general Bayesian network classifiers. In *Proc. of IEEE International Conference on Tools with Artificial Intelligence (ICTAI-03)*, pages 296–305, 2003.

T. Silander, P. Kontkane, and P. Myllymaki. On sensitivity of the map Bayesian network structure to the equivalent sample size parameter. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pages 360–367, 2007.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452, 2006.

M. Singh and M. Valtorta. An algorithm for the construction of Bayesian network structures from data. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 259–265, 1993.

D. J. Spiegelhalter and S. L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

P. Spirtes. Detecting causal relations in the presence of unmeasured variables. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-91)*, pages 392–397, 1991.

P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

P. Spirtes, C. Glymour, and R. Scheines. Causality from probability. In J. Tiles, G. McKee, and G. Dean, editors, *Evolving Knowledge in the Natural and Behavioral Sciences.* Pitman Publishing, London, 1990.

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

P. Spirtes and C. Meek. Learning Bayesian networks with discrete variables from data. In *Proc. of International Conference on Knowledge Discovery and Data Mining*, pages 294–299, 1995.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 499–506, 1995.

P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In C. Glymour and G. F. Cooper, editors, *Computation, Causation, and Discovery*, pages 211–252. AAAI/MIT Press, Cambridge, MA, 1999.

H. Steck. Learning the Bayesian network structure: Dirichlet prior versus data. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 511–518, 2008.

H. Steck and V. Tresp. Bayesian belief networks for data mining. In *Workshop Data Mining und Data Warehousing als Grundlage Moderner Entscheidungsunterst´utzender Systeme (DMDW-99)*, pages 145–154, 1999.

J. Suzuki. A construction of Bayesian networks from databases based on an MDL principle. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 266–273, 1993.

J. Suzuki. Learning Bayesian belief networks based on the minimum descricption length principle: An efficient algorithm using the B&B technique. In *Proc. of International Conference on Machine Learning (ICML-96)*, pages 462–470, 1996.

H. Tazaki. *Thermodynamics.* Baifukan, Tokyo, in Japanese, 2000.

J. Tian. A branch-and-bound algorithm for MDL learning Bayesian networks. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-00)*, pages 580–588, 2000.

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

N. Ueda and R. Nakano. Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems 7 (NIPS 7)*, pages 545–552, 1995.

M. Ueno. Learning likelihood-equivalence Bayesian networks using an empirical Bayesian approach. *Behaviormetrika*, 35(2):115–135, 2008.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proc. of Workshop on Uncertainty in Artificial Intelligence (UAI-88)*, pages 352–359, 1988.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-90)*, pages 220–227, 1990.

T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-92)*, pages 323–330, 1992.

K. Watanabe, M. Shiga, and S. Watanabe. Upper bound for variational free energy of Bayesian networks. *Machine Learning*, 75(2):199–215, 2009.

S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13:899–933, 2001.

N. Wermuth and S. Lauritzen. Graphical and recursive models for contingency tables. *Biometrika*, 70:537–552, 1983.

S. Yang and K. C. Chang. Comparison of score metrics for Bayesian network learning. *IEEE Trans. on Systems, Man and Cybernetics Part A: Systems and Humans*, 32(3): 419–428, 2002.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Information Theory*, 51(7):2282–2312, 2005.

J. Zhang. *Causal Inference and Reasoning in Causally Insufficient Systems*. Ph.D. thesis, Carnegie Mellon University, 2006.

J. Zhang and P. Spirtes. A transformational characterization of markov equivalence for directed acyclic graphs with latent variables. In *Proc. of Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 667–674, 2005.

List of Publications

Journal Papers

1. Takashi Isozaki, Noriji Kato, and Maomi Ueno. "Data Temperature" in Minimum Free Energies for Parameter Learning of Bayesian Networks, International Journal on Artificial Intelligence Tools, vol. 18, No. 5, pp. 653-671, 2009.

International Conferences (Refereed)

1. Takashi Isozaki, Noriji Kato, and Maomi Ueno. Minimum Free Energies with "Data Temperature" for Parameter Learning of Bayesian Networks, Proceedings of 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2008), vol. 1, pp. 371-378, 2008.
2. Takashi Isozaki and Maomi Ueno. Minimum Free Energy Principle for Constraint-Based Learning Bayesian Networks, Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2009), Part1, LNAI 5781, pp. 612-627, 2009.

Invited Talks and Papers

1. Takashi Isozaki. Learning Bayesian Networks using Minimum Free Energy Principle, Proceedings of 75th SIG-FPAI, pp. 37-42, 2009.

Other Refereed Conferences and Workshops

1. Takashi Isozaki and Noriji Kato. Robust Parameter Learning of Bayesian Networks by Using The Principle of Minimum Free Energy and "Data Temperature", Proceedings of Workshop on Information-Based Induction Science (IBIS2007) (in Japanese), pp. 61-66, 2007.