Université de Liège
Faculté des Sciences Appliquées
Département d'Électricité, Électronique et Informatique

Contributions to Bayesian Network Learning

Thèse présentée par
Vincent Auvray
en vue de l'obtention du titre de
Docteur en Sciences de l'Ingénieur

Année académique 2006-2007

Contributions to Bayesian Network Learning

Vincent Auvray

Acknowledgements

This dissertation is the fruit of several years of research that would not have been carried out without support. First and foremost, I would like to express my deep gratitude to my advisor Prof. Louis Wehenkel for the time and effort he invested in me. His trust in my abilities and his patience during the dissertation long gestation were really appreciated. I would also like to thank my colleagues and friends from the Montefiore Institute and the Bioinformatics unit of the GIGA-Research for creating a propitious work environment. In particular, I would like to mention Axel Legay, Pierre Geurts, and Sarah Hansoul whose help and feedback allowed me to clarify my ideas and greatly improve this work. I am also indebted to Prof. Ian Dobson who kindly welcomed me to the University of Wisconsin-Madison. Working under his supervision allowed me to grow as a researcher and eased my qualms about pursuing such a career. My parents and family also deserve my gratitude. I would like to acknowledge the sacrifices made by my wife Samantha Giessinger and thank her for her unwavering support. This dissertation is dedicated to her. Finally, I would like to thank the Belgian Fonds National de la Recherche Scientifique and the University of Liège for their crucial financial support.

Contents

Introduction

The probabilistic approach to modeling describes a domain with random variables and represents knowledge about this domain by a joint probability distribution on the variables. A probabilistic model allows to reason in uncertain conditions: after observing the values of certain variables, one can infer the conditional distribution of other variables and consequently make rational decisions. Many problems can be formulated within this framework as such inference problems. For example, in a medical domain, symptoms, genotype, and diseases can all be modeled as random variables. Then, a probabilistic model can help pose a diagnosis by computing the probability of a disease given the observation of some symptoms or the genotype.

A probabilistic model can be specified by a domain expert or learned from data. In the latter case, one assumes that some data generated by a process underlying the domain is accessible, and the goal of learning is to construct from the data a probabilistic model of this process. For example, in a medical domain, one may have medical records of patients.

Because of dimensionality problems, multivariate distributions — and thus probabilistic models — are difficult to represent, manipulate, and learn without first imposing some constraints. In particular, without assumption about its shape, a distribution on discrete random variables requires a number of parameters exponential in the number of variables in order to be represented.

Graphical probabilistic models explicitely encode with a graph certain structural properties of the joint distributions they represent. They possess two components: a graph and a parameter. The graph, or structure, encodes in a compact and intuitive way a set of marginal and conditional independence relations between the random variables of the domain. Such independence relations typically assert that each variable is directly influenced by only a few other variables, and they seem to hold in a wide variety of domains. For example, a disease may only be linked to a small set of genes, instead of the whole genome. The parameter specifies a distribution satisfying the independence assumptions of the structure via a parametrization map defined on a parameter space. With graphical models, two learning problems are typically considered: structure learning, where a structure whose independence assumptions hold in the data generating distribution is searched, and parameter learning, where a structure is given and a parameter mapped to the data generating distribution is searched. Learning the structure may be very informative about the domain as it may discover structural properties.

A Bayesian network is a special type of graphical probabilistic model whose graph is directed and acyclic and whose parameter specifies a conditional distribution for each variable given its parents in the graph. Bayesian networks allow to represent distributions compactly and to construct efficient inference and learning algorithms. Distinct directed acyclic graphs may sometimes encode the same set of independence relations and may thus be considered equivalent. To avoid considering equivalent structures, structure learning is sometimes formulated in terms of learning equivalence classes of structures, which are represented by so-called essential graphs in this dissertation.

Learning a Bayesian network structure (or an essential graph) is often cast as an optimisation problem. Given a set of candidate structures and some data, a scoring metric that ranks the structures is defined, and an optimal structure is searched, typically in a greedy way. Assuming that each candidate structure is assigned a set of neighboring structures, a greedy search algorithm explores the structure space iteratively by moving from a current structure to the highest scoring neighbor until a local optimum of the scoring metric is reached. Depending on the space of candidate structures, scoring metric, and initial structure, the greedy search may get stuck in a local, rather than global, optimum. A first contribution of this dissertation is an efficient algorithm that computes the neighborhood of essential graphs known as the inclusion boundary. The inclusion relation between sets of independence relations induces a partial order on sets of essential graphs (and sets of Bayesian network structures). Given a set of essential graphs, the union of the set of least upper bounds and the set of greatest lower bounds of an essential graph is its inclusion boundary. Under several technical assumptions, a greedy search algorithm using the inclusion boundary neighborhood possesses nice properties, and may even return a global optimum. Besides the actual computation of the inclusion boundary, this dissertation also demonstrates how to efficiently evaluate the score difference between an essential graph and one of its neighbors.

Like structure learning, parameter learning is often formulated as an optimization problem. In this case, an element of the parameter space maximizing an appropriate objective function such as the posterior parameter density or the data likelihood is searched. Sometimes, parameter learning is also interpreted as a projection problem that can be intuitively described and solved as follows. First, a distribution maximizing some unconstrained version of the original optimization problem is found. Then, this distribution is somehow projected onto the image of the parametrization map. Finally, a learned parameter is chosen among the parameters mapped to the projected distribution. To apply this procedure and design a suitable projection function, one may impose the constraint that a distribution already in the image of the parametrization map should be projected onto itself. Also, one should be able to compute the fibers of the parametrization map, i.e. the preimages of elements in the image of the parametrization map. This dissertation implements these well-known ideas to learn the parameters of the special class of Bayesian networks known as discrete Naives Bayes models with hidden class variable, and its contribution can be summarized as follows. First, assumptions implying that the preimage of a Naive Bayes distribution is finite are identified. Then, algorithms that compute the preimage of a distribution satisfying those assumptions by enumerating a finite superset of the preimage are proposed. Unfortunately, the superset may be very large, resulting in algorithms with high computational complexity. Finally, the algorithms computing fibers are converted into projection algorithms suitable for parameter learning by extending their applicability to distributions sufficiently close to the set of Naive Bayes distributions considered and ensuring their continuity. These projection algorithms should be considered preliminary: they share the high computational complexity of the fibers algorithms, have many parameters that need to be chosen, only work when the distribution to project is sufficiently close to the set of Naive Bayes distributions, and have not been extensively tested in practice. On the bright side, they also have nice asymptotic properties: under appropriate hypotheses and in the limit of a large dataset, they return an optimal parameter. Although the above description and the layout of the dissertation may not reflect it, we consider that our main contribution to the study of discrete Naive Bayes models with hidden class variable consists of the technical results and theorems behind our fibers and projection algorithms.

The dissertation is organized as follows. Background material is presented in the first two chapters. Chapter 1 introduces discrete and Gaussian Bayesian networks. Basic material such as parametric and implicit definition, independence relations encoded by a Bayesian network structure, dimension, and Bayesian networks with hidden variables is covered. More specialized notions such as inclusion and parameter optimality of a structure, equivalence of structures, and the inclusion relation between structures are also presented. Chapter 2 presents a Bayesian approach to structure and parameter learning. These two chapters do not constitute a review of the existing litterature on Bayesian networks. Instead, they form a coherent introduction to the topic, but do not provide much more than is necessary to develop the contributions of the dissertation. These contributions are gathered in the last two chapters. Chapter 3 develops efficient algorithms that construct the inclusion boundary of an essential graph and compute the difference in score between an essential graph and one element of its inclusion boundary. Chapter 4 discusses parameter learning in discrete Naive Bayes models with hidden class variable by computing fibers of the parametrization map. A few technical definitions and results are collected in the Appendix. An index, a list of theorems, a list of definitions, and a list of figures are also included at the end of the dissertation to help the reader sort through all the results, definitions, and notations.

Chapter 1

Bayesian Networks

1.1 Introduction

This chapter introduces Bayesian networks and Bayesian network models, laying the groundwork for subsequent chapters. A Bayesian network represents a probability density or distribution over a set of random variables with a graph and a set of conditional densities. A Bayesian network model is a set of densities represented by Bayesian networks sharing the same graph. The graph of a Bayesian network encodes a set of independence relations holding in the density represented.

Section 1.2 recalls some elements of probability theory and introduces notations used throughout the dissertation. Section 1.3 defines and illustrates Bayesian networks over discrete or continuous random variables. In subsequent chapters, our developments are restricted to discrete variables. Section 1.4 defines parametrically two particular classes of Bayesian network models: discrete and Gaussian. Section 1.5 defines and discusses the independence relations associated to a Bayesian network graph, leading to an alternative implicit definition of discrete and Gaussian Bayesian network models. Section 1.6 explores the link between the independence relations represented by graphs and their topological properties. Section 1.7 focuses on sets of densities obtained from Bayesian network models by marginalization. Section 1.8 provides notions useful to decide which Bayesian network model is a good candidate to represent a given density.

1.2 Elements of Probability Theory

This section introduces some elements of probability theory and the notations used in this dissertation. First, σ-fields and measures are defined. Then, random variables and vectors are presented. Most definitions are specialized to discrete variables or continuous variables with density w.r.t. Lebesgue measure. For additional details and comments, see [Bil79].

1.2.1 Measures

Definition 1. A class $\mathcal{F}$ of subsets of a set Ω is a *σ-field* (or *σ-algebra*) if

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F}$ implies $\Omega \setminus A \in \mathcal{F}$
3. $A_1, A_2, \dots \in \mathcal{F}$ implies $A_1 \cup A_2 \cup \dots \in \mathcal{F}$.

Example 1. The largest σ-field in Ω is the power set 2^Ω, i.e. the set of all subsets of Ω. The smallest σ-field in Ω is $\{\emptyset, \Omega\}$.

Remark 1. An intersection of σ-fields in Ω is a σ-field in Ω.

Definition 2. The *σ-field generated* by a class $\mathcal{A}$ of subsets of Ω is the intersection of all the σ-fields in Ω containing $\mathcal{A}$.

Example 2. If $a_i \leq b_i \in \mathbb{R}$ for $i \in \{1, \dots, k\}$, the set

$$\Big\{(x_1, \dots, x_k) \in \mathbb{R}^k \Big| a_i < x_i \leq b_i \text{ for } i \in \{1, \dots, k\}\Big\} \tag{1.1}$$

is a *bounded rectangle* in $\mathbb{R}^k$. The class $\mathcal{R}^k$ of *k-dimensional Borel sets* is the σ-field in $\mathbb{R}^k$ generated by the class of bounded rectangles in $\mathbb{R}^k$.

Definition 3. If $\mathcal{F}$ is a σ-field in Ω, the pair $(\Omega, \mathcal{F})$ is a *measurable space*.

Definition 4. A *measure* μ on a measurable space $(\Omega, \mathcal{F})$ is a function on $\mathcal{F}$ that satisfies

1. $\mu(A) \in [0, \infty]$ for $A \in \mathcal{F}$
2. $\mu(\emptyset) = 0$
3. if $A_1, A_2, \dots$ is a disjoint sequence of sets in $\mathcal{F}$, then

$$\mu(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mu(A_k). \tag{1.2}$$

Example 3. The *k-dimensional Lesbegue measure* λ_k is the unique measure on $\mathcal{R}^k$ such that

$$\lambda_k\Big(\{(x_1, \dots, x_k) \in \mathbb{R}^k \Big| a_i < x_i \leq b_i \text{ for } i \in \{1, \dots, k\}\}\Big) = \prod_{i=1}^{k} (b_i - a_i) \tag{1.3}$$

for all bounded rectangles in $\mathbb{R}^k$.

Definition 5. If μ is a measure on $(\Omega, \mathcal{F})$, a set $A \in \mathcal{F}$ is *negligible w.r.t* μ if $\mu(A) = 0$.

Definition 6. A measure P on $(\Omega, \mathcal{F})$ is a *probability measure* if $P(\Omega) = 1$.

Definition 7. If P is a probability measure on $(\Omega, \mathcal{F})$, the triple $(\Omega, \mathcal{F}, P)$ is a *probability space*.

Definition 8. A *support* of a probability measure P on $(\Omega, \mathcal{F})$ is a set $A \in \mathcal{F}$ such that $P(A) = 1$.

Definition 9. If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are measurable spaces, a function $f : \Omega \to \Omega'$ is *measurable* $\mathcal{F}/\mathcal{F}'$ if $\{\omega \in \Omega | f(\omega) \in A'\} \in \mathcal{F}$ for every $A' \in \mathcal{F}'$.

Definition 10. If $(\Omega, \mathcal{F}, P)$ is a probability space, $(\Omega', \mathcal{F}')$ is a measurable space, $f : \Omega \to \Omega'$ is measurable $\mathcal{F}/\mathcal{F}'$, and $A \in \mathcal{F}'$, the proposition $f \in A$ *holds with probability one* if the set $\{\omega \in \Omega | f(\omega) \notin A\}$ is negligible w.r.t. P.

Example 4. If $(\Omega, \mathcal{F}, P)$ is a probability space and $f : \Omega \to \mathbb{R}$ and $g : \Omega \to \mathbb{R}$ are measurable $\mathcal{F}/\mathcal{R}^1$, then $\{\omega \in \Omega | f(\omega) \neq g(\omega)\} \in \mathcal{F}$ and we say that f and g are *equal with probability one* if $\{\omega \in \Omega | f(\omega) \neq g(\omega)\}$ is negligible w.r.t. P.

1.2.2 Random Variables

Random variables may be defined as follows.

Definition 11. If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are measurable spaces, a *random variable* X is a function $X : \Omega \to \Omega'$ measurable $\mathcal{F}/\mathcal{F}'$.

Sometimes, the definition of random variables assumes that a probability measure P on $(\Omega, \mathcal{F})$ is given. As Proposition 1.1 will show, such a measure P induces a distribution for X. In this dissertation, random variables are defined without reference to a probability measure because sets of distributions for a fixed random variable are manipulated.

Definition 12. If $X : \Omega \to \Omega'$ is a random variable, the set $\mathcal{X} = \Omega'$ is the set of *possible values* (or *states*) of X.

Remark 2. A random variable is denoted by an upper-case token (e.g. X, X_i). A possible value is denoted by a lower-case token (e.g. x, x_i).

This dissertation only deals with real random variables, discrete random variables, real random vectors, and discrete random vectors.

Definition 13. If $(\Omega, \mathcal{F})$ is a measurable space, a *real random variable* is a function $X : \Omega \to \mathbb{R}$ measurable $\mathcal{F}/\mathcal{R}^1$.

Definition 14. A *k-dimensional real random vector* $(X_1, \dots, X_k)$ is a k-tuple of real random variables defined on the same set Ω.

Remark 3. If $(X_1, \dots, X_k)$ is a real random vector on Ω, then $X : \Omega \to \mathbb{R}^k : \omega \mapsto (X_1(\omega), \dots, X_k(\omega))$ is a random variable measurable $\mathcal{F}/\mathcal{R}^k$.

Definition 15. If $\mathcal{X}$ is a countable set and $(\Omega, \mathcal{F})$ is a measurable space, a *discrete random variable* is a function $X : \Omega \to \mathcal{X}$ measurable $\mathcal{F}/2^{\mathcal{X}}$.

Definition 16. A *k-dimensional discrete random vector* $(X_1, \dots, X_k)$ is a k-tuple of discrete random variables defined on the same set Ω.

Remark 4. If $(X_1, \dots, X_k)$ is a discrete random vector on Ω, then $X : \Omega \to \mathcal{X}_1 \times \dots \times \mathcal{X}_k : \omega \mapsto (X_1(\omega), \dots, X_k(\omega))$ is a discrete random variable.

Remark 5. The properties of random vectors that are defined in this dissertation do not depend on the precise order of their components. A finite set of random variables is thus often considered a random vector and vice-versa.

Distributions

Definition 17. If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are measurable spaces, a *distribution* of a random variable $X : \Omega \to \Omega'$ is a probability measure on $\mathcal{F}'$.

A probability measure on $(\Omega, \mathcal{F})$ induces a distribution of X.

Proposition 1.1. *If $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ are measurable spaces, $X : \Omega \to \Omega'$ is a random variable, and P is a probability measure on $(\Omega, \mathcal{F})$, then*

$$\mu(A) = P\Big(\{\omega \in \Omega \Big| X(\omega) \in A\}\Big), \quad A \in \mathcal{F}' \tag{1.4}$$

is a distribution of X.

By Remarks 3, 4, and 5, vectors and sets of discrete or real random variables may be considered random variables. Hence, the notion of distribution is also defined for them.

Definition 18. Let $f : \mathbb{R}^k \to \mathbb{R}$ be a nonnegative and measurable $\mathcal{R}^k/\mathcal{R}^1$ function. A real random vector $(X_1, \dots, X_k)$ with distribution μ has *density f (w.r.t. Lebesgue measure)* if

$$\mu(A) = \int_A f(x_1, \dots, x_k) dx_1 \dots dx_k, \quad A \in \mathcal{R}^k. \tag{1.5}$$

Definition 19. A real random vector X with distribution μ is *continuous* if μ has a density w.r.t. Lebesgue measure.

Remark 6. A distribution μ of a discrete random vector $X = (X_1, \dots, X_k)$ is completely determined by the values $\mu(\{x\}), x \in \mathcal{X}$:

$$\mu(A) = \sum_{x \in A} \mu(\{x\}), \quad A \in 2^{\mathcal{X}}. \tag{1.6}$$

In the sequel, $\mu(\{x\})$ is denoted $\mu(x)$ to simplify notations.

Marginal Distributions

Definition 20. If $X = \{X_1, \dots, X_k\}$ is a set of random variables with distribution μ_X and $X' \subseteq X$, the *marginal distribution* $\mu_{X'}$ is the distribution of X' defined by

$$\mu_{X'}(A) = \mu_X(A \times_{X_i \in (X \setminus X')} \mathcal{X}_i). \tag{1.7}$$

Remark 7. To simplify notations, the distributions μ_X and $\mu_{X'}$ are usually denoted by the same symbol, e.g. μ.

Remark 8. If $X = \{X_1, \dots, X_j, X_{j+1}, \dots, X_k\}$ is a set of discrete random variables with distribution μ, the marginal distribution of $X' = \{X_1, \dots, X_j\}$ is completely specified by

$$\mu(x_1, \dots, x_j) = \sum_{(x_{j+1},\dots,x_k) \in \mathcal{X}_{j+1} \times \dots \times \mathcal{X}_k} \mu(x_1, \dots, x_k), \quad (x_1, \dots, x_j) \in \mathcal{X}'. \tag{1.8}$$

Definition 21. If $X = \{X_1, \dots, X_j, X_{j+1}, \dots, X_k\}$ is a set of continuous real random variables with density f_X, the *marginal density* $f_{X'}$ of X' is defined by

$$f_{X'}(x_1, \dots, x_j) = \int_{\mathbb{R}^{k-j}} f_X(x_1, \dots, x_k) dx_{j+1} \dots dx_k, \quad (x_1, \dots, x_j) \in \mathbb{R}^j. \tag{1.9}$$

Remark 9. A density and its marginal densities are usually denoted by the same symbol.

Remark 10. The marginal density of X' is a density for the marginal distribution of X'.

Conditional Distributions

Definition 22. If X and Y are disjoint subsets of a set of discrete random variables, μ is a distribution of Z and $y \in \mathcal{Y}$ satisfies $\mu(y) \neq 0$, the *conditional distribution of X given y* is the probability measure $\mu(\cdot|y)$ on $2^{\mathcal{X}}$ specified by

$$\mu(x|y) = \frac{\mu(x, y)}{\mu(y)}, \quad x \in \mathcal{X}. \tag{1.10}$$

Definition 23. If X and Y are disjoint subsets of a set of continuous real random variables with density f and $y \in \mathbb{R}^{|Y|}$ satisfies $f(y) \neq 0$, the *conditional density of X given y* is the function $f(\cdot|y)$ defined on $\mathbb{R}^{|X|}$ by

$$f(x|y) = \frac{f(x, y)}{f(y)}, \quad x \in \mathbb{R}^{|X|}. \tag{1.11}$$

Independence

Definition 24. If $X_1, \dots, X_k$ are disjoint subsets of a set of discrete random variables with distribution μ, the sets $X_1, \dots, X_k$ are *(marginally) independent* if

$$\mu(x_1, \dots, x_k) = \mu(x_1) \dots \mu(x_k) \tag{1.12}$$

for all $(x_1, \dots, x_k) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_k$.

Definition 25. If $X_1, \dots, X_k$ are disjoint subsets of a set of continuous random variables with density f, the sets $X_1, \dots, X_k$ are *(marginally) independent* if

$$f(x_1, \dots, x_k) = f(x_1) \dots f(x_k) \tag{1.13}$$

for all $(x_1, \dots, x_k) \in \mathbb{R}^{|X_1|} \times \dots \times \mathbb{R}^{|X_k|}$.

Definition 26. If $X_1, \dots, X_k, Y$ are disjoint subsets of a set of discrete random variables with distribution μ, the sets $X_1, \dots, X_k$ are *(conditionally) independent given* Y if

$$\mu(x_1, \dots, x_k | y) = \mu(x_1 | y) \dots \mu(x_k | y) \tag{1.14}$$

for all $(x_1, \dots, x_k) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_k$ and all $y \in \mathcal{Y}$ satisfying $\mu(y) \neq 0$.

Definition 27. If $X_1, \dots, X_k, Y$ are disjoint subsets of a set of continuous random variables with density f, the sets $X_1, \dots, X_k$ are *(conditionally) independent given* Y if

$$f(x_1, \dots, x_k | y) = f(x_1 | y) \dots f(x_k | y) \tag{1.15}$$

for all $(x_1, \dots, x_k) \in \mathbb{R}^{|X_1|} \times \dots \times \mathbb{R}^{|X_k|}$ and all $y \in \mathbb{R}^{|Y|}$ satisfying $f(y) \neq 0$.

Remark 11. The conditional independence of X and Y given Z is denoted $X \perp Y | Z$. The marginal independence of X and Y is denoted $X \perp Y$ (or sometimes $X \perp Y | \emptyset$). With this notation, arbitrary independence relations between sets of discrete or continuous variables can be expressed:

- $X_1, \dots, X_k$ are marginally independent if, and only if,

  $$X_i \perp (X_{i+1} \cup \dots \cup X_k) \tag{1.16}$$

  for $i \in \{1, \dots, k-1\}$;

- $X_1, \dots, X_k$ are conditionally independent given Y if, and only if,

  $$X_i \perp (X_{i+1} \cup \dots \cup X_k) | Y \tag{1.17}$$

  for $i \in \{1, \dots, k-1\}$.

Expected Value and the Strong Law of Large Numbers

Definition 28. The *expected* (or *mean*) value $\langle X \rangle$ of a real and discrete random variable X with distribution μ is

$$\langle X \rangle = \sum_{x \in X} x\mu(x). \tag{1.18}$$

Definition 29. The *expected* (or *mean*) value $\langle X \rangle$ of a continuous real random variable X with density f is

$$\langle X \rangle = \int_{-\infty}^{\infty} x f(x) dx. \tag{1.19}$$

Remark 12. If the sum in (1.18) does not converge or the integral in (1.19) does not exist, the expected value is not defined.

Khinchine's version of the strong law of large numbers states the following (from [Bil79]).

Theorem 1.2. *Suppose that $X_1, X_2, \ldots$ is a sequence of independent and identically distributed real random variables whose common expected value exists and is equal to m. We have*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} X_i = m \tag{1.20}$$

with probability one.

1.3 Bayesian Networks

This section defines and illustrates Bayesian networks. First, we introduce elementary graphical notions and notations.

Definition 30. A *graph* is a pair (V, E) where V is a non-empty and *finite* set of *vertices* and E is a subset of $(V \times V) \setminus \{(a, a) | a \in V\}$.

Definition 31. A graph $G = (V, E)$ has

- an *edge* between a and b, denoted $a \cdots b \in G$, if $(a, b) \in E$ or $(b, a) \in E$;
- an *undirected edge* (or *line*) between a and b, denoted $a - b \in G$, if $(a, b) \in E$ and $(b, a) \in E$;
- a *directed edge* (or *arrow*) from a to b, denoted $a \to b \in G$, if $(a, b) \in E$ and $(b, a) \notin E$.

Definition 32. If $G = (V, E)$ is a graph, the set $pa_G(v)$ of *parents* of $v \in V$ is

$$pa_G(v) = \{u \in V \big| u \to v \in G\}. \tag{1.21}$$

Remark 13. If G is determined by the context, $pa_G(v)$ is simply denoted $pa(v)$.

Definition 33. If $G = (V, E)$ is a graph, a *path* is a sequence $v_0, \dots, v_n$ of *distinct* vertices such that $v_i - v_{i+1} \in G$ or $v_i \to v_{i+1} \in G$ for all $i \in \{0, \dots, n-1\}$.

Definition 34. If $G = (V, E)$ is a graph, a path $v_0, \dots, v_n$ is *directed* if $v_i \to v_{i+1} \in G$ for a least one $i \in \{0, \dots, n-1\}$. Otherwise, it is *undirected.*

Definition 35. If $G = (V, E)$ is a graph, a *cycle of length n* is a path $v_0, \dots, v_n$ with the modification that $v_0 = v_n$.

Definition 36. If $G = (V, E)$ is a graph, a cycle $v_0, \dots, v_n$ is *directed* if $v_i \to v_{i+1} \in G$ for a least one $i \in \{0, \dots, n-1\}$.

Bayesian networks are defined using a special class of graphs: directed acyclic graphs. Other classes, such as undirected graphs and chain graphs, will be encountered further in this dissertation.

Definition 37. A *directed acyclic graph* (DAG) is a graph without line or cycle.

Definition 38. If $X = \{X_v\}_{v \in V}$ is a set of random variables indexed by a set V, $x = (x_v)_{v \in V} \in \mathcal{X}$, and $U \subseteq V$, let $X_U = \{X_v\}_{v \in U}$ and let $x_U = (x_v)_{v \in U} \in \mathcal{X}_U$.

In the above definition, a singleton $U = \{u\} \subseteq V$ is often denoted by u. Therefore, if $x = (x_v)_{v \in V} \in \mathcal{X}$ and $u \in V$, then x_u denotes the value $x_v \in \mathcal{X}_v$ such that $v = u$.

The recursive factorization property connects probability theory and graph theory. For sets of discrete variables or sets of continuous variables, it is defined as follows (see [CDLS99] for a more general formulation).

Definition 39 (*Recursive factorization for discrete variables*). Let X be a finite and non-empty set of discrete random variables, and let D be a DAG whose vertex set V is in bijection with X. A probability distribution P of X *factorizes recursively according to D* if there exist non-negative functions $k_v(\cdot, \cdot)$, $v \in V$ defined on $\mathcal{X}_v \times \mathcal{X}_{pa(v)}$ such that

$$\sum_{x_v \in \mathcal{X}_v} k_v(x_v, x_{pa(v)}) = 1 \tag{1.22}$$

for all $x_{pa(v)} \in \mathcal{X}_{pa(v)}$ and

$$P(x) = \prod_{v \in V} k_v(x_v, x_{pa(v)}), \quad x \in \mathcal{X}. \tag{1.23}$$

Remark 14. If a distribution P for discrete variables factorizes recursively, then $k_v(x_v, x_{pa(v)}) = P(x_v | x_{pa(v)})$ for $x_v \in \mathcal{X}_v$ and $x_{pa(v)} \in \mathcal{X}_{pa(v)}$ such that $P(x_{pa(v)}) \neq 0$.

Definition 40 (*Recursive factorization for continuous variables*). Let X be a finite and non-empty set of continuous random variables, and let D be a DAG whose vertex set V is in bijection with X. A probability distribution P of X *factorizes recursively according to* D if there exist non-negative functions $k_v(\cdot,\cdot)$, $v \in V$ defined on $\mathbb{R} \times \mathbb{R}^{|pa(v)|}$ such that

$$\int_{-\infty}^{\infty} k_v(x_v, x_{pa(v)}) dx_v = 1 \tag{1.24}$$

for all $x_{pa(v)} \in \mathbb{R}^{|pa(v)|}$ and P has density p given by

$$p(x) = \prod_{v \in V} k_v(x_v, x_{pa(v)}), \quad x \in \mathbb{R}^{|X|}. \tag{1.25}$$

Remark 15. If a distribution P with density p for continuous variables factorizes recursively, then, with probability one, $k_v(x_v, x_{pa(v)}) = p(x_v|x_{pa(v)})$ for $x_v \in \mathbb{R}$ and $x_{pa(v)} \in \mathbb{R}^{|pa(v)|}$ such that $p(x_{pa(v)}) \neq 0$.

Remark 16. In this dissertation, a (conditional) distribution of discrete random variables is often referred to as a (conditional) density. This simplifies notations and allows to simultaneously define notions and properties for discrete or continuous variables.

A Bayesian network is a graphical representation of a density that factorizes recursively. It is defined as follows.

Definition 41 (*Bayesian network*). Let X be a non-empty and finite set of random variables, let D be a DAG whose vertex set V is in bijection with X, and let θ be a set of conditional densities $\{p(x_v|x_{pa(v)})\}_{v \in V}$. A *Bayesian network* (BN) B is a pair (D, θ), where D is the *structure* and θ are the *parameters*, that represents the density

$$p_B(x) = \prod_{v \in V} p(x_v|x_{pa(v)}), \quad x \in \mathcal{X}. \tag{1.26}$$

Remark 17. Sometimes, the structure of a Bayesian network has a causal interpretation where an arrow $u \to v \in D$ means that X_u is a direct cause of X_v (see [Nea03] for an introduction). In this dissertation, structures are not interpreted causally.

Remark 18. Often, the sets X and V are not distinguished because of the bijection between them. For example, a Bayesian network structure is said to be over a set X of random variables.

Let us present two examples of Bayesian networks. The first is taken from [Nea03]. The second is adapted from [Pea88].

Example 5. Let $X = \{H, B, L, F, C\}$ be a set of binary random variables indicating whether a patient has a smoking history (H), bronchitis (B), lung cancer (L), experiences fatigue (F), and whether an X-ray of the patient's chest tests positive for

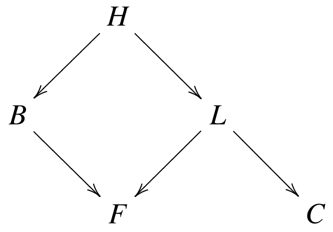


Figure 1.1: A Bayesian network structure whose vertex set is identified with $\{H, B, L, F, C\}$

$H:$	$p(H = t)$	$= 0.2$	$F:$	$p(F = t \mid B = t, L = t)$	$= 0.75$
				$p(F = t \mid B = t, L = f)$	$= 0.1$
$B:$	$p(B = t \mid H = t)$	$= 0.25$		$p(F = t \mid B = f, L = t)$	$= 0.5$
	$p(B = t \mid H = f)$	$= 0.05$		$p(F = t \mid B = f, L = f)$	$= 0.05$
$L:$	$p(L = t \mid H = t)$	$= 0.003$	$C:$	$p(C = t \mid L = t)$	$= 0.6$
	$p(L = t \mid H = f)$	$= 0.00005$		$p(C = t \mid L = f)$	$= 0.02$

Table 1.1: Parameters associated to the structure given in Figure 1.1

lung cancer (C). The structure given in Figure 1.1 and the conditional distributions given in Table 1.1 specify together a Bayesian network B and thus a probability distribution p_B given by (1.26). For example, we have $p_B(H = f, B = f, L = t, F = t, C = t) = p(H = f)p(B = f|H = f)p(L = t|H = f)p(F = t|B = f, L = t)p(C = t|L = t) = 0.8 \times 0.95 \times 0.00005 \times 0.5 \times 0.6 = 1.14 \times 10^{-5}$.

Example 6. Let $X = \{P, M, W, D_1, D_2\}$ be a set of real variables where P measures the production cost of a given car, M the marketing cost, W the wholesale price, D_1 the asking price of a first car dealer, and D_2 the asking price of a second dealer. Figure 1.2 shows a structure over X. To obtain a Bayesian network, we may suppose that each conditional density $p(x_v|x_{pa(v)})$, $v \in V$ is Gaussian (see (A.1) in Appendix A) with a mean that is a linear combination of the values of the parent

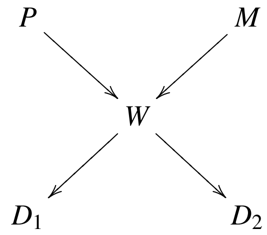


Figure 1.2: A Bayesian network structure over $X = \{P, M, W, D_1, D_2\}$

variables and a constant variance, that is

$$p(x_v|x_{pa(v)}) = \mathcal{N}(x_v|\alpha_v + \sum_{u\in pa(v)} \beta_{v,u}x_u, \sigma_v^2). \tag{1.27}$$

For example, we can specify $p(D_1|W = w) = \mathcal{N}(D_1|1000 + w, 300^2)$ and $p(W|P = p, M = m) = \mathcal{N}(W|800 + 1.2p + m, 500^2)$.

1.4 Parametric Bayesian Network Models

This section defines a special class of statistical models: *Bayesian network models*. In particular, it defines parametrically *discrete* and *Gaussian* Bayesian network models and their dimension.

1.4.1 Statistical Models

Definition 42. A *statistical model* (or *family*) $\mathcal{M}$ for a set X of random variables is a set of densities for X.

The notion of probabilistic model used in the introduction and the above notion of statistical model are different. In the sequel, the word model refers to a *statistical* model.

A statistical model $\mathcal{M}$ can be specified *parametrically* as the image $\mathcal{M} = f(\Theta)$ of a *parameter space* Θ through a *parametrization map* f defined on Θ. For example, the set of strictly positive distributions of a discrete random variables can be described parametrically with the following notions.

Definition 43. If X is a discrete random variable with finite $\mathcal{X}$, let

$$S_X^+ = \left\{(p_x)_{x\in\mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|} \middle| \sum_{x\in\mathcal{X}} p_x = 1, (\forall x \in \mathcal{X} : p_x > 0)\right\} \tag{1.28}$$

and let f_X be the function defined on S_X^+ by $f_X((p_x)_{x\in\mathcal{X}}) = p$ with

$$p(x) = p_x, \quad x \in \mathcal{X}. \tag{1.29}$$

Remark 19. The function f_X is injective. Also, a vector $(p_x)_{x\in\mathcal{X}} \in S_X^+$ has only $|\mathcal{X}| - 1$ independent components since $\sum_{x\in\mathcal{X}} p_x = 1$. In fact, the set S_X^+ is a smooth manifold in $\mathbb{R}^{|\mathcal{X}|}$ of dimension $|\mathcal{X}| - 1$.

Example 7. Exponential families are defined parametrically in Appendix A.

1.4.2 Bayesian Network Models

Informally, a Bayesian network model with structure D is a set of densities represented by Bayesian networks sharing the structure D.

Definition 44 (*Bayesian network model*). If X is a non-empty and finite set of random variables and D is a DAG whose vertex set V is in bijection with X, a *Bayesian network model with structure* D is a set of densities for X factorizing recursively according to D.

Following the examples of Section 1.3, let us introduce two important classes of Bayesian network models.

Discrete Bayesian Network Models

This section defines parametrically the class of discrete Bayesian network models. All the discrete random variables considered are supposed to have a finite set of possible values. The parameter space is defined as follows.

Definition 45. If X is a non-empty and finite set of discrete random variables and D is a DAG whose vertex set V is in bijection with X, let

$$\Theta_{d,D} = \times_{v \in V} (S^{+}_{X_v})^{|\mathcal{X}_{pa(v)}|}. \tag{1.30}$$

Remark 20. The set $\Theta_{d,D}$ is a smooth manifold of dimension $\sum_{v \in V}(|\mathcal{X}_v| - 1)|\mathcal{X}_{pa(v)}|$.

The parametrization map is defined as follows.

Definition 46. If X is a non-empty and finite set of discrete random variables and D is a DAG whose vertex set V is in bijection with X, let $f_{d,D}$ be the function defined on $\Theta_{d,D}$ by

$$f_{d,D}\Big(\Big(((\theta_{x_v}^{X_v, x_{pa(v)}})_{x_v \in \mathcal{X}_v})_{x_{pa(v)} \in \mathcal{X}_{pa(v)}}\Big)_{v \in V}\Big) = p \tag{1.31}$$

with

$$p(x) = \prod_{v \in V} \theta_{x_v}^{X_v, x_{pa(v)}}, \quad x \in \mathcal{X}. \tag{1.32}$$

Definition 47 (*Discrete Bayesian network model*). If X is a non-empty and finite set of discrete random variables and D is a DAG whose vertex set V is in bijection with X, the *discrete Bayesian network model* $\mathcal{M}_d(D)$ with structure D is the statistical model $f_{d,D}(\Theta_{d,D})$.

Remark 21. The positivity requirement on the local conditional distributions excludes functional relations among the variables. However, the probabilities can be as close to 0 as desired. On the other hand, this requirement also ensures that the parametrization map is injective. Indeed, if $q = f(\theta)$, we have

$$\theta_{x_v}^{X_v, x_{pa(v)}} = \frac{q(x_v, x_{pa(v)})}{q(x_{pa(v)})}. \tag{1.33}$$

Gaussian Bayesian Network Models

This section defines parametrically the class of Gaussian Bayesian network models. Let $\mathbb{R}_{>0}$ denote the set of strictly positive real numbers. The parameter space is defined as follows.

Definition 48. If X is a non-empty and finite set of real random variables and D is a DAG whose vertex set V is in bijection with X, let

$$\Theta_{g,D} = \times_{v \in V} \left(\mathbb{R}_{>0} \times \mathbb{R}^{|pa(v)|+1} \right). \tag{1.34}$$

Remark 22. The set $\Theta_{g,D}$ is a smooth manifold of dimension $\sum_{v \in V}(|pa(v)| + 2)$.

The parametrization map is defined as follows.

Definition 49. If X is a non-empty and finite set of real random variables and D is a DAG whose vertex set V is in bijection with X, let $f_{g,D}$ be the function defined on $\Theta_{g,D}$ by

$$f_{g,D}\Big((\sigma_v^2, \alpha_v, \beta_{v,1}, \ldots, \beta_{v,|pa(v)|})_{v \in V}\Big) = p \tag{1.35}$$

with

$$p(x) = \prod_{v \in V} \mathcal{N}(x_v | \alpha_v + \sum_{u \in pa(v)} \beta_{v,u} x_u, \sigma_v^2), \quad x \in \mathbb{R}^{|X|}. \tag{1.36}$$

Definition 50 (*Gaussian Bayesian network model*). If X is a non-empty and finite set of real random variables and D is a DAG whose vertex set V is in bijection with X, the *Gaussian Bayesian network model* $\mathcal{M}_g(D)$ with structure D is the statistical model $f_{g,D}(\Theta_{g,D})$.

Remark 23. A Gaussian Bayesian network model represents a multivariate Gaussian density (see (A.2) in Appendix A and [SK89]). Moreover, the parametrization map is injective.

There are many other interesting classes of BN models. With discrete variables, it is possible to constrain the conditional distributions. For example, parameters can be shared (see [NMR06]). Let us also mention *sigmoid* BN models (see [Nea92]) or *noisy-OR* BN models (see [Pea88]). With continuous variables, variants of Gaussian BN models can be defined by choosing a mean that does not depend linearly on the parents, for example by using a sigmoid map. Finally, some BN models, such as *conditional Gaussian* BN models (see [CDLS99]), mix discrete and continuous variables.

1.4.3 Dimension

This section defines the dimension of discrete and Gaussian Bayesian network models. The dimension is a geometric property that is used for learning (see Section 2.5.1) and that intuitively measures the size or complexity of the model.

Discrete Bayesian Network Models

The dimension of a discrete Bayesian network model is defined as follows.

Definition 51. The *dimension* $d(\mathcal{M}_d(D))$ of a discrete BN model $\mathcal{M}_d(D)$ is

$$d(\mathcal{M}_d(D)) = \sum_{v \in V} (|\mathcal{X}_v| - 1)|\mathcal{X}_{pa(v)}|. \tag{1.37}$$

The dimension has multiple interpretations. First, it is the dimension of the parameter space $\Theta_{d,D}$. Second, the following proposition holds (see [GHKM01] and Appendix A).

Proposition 1.3. *A discrete Bayesian network model $\mathcal{M}_d(D)$ is a curved exponential model of dimension $d(\mathcal{M}_d(D))$.*

If X is fixed, the dimension varies with the structure as follows. The minimal dimension is $\sum_{v \in V}(|\mathcal{X}_v| - 1)$ for the structure without any arrow. From there, the dimension increases exponentially with the number of parents of each variable until it reaches the maximal dimension $|\mathcal{X}| - 1$ for a structure where all the vertices are connected by an arrow.

Remark 24. By (1.29), the vector representation of a (strictly positive) distribution p of X uses $|\mathcal{X}| - 1$ real parameters, which equals the maximal dimension. Hence, a distribution $p = f_{d,D}(\theta) \in \mathcal{M}_d(D)$ can often be represented compactly by the parameter θ.

Gaussian Bayesian Network Models

The dimension of a Gaussian Bayesian network model is defined as follows.

Definition 52. The *dimension* $d(\mathcal{M}_g(D))$ of a Gaussian Bayesian network model $\mathcal{M}_g(D)$ is

$$d(\mathcal{M}_g(D)) = \sum_{v \in V} (|pa(v)| + 2). \tag{1.38}$$

The dimension $d(\mathcal{M}_g(D))$ is the dimension of the parameter space $\Theta_{g,D}$. Also, the following proposition holds (see [GHKM01]).

Proposition 1.4. *A Gaussian BN model $\mathcal{M}_g(D)$ is a curved exponential model of dimension $d(\mathcal{M}_g(D))$.*

If X is fixed, the dimension varies with the structure as follows. The minimal dimension is $2|X|$ for the structure without any arrow. From there, the dimension increases linearly with the number of parents of each variable until it reaches the maximal dimension $|X| + \frac{1}{2}|X|(|X| + 1)$ for a structure where all the vertices are connected by an arrow.

Remark 25. The mean vector and covariance matrix of a Gaussian density p of X are represented by $|X| + \frac{1}{2}|X|(|X| + 1)$ real parameters, which equals the maximal dimension. Hence, a density $p = f_{g,D}(\theta) \in \mathcal{M}_g(D)$ can often be represented compactly by the parameter θ.

1.5 Implicit Bayesian Network Models

A statistical model $\mathcal{M}$ can be specified *implicitely* as the submodel of a model $\mathcal{M}_0$ where some property holds.

Example 8. By definition, the set $\mathcal{M}$ of all strictly positive distributions of a discrete random variable X can be described implicitely as

$$\mathcal{M} = \left\{ p \in \mathcal{M}_0 \middle| p(x) > 0 \quad \text{for all } x \in \mathcal{X} \right\}. \tag{1.39}$$

where $\mathcal{M}_0$ is the set of all probability distributions of X.

This section shows how discrete and Gaussian BN models can be defined implicitely with independence models. Section 1.5.1 introduces independence models and the statistical models they specify implicitely. Section 1.5.2 defines independence models associated to DAGs, demonstrating how they can be used to encode large sets of independence relations in a compact and intuitive way. Section 1.5.3 provides the connection between the independence relations associated to a DAG and the recursive factorization property, leading to the implicit definitions of discrete and Gaussian BN models. Note that other types of graphs can also be used to represent independence relations. For example, sets represented by undirected and chain graphs are presented in [CDLS99] and [Lau96].

1.5.1 Independence Models

Definition 53. A *(conditional) independence model* I for a set X of random variables is a set of marginal and conditional independence relations between subsets of X.

This dissertation focuses on independence models defined by DAGs. However, they may be defined in other ways, e.g. algebraically (see [VS07]) or as follows.

Definition 54. The independence model $I(P)$ (resp. $I(p)$) associated to a distribution P (resp. density p) for X is the set of independence relations that hold between the subsets of X.

A set of independence relations may imply other independence relations by the axioms of probability theory.

Example 9. Consider a distribution P for X. For $A, B, C, D \subseteq X$, we have

$$(A \perp B|C \in I(P)) \Rightarrow (B \perp A|C \in I(P)) \tag{1.40}$$

$$((A \cup D) \perp B|C \in I(P)) \Rightarrow (A \perp B|C \in I(P)). \tag{1.41}$$

Intuitively, an independence model is probabilistic if it can not be augmented by such implied independence relations.

Definition 55. An independence model I is *probabilistic* if there exists a distribution P such that the independence relations holding in P are exactly those in I, that is $I(P) = I$.

Independence models can be used to define statistical models implicitely.

Definition 56. If I is an independence model I on a set X of random variables and $\mathcal{M}_0$ is a statistical model for X, let $\mathcal{M}(I, \mathcal{M}_0)$ be the submodel of $\mathcal{M}_0$ such that all the independence relations in I hold, that is

$$\mathcal{M}(I, \mathcal{M}_0) = \left\{ p \in \mathcal{M}_0 \middle| I \subseteq I(p) \right\}. \tag{1.42}$$

Remark 26. If $I \subseteq I'$, then $\mathcal{M}(I', \mathcal{M}_0) \subseteq \mathcal{M}(I, \mathcal{M}_0)$.

1.5.2 Independence Models Associated to DAGs

In order to define the independence model associated to a DAG, necessary notions and notations are first introduced.

Definition 57. If $G = (V, E)$ is a graph, the *descendants* of a vertex $v \in V$ is the set of vertices that can be reached by a path starting at v.

Definition 58. If $G = (V, E)$ is a graph, a *trail* is a sequence $v_0, \dots, v_n$ of distinct vertices such that $v_i \cdots v_{i+1} \in G$ for all $i = 0, \dots, n-1$.

Definition 59. If $G = (V, E)$ is a graph, a vertex w_i of a trail $\tau = w_0, \dots, w_n$ is a *collider* of τ if $0 < i < n$, $w_{i-1} \longrightarrow w_i \in G$, and $w_{i+1} \longrightarrow w_i \in G$.

Definition 60. A trail τ between two vertices u and v is *blocked* by a set C of vertices if τ contains

- a non-collider $w \in C$ or
- a collider w such that neither w nor any of its descendants belongs to C.

Definition 61 (*d-separation criterion*). If A, B, and C are disjoint subsets of vertices of a DAG D, then C *d-separates* A *and* B if all the trails between vertices of A and B are blocked by C.

Example 10. In the DAG of Figure 1.3, $\{X_1, X_2\}$ and $\{X_5\}$ are d-separated by $\{X_3\}$, while $\{X_1\}$ and $\{X_2, X_4\}$ are not d-separated by $\{X_5\}$.

The independence model $I(D)$ associated to a DAG D is defined as follows.

Definition 62. If D is a DAG over a set X of random variables, the independence model $I(D)$ is the set of all the independence relations $X_A \perp X_B | X_C$ such that X_A and X_B are d-separated by X_C in D, that is

$$I(D) = \left\{ (X_A \perp X_B | X_C) \middle| X_A \text{ and } X_B \text{ are d-separated by } X_C \right\}. \tag{1.43}$$

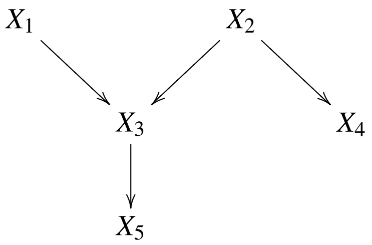


Figure 1.3: d-separation criterion

Definition 63. A DAG D over X is *faithfull* to a density p for X if $I(D) = I(p)$.

Definition 64. An independence model I for X is *DAG isomorph* if there exists a DAG D over X such that $I = I(D)$.

Remark 27. DAG isomorph independence models are probabilistic (see [Mee95] and [SGS01]).

Remark 28. A DAG isomorph independence model I satisfies the *composition property* (see [Pea88]):

$$(A \perp B|C \in I) \wedge (A \perp D|C \in I) \Rightarrow (A \perp (B \cup D)|C \in I) \tag{1.44}$$

for subsets $A, B, C, D \subseteq X$. This property holds because d-separation of sets is defined in terms of d-separation between pairs of vertices in each set.

Not all independence models are DAG isomorph. The following example builds a probabilistic independence model where the composition property does not hold and is thus not DAG isomorph.

Example 11. Consider an experiment where two fair coins are tossed and define three binary random variables C_1, C_2 and S such that $C_1 = h$ (resp. $C_2 = h$) if the first (resp. second) coin falls heads up and $S = t$ if both coins have the same side up. The fairness assumption implies that $P(C_1 = h) = P(C_2 = h) = 0.5$. The independence relations $\{C_1\} \perp \{C_2\}$ and $\{C_1\} \perp \{S\}$ hold in P, but obviously $\{C_1\} \not\perp \{C_2, S\}$.

Empty and Complete DAGs

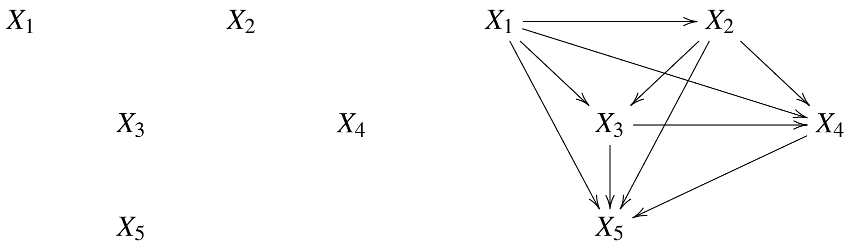
Empty and complete DAGs are especially noteworthy.

Definition 65. A graph is *complete* if there exists an edge between each pair of distinct vertices.

Definition 66. A graph is *empty* if it has no edge.

Remark 29. There are $|V|!$ distinct complete DAGs with a given vertex set V, but only one empty graph.

Example 12. Figure 1.4 shows the empty DAG and a complete DAG over the vertices $\{X_1, \dots, X_5\}$.



(a) empty DAG (b) complete DAG

Figure 1.4: Special types of DAGs

No d-separation holds in a complete DAG D_c, i.e. $I(D_c) = \phi$, while all the possible d-separations hold in an empty DAG D_e. Hence, we have

$$I(D_c) \subseteq I(D) \subseteq I(D_e), \tag{1.45}$$

provided all the DAGs are defined on the same vertex set. Section 1.6 discusses further the inclusion relations between DAG independence models.

Extensions

Larger classes of independence models can be obtained by modifying slightly Definition 62.

Definition 67. If D is a DAG over $X \cup H$, the independence model $I_H(D)$ is the subset of $I(D)$ that contains the independence relations between the subsets of X only, i.e.

$$I_H(D) = \Big\{(A \perp B|C) \in I(D) \Big| A, B, C \subseteq X\Big\}. \tag{1.46}$$

Definition 68. If D is a DAG over $X \cup S$, the independence model $I_S(D)$ is the subset of $I(D)$ that contains the independence relations conditioned on S, i.e.

$$I_S(D) = \Big\{(A \perp B|C) \Big| (A \perp B|C \cup S) \in I(D)\Big\}. \tag{1.47}$$

The above independence models satisfy the composition property. To show that non-DAG isomorph independence models can be obtained, let us introduce the following lemma.

Lemma 1.5. *Let D be a DAG whose vertex set indexes a set X of random variables. The edge $u \cdots v \in D$ if, and only if, $I(D)$ does not contain any independence relation between X_u and X_v.*

Proof.

1. Suppose that $u \cdots v \in D$. The trail u, v can not be blocked, and thus $I(D)$ does not contain any independence relation between X_u and X_v.

2. Suppose that $u \cdots v \notin D$. Let us show that each trail in the set T of all the trails between u and v is blocked by $C_1 \cup C_2$ where

$$C_1 = \{t \in V | \exists (u, t, \dots, v) \in T_1\}, \tag{1.48}$$
$$C_2 = \{t \in V | \exists (u, \dots, t, v) \in T_2\}, \tag{1.49}$$

and

$$T_1 = \{(u, t, \dots, v) \in T | t \to u \in D\}, \tag{1.50}$$
$$T_2 = \{q(u, \dots, t, v) \in T | t \to v \in D\}. \tag{1.51}$$

This will imply that $\{X_u\} \perp \{X_v\} | X_{C_1 \cup C_2} \in I(D)$. Consider a trail $\tau = (u, t_1, \dots, t_n, v) \in T$.

(a) If $\tau \in T_1$, then $t_1 \in C_1$ is a non-collider and τ is blocked by $C_1 \cup C_2$.

(b) If $\tau \in T_2$, then $t_n \in C_2$ is a non-collider and τ is blocked by $C_1 \cup C_2$.

(c) Suppose that $\tau \in T \setminus (T_1 \cup T_2)$. Let t_α be the first collider in τ and let t_β be the last collider in τ. Let us show by contradiction that there are no vertices $d_\alpha, d_\beta \in C_1 \cup C_2$ such that d_α is a descendant of t_α and d_β is a descendant of t_β. Suppose that $t_\alpha, p_1, \dots, p_\alpha, d_\alpha$ and $t_\beta, q_1, \dots, q_\beta, d_\beta$ are paths in D. If $d_\alpha \in C_1$, then $u, t_1, \dots, t_\alpha, p_1, \dots, p_\alpha, d_\alpha, u$ is a cycle. If $d_\beta \in C_2$, then $v, t_n, \dots, t_\beta, q_1, \dots, q_\beta, d_\beta, v$ is a cycle. If $d_\alpha \in C_2$ and $d_\beta \in C_1$, then

$$u, t_1, \dots, t_\alpha, p_1, \dots, p_\alpha, d_\alpha, v, t_n, \dots, t_\beta, q_1, \dots, q_\beta, d_\beta, u \tag{1.52}$$

is a cycle. By acyclicity of D, there are no such vertices d_α and d_β. Hence, τ is blocked by $C_1 \cup C_2$. □

Example 13. Consider the independence model $I_H(D)$ encoded by the DAG of Figure 1.5, and suppose there exists a DAG G over X such that $I(G) = I_H(D)$. By Lemma 1.5, we have $X_1 \cdots X_2 \in G$, $X_2 \cdots X_3 \in G$, $X_3 \cdots X_4 \in G$, $X_1 \cdots X_3 \notin G$ and $X_2 \cdots X_4 \notin G$. Such a DAG encodes $\{X_1\} \perp \{X_3\} | \{X_2\}$ or $\{X_2\} \perp \{X_4\} | \{X_3\}$, but neither relation is in $I_H(D)$.

Example 14. Consider the independence model $I_S(D)$ encoded by the DAG of Figure 1.6, and suppose that there exists a DAG G over X such that $I(G) = I_S(D)$. By Lemma 1.5, G has the edges $X_1 \cdots X_2$, $X_2 \cdots X_3$, $X_3 \cdots X_4$ and $X_1 \cdots X_4$. To avoid having a cycle, G must have at least two arrows pointing towards the same variable, say X_1 (the other cases are similar). In that case, $\{X_2\} \perp \{X_4\} | \{X_1, X_3\}$ is in $I_S(D)$ but not in $I(G)$.

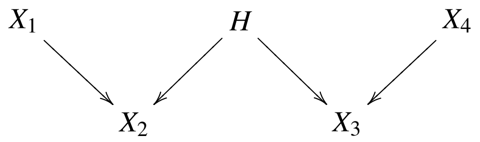


Figure 1.5: A DAG D over $\{H\} \cup \{X_1, X_2, X_3, X_4\}$ encoding a non-DAG isomorph independence model $I_H(D)$ for $\{X_1, X_2, X_3, X_4\}$

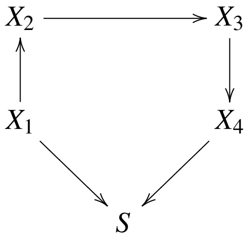


Figure 1.6: A DAG D over $\{S\} \cup \{X_1, X_2, X_3, X_4\}$ encoding a non-DAG isomorph independence model $I_S(D)$ for $\{X_1, X_2, X_3, X_4\}$

1.5.3 Implicit Definition

The following theorem (see e.g. [CDLS99]) provides the link between the recursive factorization property along a DAG D and the independence model $I(D)$. It states that a density p belongs to some Bayesian network model with structure D if, and only if, $I(D) \subseteq I(p)$.

Theorem 1.6. *Let X be a set of discrete variables or a set of continuous variables, and let D be a DAG over X. A distribution P for X admits a recursive factorization according to D if, and only if, the independence relations encoded by D hold in P, that is $I(D) \subseteq I(P)$.*

As a consequence, discrete and Gaussian Bayesian network models admit implicit definitions (see [CDLS99] and [SK89]).

Corollary 1.7. *If X is a set of discrete variables, $\mathcal{M}_0$ is the set of all strictly positive distributions for X and D is a DAG over X, then*

$$\mathcal{M}_d(D) = \mathcal{M}(I(D), \mathcal{M}_0). \tag{1.53}$$

Remark 30. The discrete BN model associated to a complete structure over X is the set of all strictly positive distributions for X. Hence, any strictly positive distribution belongs to some discrete BN model.

Corollary 1.8. *If X is a set of continuous real variables, $\mathcal{M}_0$ is the set of Gaussian densities for X and D is a DAG over X, then*

$$\mathcal{M}_g(D) = \mathcal{M}(I(D), \mathcal{M}_0). \tag{1.54}$$

Remark 31. The Gaussian BN model associated to a complete structure over X is the set of Gaussian densities for X. Hence, any Gaussian density belongs to some Gaussian BN model.

1.6 Equivalence and Inclusion of DAGs

This section studies the partial order on DAGs induced by the inclusion relations between the associated independence models. The results presented constitute the basis of a structure learning algorithm (see Section 2.5.3 and Chapter 3).

Definition 69. If X is a set of vertices or random variables, the set of DAGs over X is denoted $\mathcal{B}(X)$.

Inclusion and equivalence between DAGs is defined as follows.

Definition 70. If $G, H \in \mathcal{B}(X)$, we say that

- G is *independence included* in H, denoted $G \leq_I H$, if $I(H) \subseteq I(G)$
- G is *strictly independence included* in H, denoted $G <_I H$, if $I(H) \subsetneq I(G)$
- G and H are *independence* (or *Markov*) *equivalent*, denoted $G =_I H$, if $I(H) = I(G)$.

If $\mathcal{S}$ is a set of DAGs over X, for example $\mathcal{S} = \mathcal{B}(X)$, independence inclusion and independence equivalence induce a partial order on $\mathcal{S}$ and define a partition of $\mathcal{S}$ into equivalence classes.

As the following proposition shows, the order relation induced by $\leq_I$ carries over to sets of discrete or Gaussian Bayesian network models.

Proposition 1.9. *Given $G, H \in \mathcal{B}(X)$, the following propositions are equivalent:*

(a) $G \leq_I H$

(b) $\mathcal{M}_d(G) \subseteq \mathcal{M}_d(H)$

(c) $\mathcal{M}_g(G) \subseteq \mathcal{M}_g(H)$.

Proof.

1. Let us show that (a) and (b) are equivalent. By (1.53), (a) implies (b). Suppose that (b) holds. There exists $p \in \mathcal{M}_d(G)$ such that p is faithfull to G (see [Mee95]). By (1.53), $p \in \mathcal{M}_d(H)$ implies $I(H) \subseteq I(p) = I(G)$ and (a) thus holds.

2. Let us show that (a) and (c) are equivalent. By (1.54), (a) implies (c). Suppose that (c) holds. There exists $p \in \mathcal{M}_g(G)$ such that p is faithfull to G (see [SGS01]). By (1.54), $p \in \mathcal{M}_d(H)$ implies $I(H) \subseteq I(p) = I(G)$ and (a) thus holds. □

Section 1.6.1, describes independence equivalence and inclusion graphically. Section 1.6.2 introduces a graphical representation of equivalence classes.

1.6.1 Graphical Characterization of $\leq_I$ and $=_I$

First, let us define additional graphical notions.

Definition 71. If $G = (V, E)$ is a graph, a *v-structure* is a pair $(h, \{t_1, t_2\})$ such that $h, t_1, t_2 \in V$ are distinct vertices, $t_1 \rightarrow h \in G$, $t_2 \rightarrow h \in G$, and $t_1 \cdots t_2 \notin G$.

Definition 72. If G is a graph, let $v(G)$ be the set of v-structures of G.

Definition 73. An *undirected graph* is a graph without arrow.

Definition 74. The *skeleton* $S(G)$ of a graph G is the undirected graph obtained from G by converting every arrow into a line.

Definition 75. If G is a graph, an arrow $u \rightarrow v \in G$ is *covered* if $pa_G(v) = pa_G(u) \cup \{u\}$. Otherwise, it is *protected*.

Definition 76. The addition of an arrow to a DAG is *legal* if the resulting graph is still a DAG, that is no cycle is created.

One can easily check whether two DAGs over X are independence equivalent with the following theorem (see [Pea88]).

Theorem 1.10. *Two DAGs are independence equivalent if, and only if, they have the same skeleton and v-structures.*

Example 15. The DAGs of Figure 1.7 are independence equivalent because they have the same skeleton $X_1 - X_3 - X_2$ and no v-structure.

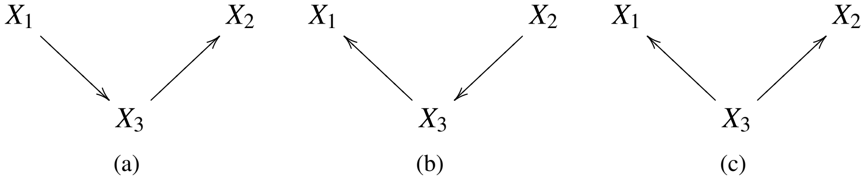


Figure 1.7: Independence equivalent DAGs

Theorems 1.11 and 1.12 (from [Chi95]) show that there exists a transformational characterization of independence equivalence.

Theorem 1.11. *Let D be a DAG and D' be the graph obtained by reversing the arrow $u \to v \in D$. The graph D' is a DAG that is independence equivalent to D if, and only if, $u \to v$ is covered in D.*

Hence, a sequence of covered arrow reversals in D produces an independence equivalent DAG H. The converse assertion also holds.

Theorem 1.12. *Let D and H be a pair of DAGs such that $D =_I H$ and for which there are r arrows in D that have opposite orientation in H. There exists a sequence of r distinct covered arrow reversals in D such that, after all the reversals, $D = H$.*

There exists a similar transformational characterization of independence inclusion which generalizes Theorem 1.12. First, observe that if D' is the result of a legal arrow addition to D, then $D <_I D'$. Hence, a sequence of covered arrow reversals and legal arrow additions in a DAG D results in a DAG H such that $D \leq_I H$. The converse assertion also holds (see [Chi02b] for a constructive proof).

Theorem 1.13. *Let D and H be a pair of DAGs such that $D \leq_I H$, let r be the number of arrows in H that have opposite orientation in D, and let m be the number of arrows in H that do not exist in either orientation in D. There exists a sequence of at most $r + 2m$ covered arrow reversals and legal arrow additions in D such that, after all the reversals and additions, $D = H$.*

Example 16. Let $X = \{X_1, X_2, X_3\}$. There are 25 distincts DAGs in $\mathcal{B}(X)$, but only 11 equivalence classes. The partial order on $\mathcal{B}(X)$ induced by $\leq_I$ is illustrated in Figure 1.8 (from [CK03]).

Theorems 1.11 and 1.13 have an immediate corollary used in Section 1.8.

Corollary 1.14. *Let G and H be DAGs over a set X of discrete (resp. continuous) variables. If $H <_I G$, then $d(\mathcal{M}_d(H)) < d(\mathcal{M}_d(G))$ (resp. $d(\mathcal{M}_g(H)) < d(\mathcal{M}_g(G))$.*

1.6.2 Essential Graphs

Besides their representation as sets of DAGs, equivalence classes in $\mathcal{B}(X)$ induced by $=_I$ may also be represented graphically. By Theorem 1.10, the elements of a Markov equivalence class of structures $C \subseteq \mathcal{B}(X)$ have the same skeleton and only differ in the orientation of their arrows.

Definition 77. An arrow $u \to v$ is *compelled* in a Markov equivalence class of structures C if it exists in every DAG of C. An arrow that is not compelled is *reversible*.

Remark 32. By Theorem 1.10, arrows participating in a v-structure are compelled.

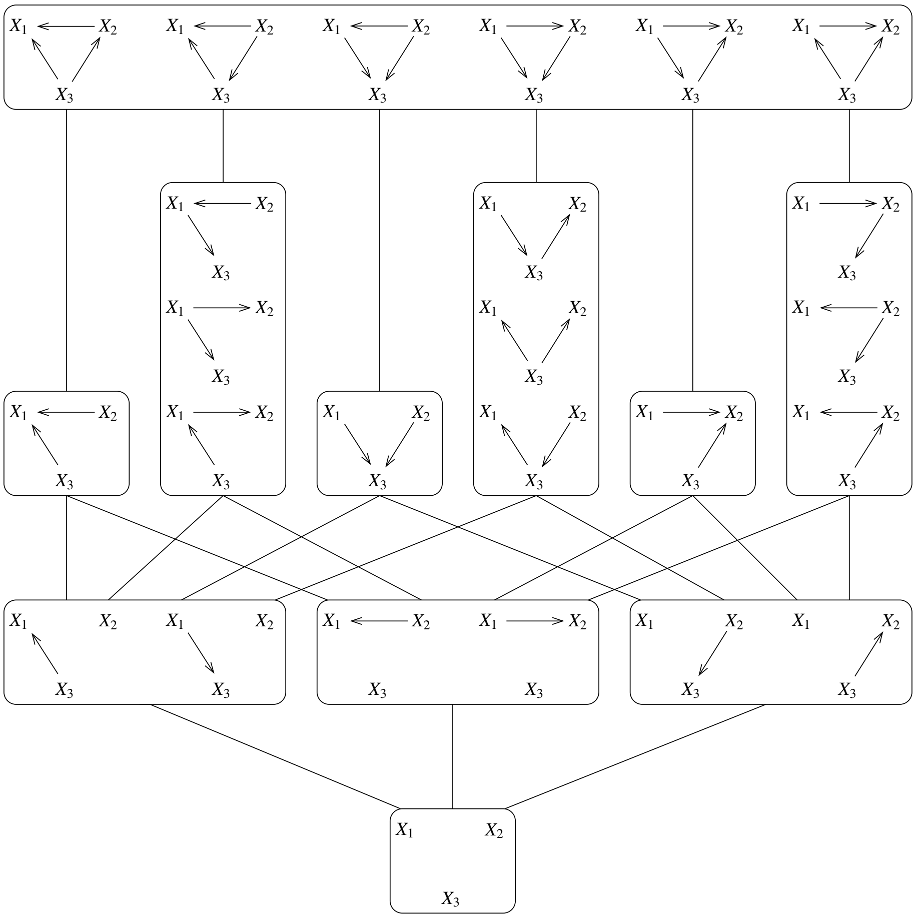


Figure 1.8: Inclusion order on $\mathcal{B}(X)$. For any $D, H \in \mathcal{B}(X)$, $D =_I H$ if, and only if, D and H are contained in the same box, while $D <_I H$ if, and only if, there is a downward path from the box containing H to the box containing D.

Example 17. Consider the equivalence class associated to the DAG of Figure 1.9. The arrows $X_1 \to X_3$ and $X_2 \to X_3$ are compelled because they are part of the v-structure $(X_3, \{X_1, X_2\})$. The arrow $X_3 \to X_4$ is also compelled because its reversal would create the v-structures $(X_3, \{X_1, X_4\})$ and $(X_3, \{X_2, X_4\})$.

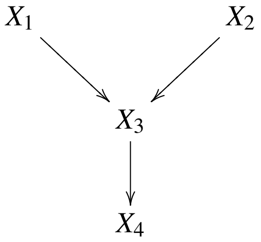


Figure 1.9: A DAG where all the arrows are compelled

The following graph can be associated to a Markov equivalence class.

Definition 78. The *essential graph* $E(C)$ (or *completed partially directed acyclic graph* or *DAG pattern*) of a Markov equivalence class $C \subseteq \mathcal{B}(V)$ is the graph over V with an arrow for every compelled arrow in C and a line for every reversible arrow in C, i.e.

$$E(C) = \Big(V, \bigcup_{(V,E)\in C} E\Big). \tag{1.55}$$

Remark 33. If $C \subseteq \mathcal{B}(V)$ is a Markov equivalence class of structures and $D \in C$, then $v(E(C)) = v(D)$ and $S(E(C)) = S(D)$ by Theorem 1.10.

As the following proposition shows, essential graphs can be used to represent Markov equivalence classes of structures.

Proposition 1.15. *If* $E(C_1) = E(C_2)$*, then* $C_1 = C_2$.

Proof. Let us show that $C_1 \neq C_2$ implies $E(C_1) \neq E(C_2)$. Let $D_1 \in C_1$ and $D_2 \in C_2$. By Theorem 1.10, $C_1 \neq C_2$ implies that $S(D_1) = S(E(C_1)) \neq S(D_2) = S(E(D_2))$ or $v(D_1) = v(E(C_1)) \neq v(D_2) = v(E(C_2))$. Hence, $E(C_1) \neq E(C_2)$. □

Definition 79. The set of essential graphs over X is denoted $\mathcal{E}(X)$.

Definition 80. If $E \in \mathcal{E}(X)$, let $[E] \subseteq \mathcal{B}(X)$ be the Markov equivalence class of structures such that $E([E]) = E$.

Remark 34. If $D \in \mathcal{B}(X)$ and $E \in \mathcal{E}(X)$, then $D \in [E]$ if, and only if, D and E have the same skeleton and v-structures.

Remark 35. Chapter 3 (see also [AMP97] and [Chi02b]) presents efficient algorithms to build the essential graph of an equivalence class and to find a member of an equivalence class represented by an essential graph.

Many notions defined for DAGs only depend on the independence model associated to them. They can thus be extended to essential graphs. If E is an essential graph and $G \in [E]$, the independence model $I(E) = I(G)$ and the statistical models $\mathcal{M}_d(E) = \mathcal{M}_d(G)$ and $\mathcal{M}_g(E) = \mathcal{M}_g(G)$ are defined. Similarly, independence inclusion and independence equivalence between essential graphs are also defined.

Example 18. The partial order on $\mathcal{E}(\{X_1, X_2, X_3\})$ induced by $\leq_I$ is illustrated in Figure 1.10.

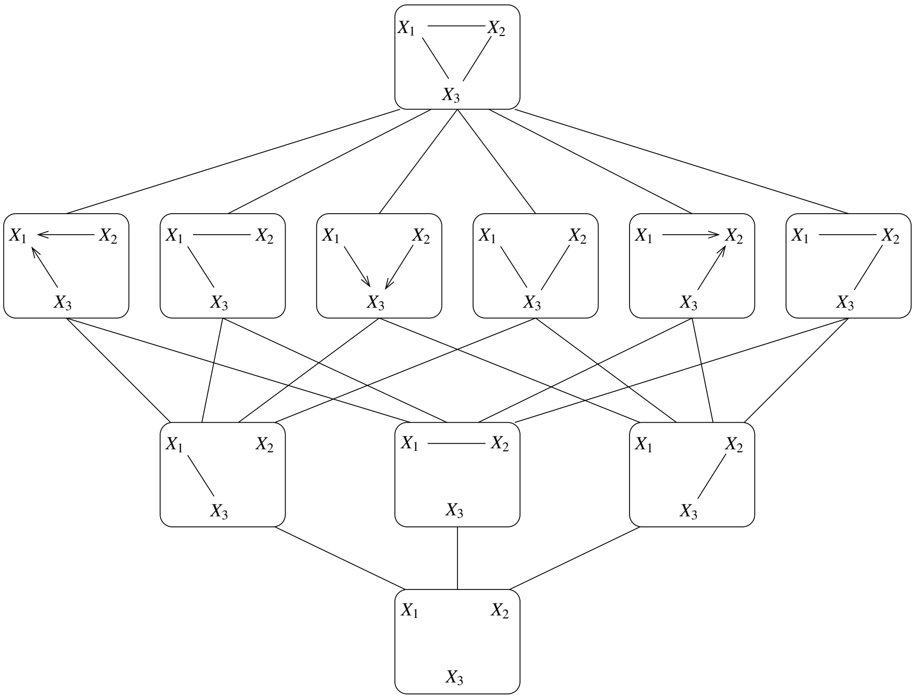


Figure 1.10: Inclusion order on $\mathcal{E}(\{X_1, X_2, X_3\})$

1.7 Bayesian Network Models with Hidden Variables

Bayesian network models with hidden variables are statistical models derived from Bayesian network models. Their study is relevant to structure and parameter learning with hidden variables (see Chapter 2 and Chapter 4).

1.7.1 Hidden and Observable Variables

Let us introduce the notion of hidden and observable variables by an example.

EXAMPLE 19. Consider a Bayesian network B for $\{X_1, X_2, X_3, X_4, X_5, X_6, H_1\}$ with structure given in Figure 1.11 and representing a density p_B. Suppose that B is used for medical diagnosis, and

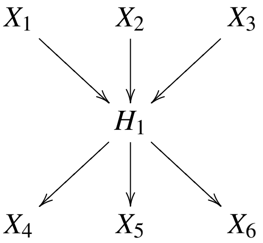


Figure 1.11: A structure with one hidden variable

- X_1, X_2 and X_3 represent the medical history of a patient
- H_1 is a variable of interest about the patient, e.g. it indicates a disease
- X_4, X_5 and X_6 are symptoms or tests results.

In practice, the variable H_1 may never be observed directly because it is too difficult, dangerous or costly to do so (e.g. H_1 measures the weight of a fetus, or it indicates the presence or absence of a brain tumor) or because it is intrinsically non-observable (e.g. H_1 does not correspond to a physical event, but rather a mental construction like a syndrome). In that case, B may be treated as a representation of the marginal density for $\{X_1, X_2, X_3, X_4, X_5, X_6\}$ of p_B.

A variable whose value is never observed is *hidden* (or *latent*), while a variable that is sometimes or always observed is *observable* (or *manifest*). These notions depend on the context: a variable that is hidden in one situation may be observable in another and vice-versa. In Section 2.3.2, the notion of context is defined by a dataset.

1.7.2 Bayesian Network Models with Hidden Variables

As Example 19 showed, hidden variables can be marginalized. This leads to the following statistical models.

Definition 81. Let X and H be non-empty sets partitioning a finite set Y of random variables, and let $\mathcal{M}$ be a Bayesian network model for Y. The *Bayesian network model* $\mathcal{M}_H$ *with hidden variables* H associated to $\mathcal{M}$ is the statistical model obtained by marginalization of the hidden variables.

Remark 36. For discrete variables, we have

$$\mathcal{M}_H = \Big\{p(x) = \sum_{h\in\mathcal{H}} p(x,h) \,\Big|\, p(x,h) \in \mathcal{M}\Big\}. \tag{1.56}$$

For continuous variables, we have

$$\mathcal{M}_H = \Big\{p(x) = \int_{\mathcal{H}} p(x,h)dh \,\Big|\, p(x,h) \in \mathcal{M}\Big\}. \tag{1.57}$$

Remark 37. A parametrization map f_H for $\mathcal{M}_H$ can be obtained by composing a parametrization map f for $\mathcal{M}$ and the marginalization operation. In general, the injectivity of f does not imply the injectivity of f_H.

Let us give two examples of Bayesian network models with hidden variables associated to discrete Bayesian network models. They are generalized in Section 2.3.2.

Example 20. Let D be the DAG over the set of discrete variables $\{X_1, X_2, X_3\} \cup \{H\}$ given in Figure 1.12. The Bayesian network model $\mathcal{M}_H$ with hidden variable H

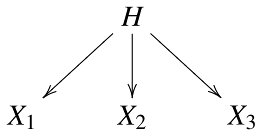


Figure 1.12: A naive Bayes structure D

associated to $\mathcal{M}_d(D)$ admits the parameter space

$$\Theta_H = \Theta_{d,D} = (S^+_{X_1})^{|\mathcal{H}|} \times (S^+_{X_2})^{|\mathcal{H}|} \times (S^+_{X_3})^{|\mathcal{H}|} \times S^+_H \tag{1.58}$$

and the parametrization map f_H given by

$$f_H\Big(\big(((\theta^{X_i,h}_{x_i})_{x_i\in\mathcal{X}_i})_{h\in\mathcal{H}}\big)_{i=1}^{3}, (\theta^H_h)_{h\in\mathcal{H}}\Big) = p \tag{1.59}$$

with

$$p(x_1,x_2,x_3) = \sum_{h\in\mathcal{H}} \theta^H_h \prod_{i=1}^{3} \theta^{X_i,h}_{x_i}, \quad (x_1,x_2,x_3) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3. \tag{1.60}$$

Example 21. Let D be the DAG over the set of discrete variables $\{X_1, \ldots, X_7\} \cup \{H_1, H_2, H_3\}$ given in Figure 1.13. The Bayesian network model $\mathcal{M}_H$ with hidden variables $\{H_1, H_2, H_3\}$ associated to $\mathcal{M}_d(D)$ admits the parameter space

$$\Theta_H = \Theta_{d,D} = \Big((S^+_{X_i})^{|\mathcal{H}_2|}\Big)_{i=1}^{3} \times (S^+_{X_4})^{|\mathcal{H}_1|} \times \Big((S^+_{X_i})^{|\mathcal{H}_3|}\Big)_{i=5}^{7} \times S^+_{H_1} \times (S^+_{H_2})^{|\mathcal{H}_1|} \times (S^+_{H_3})^{|\mathcal{H}_1|} \tag{1.61}$$

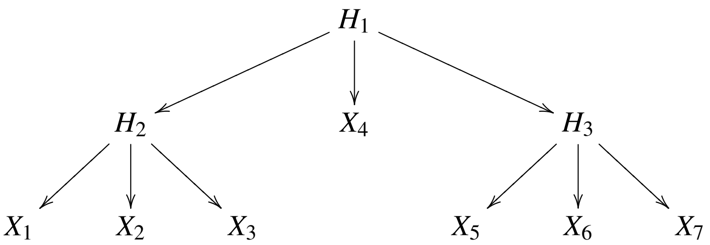


Figure 1.13: A HLC structure D

and the parametrization map f_H such that $f_H(\theta) = p$ with

$$p(x_1, \dots, x_7) = \sum_{h_1 \in \mathcal{H}_1} \sum_{h_2 \in \mathcal{H}_2} \sum_{h_3 \in \mathcal{H}_3} \theta_{h_1}^{H_1} \theta_{h_2}^{H_2,h_1} \theta_{h_3}^{H_3,h_1} (\prod_{i=1}^{3} \theta_{x_i}^{X_i,h_2}) \theta_{x_4}^{X_4,h_1} (\prod_{i=5}^{7} \theta_{x_i}^{X_i,h_3}). \tag{1.62}$$

In general, the densities of Bayesian network model with hidden variables $\mathcal{M}_H$ no longer satisfy a recursive factorization property, and it is not easy to find an implicit description of $\mathcal{M}_H$. However, $p \in \mathcal{M}_H$ implies $I_H(D) \subseteq I(p)$, and thus $\mathcal{M}_H \subseteq \mathcal{M}_0$ implies $\mathcal{M}_H \subseteq \mathcal{M}(I_H(D), \mathcal{M}_0)$. As a matter of fact, the marginalization sometimes introduces constraints that can not be expressed as independence constraints (see [GM98], [Gar04] and Corollary 4.10 in Chapter 4).

Example 22. Let $X_1, \dots, X_6$, and H_1 be binary variables, let D_1 be the DAG of Figure 1.11, and let D_2 be the DAG of Figure 1.14. One can see that $I_H(D_1) = I(D_2)$ but $d(\mathcal{M}_d(D_1)) = 17 < 59 = d(\mathcal{M}_d(D_2))$. Hence, the marginalization of H_1 introduces constraints not accounted for in $I_H(D_1)$.

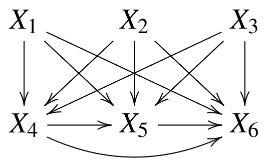


Figure 1.14: A structure D_2 without hidden variable such that $I(D_2) = I_H(D_1)$

1.7.3 Dimension

Defining a notion of dimension for discrete or Gaussian Bayesian network models with hidden variables is not as straightforward as in the case without hidden variables. In fact, there exist different notions of dimension. Let us first present the *standard dimension*.

Definition 82 (*Standard dimension*). Let $\mathcal{M}$ be a discrete or Gaussian BN model for the variables $X \cup H$ and let $\mathcal{M}_H$ be the corresponding BN model with hidden variables. The *standard dimension* $ds(\mathcal{M}_H)$ of $\mathcal{M}_H$ is the dimension of $\mathcal{M}$, i.e. the dimension of the parameter space (see (1.37) and (1.38)).

The standard dimension is not always equal to the number of independent parameters necessary to represent a density (see [GHM96] and [GHKM01]).

Example 23. Consider the structure of Figure 1.5. Suppose that H is hidden and that the other variables are observable and binary. As noted before, any distribution on these four observable variables can be parametrized by a vector of $2^4 - 1 = 15$ components (see (1.29)). If H is binary, then $ds(\mathcal{M}_H) = 11 \leq 15$. However, if H is ternary, then $ds(\mathcal{M}_H) = 28 > 15$.

In [GHKM01], the authors show that discrete and Gaussian Bayesian network models with hidden variables are stratified exponential models (see Appendix A). They can thus inherit their definition of dimension.

Definition 83 (*Effective dimension*). Let $\mathcal{M}$ be a discrete or Gaussian BN model for the variables $X \cup H$ and let $\mathcal{M}_H$ be the corresponding BN model with hidden variables. The *(effective) dimension* $d(\mathcal{M}_H)$ of $\mathcal{M}_H$ is its dimension as a stratified exponential model.

Remark 38. Computing $d(\mathcal{M}_H)$ is a non-trivial problem (see [RG03], [KZ02], [KZ03] and Appendix B).

Example 24. Consider the structure over binary variables described in Example 23. One can show that the effective dimension is smaller than the standard dimension, with $d(\mathcal{M}_H) = 9 < 11 = ds(\mathcal{M}_H)$.

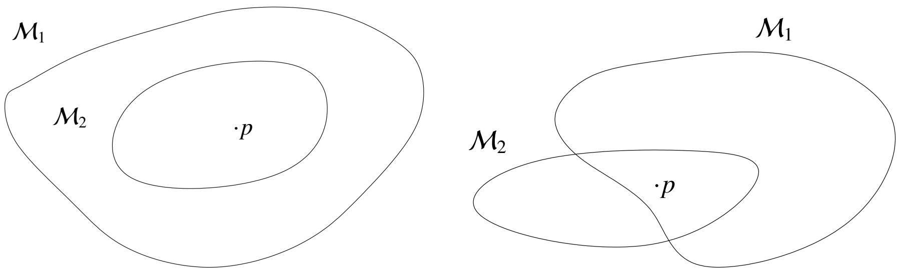
1.8 Optimality of a Bayesian Network Model

This section introduces two notions of optimality that are important to decide whether a Bayesian network model is a good candidate to represent a density. These notions are used in Section 2.5 to evaluate the quality of models learned from data.

Definition 84 (*Inclusion optimality*). Let p be a density for a set X of variables and let M be a set of statistical models for X. A model $\mathcal{M} \in M$ is *inclusion optimal w.r.t. M and p* if $p \in \mathcal{M}$ and there is no $\mathcal{M}' \in M$ such that $p \in \mathcal{M}'$ and $\mathcal{M}' \subsetneq \mathcal{M}$.

Recall that the dimension of a discrete or Gaussian BN model measures its complexity. This leads to the following definition, illustrated in Figure 1.15.

Definition 85 (*Parameter optimality*). Let p be a density for a set X of variables and let M be a set of discrete Bayesian network models or a set of Gaussian Bayesian network models for X. A model $\mathcal{M} \in M$ is *parameter optimal w.r.t. M and p* if $p \in \mathcal{M}$ and there is no $\mathcal{M}' \in M$ such that $p \in \mathcal{M}'$ and $d(\mathcal{M}') < d(\mathcal{M})$.



(a) $\mathcal{M}_1$ is neither inclusion nor parameter optimal, while $\mathcal{M}_2$ is parameter and inclusion optimal

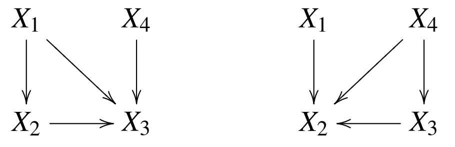
(b) $\mathcal{M}_1$ is inclusion but not parameter optimal, while $\mathcal{M}_2$ is parameter and inclusion optimal

Figure 1.15: M is composed of two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The dimension of a model is proportional to its area in the figure.

Remark 39. For discrete and Gaussian BN models without hidden variable, parameter optimality implies inclusion optimality. Indeed, $\mathcal{M}_1 \subsetneq \mathcal{M}_2$ implies that $d(\mathcal{M}_1) < d(\mathcal{M}_2)$ by Proposition 1.9 and Corollary 1.14.

Let us present two examples illustrating these notions of optimality.

Example 25 (from [CM02]). Let D_1 be the DAG given in Figure 1.5 with $|\mathcal{X}_1| = |\mathcal{X}_3| = |\mathcal{X}_4| = 2$ and $|\mathcal{X}_2| = 3$, let p be a distribution such that $I(p) = I_H(D_1)$, and let M be the set of all discrete BN models over $\{X_1, X_2, X_3, X_4\}$. The discrete BN model with structure given in Figure 1.16(a) is parameter and inclusion optimal, while the discrete BN model with structure given in Figure 1.16(b) is inclusion optimal but not parameter optimal.

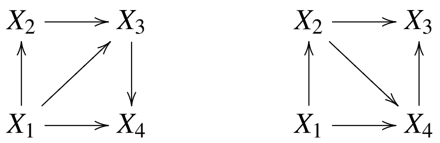


(a) parameter optimal (d = 18)

(b) inclusion optimal (d = 20)

Figure 1.16: Structure of optimal models when $I(p) = I_H(D_1)$

Example 26 (from [CM02]). Let D_2 be the DAG given in Figure 1.6 with $|\mathcal{X}_2| = |\mathcal{X}_3| = |\mathcal{X}_4| = 2$ and $|\mathcal{X}_1| = 4$, let p be a distribution such that $I(p) = I_S(D_2)$, and let M be the set of all discrete BN models over $\{X_1, X_2, X_3, X_4\}$. The discrete BN model with structure given in Figure 1.17(a) is parameter and inclusion optimal, while the discrete BN model with structure given in Figure 1.17(b) is inclusion optimal but not parameter optimal.



(a) parameter optimal ($d = 19$)

(b) inclusion optimal ($d = 23$)

Figure 1.17: Structure of optimal models when $I(p) = I_S(D_2)$

There exists a connection between the notions of faithfulness, parameter optimality, and inclusion optimality.

Proposition 1.16. *Let X be a set of discrete (resp. continuous) variables, let M be the set of discrete (resp. Gaussian) BN models with structure in a set $\mathcal{S} \subseteq \mathcal{B}(X)$, and let p be a distribution for X such that there exists a DAG $D \in \mathcal{S}$ faithfull to p and satisfying $p \in \mathcal{M}_d(D)$ (resp. $p \in \mathcal{M}_g(D)$). If $G \in \mathcal{S}$, the following propositions are equivalent*

(a) G is faithfull to p

(b) $\mathcal{M}_d(G)$ (resp. $\mathcal{M}_g(G)$) is parameter optimal w.r.t. M and p

(c) $\mathcal{M}_d(G)$ (resp. $\mathcal{M}_g(G)$) is inclusion optimal w.r.t. M and p.

Proof. Let $\mathcal{M}(G)$ denote the discrete (resp. Gaussian) BN model associated to a DAG G over the set X of discrete (resp. continuous) variables.

1. Let us show that (a) implies (b). By Corollary 1.14, $\mathcal{M}(G)$ has minimal dimension over the set of DAGs H such that $I(H) \subseteq I(p) = I(G)$. Hence, $\mathcal{M}(G)$ is parameter optimal.

2. By Remark 39, (b) implies (c).

3. Let us show that (c) implies (a). Since $p \in \mathcal{M}(G)$, we have $I(G) \subseteq I(p)$ by the implicit definition of $\mathcal{M}(G)$. Since $p \in \mathcal{M}(D)$ and G is optimal, $\mathcal{M}(D)$ is not a proper subset of $\mathcal{M}(G)$. By Proposition 1.9, $I(G)$ is thus not a proper subset of $I(D)$. By faithfulness of D, $I(G)$ is not a proper subset of $I(p)$. Hence, $I(G) = I(p)$. □

Finding a parameter optimal model may be difficult. First, M may contain too many elements to be enumerated in reasonable time. For example, the number of DAGs over n vertices grows very quickly with n (see [Rob77]).

Proposition 1.17. *The number $f(n)$ of DAGs over n vertices is given by the following recurrence:*

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \quad \text{for } n > 0, \tag{1.63}$$

$$f(0) = 1. \tag{1.64}$$

Example 27. We have $f(2) = 3$, $f(3) = 25$, $f(5) = 29000$ and $f(10) \approx 4.2 \times 10^{18}$.

For large $|X| = n$, one can show that $|\mathcal{B}(X)| = f(n) = 2^{O(n^2 \log n)}$ (see [FK03]). The following subsets of $\mathcal{B}(X)$ can also be considered.

Definition 86. If $k \geq 0$, let $\mathcal{B}_k(X)$ be the set of DAGs over X such that each vertex has at most k parents.

Even in the case $M = \mathcal{B}_k(X)$, enumeration may still not be possible: for large $n = |X|$, one can show that $|\mathcal{B}_k(X)| = 2^{O(kn \log n)}$ (see [FK03]). Finally, the following complexity result holds (see [CMH03]).

Theorem 1.18. *Given a finite set X of discrete variables, a distribution p for X, and integers $k \geq 3$ and d, the problem of deciding if there exists a discrete Bayesian network model with structure in $\mathcal{B}_k(X)$ that contains p and has a dimension less than or equal to d is NP-hard.*

Hence, identifying a parameter optimal discrete Bayesian network model with structure in $\mathcal{B}_k(X)$ is NP-hard for $k \geq 3$.

Chapter 2

Learning Bayesian Networks

2.1 Introduction

There are many different ways to precisely define and approach the problem of learning Bayesian networks from data (see e.g. [Nea03] or [NWL$^+$04]). This chapter presents a Bayesian approach to learning Bayesian networks over discrete variables and discrete Bayesian network models. It does not attempt to cover all the existing approaches and variants but rather introduces material relevant for subsequent chapters of the dissertation.

Learning Bayesian networks can be seen as an instance of learning statistical models. The problem of learning statistical models can be described informally as follows. First, let us define the notions of observation and dataset.

Definition 87. An *observation* o is a value $o \in \mathcal{O}$ of a set of random variables O.

Definition 88. A *dataset* (or *sample*) d of size n is a n-tuple $(o[1], \ldots, o[n])$ of observations.

A distribution g for a set X of random variables can generate an observation o of a subset of variables $O \subseteq X$. Given a dataset d of independent observations generated by an unknown distribution g, the following problems arise:

- if M is a set of statistical models, find a low-dimensional model $\mathcal{M} \in M$ containing a good approximation of g (in some sense);

- if $\mathcal{M}$ is a statistical model, find a good approximation $q \in \mathcal{M}$ of g (in some sense).

Remark 40. Because g is unknown, the quality of the approximation p has to be estimated from d only.

In the context of Bayesian networks, these learning problems can be formulated as follows. Given a dataset d generated by g and a set $\mathcal{S}$ of DAGs (or essential graphs), the *structure learning* problem is to find a structure $D \in \mathcal{S}$ such that

$\mathcal{M}_d(D)$ has small dimension and contains a good approximation of g. Given a dataset d generated by g and a structure D, the *parameter learning* problem is to find a parameter $\theta \in \Theta_{d,G}$ such that $f_{d,G}(\theta)$ is a good approximation of g.

The rest of the chapter is organised as follows. Section 2.2 outlines a Bayesian approach to solve the above learning problems. Section 2.3 discusses the adaptation of the Bayesian framework to discrete Bayesian network models. Section 2.4 presents parameter learning. Section 2.5 presents structure learning.

2.2 A Bayesian Approach to Learning Statistical Models

A Bayesian approach to learning statistical models can be outlined as follows. For the interested reader, Bayesian methods are discussed in [Jay03]. In a few words, a Bayesian approach combines a *prior distribution* on (subsets of) a *hypothesis space* with some *data* to obtain a *posterior distribution*.

Suppose that a dataset d generated by an unknown distribution g is given. Also, suppose that M is a finite or countable set of statistical models such that $g \in \cup_{\mathcal{M} \in M} \mathcal{M}$ and each $\mathcal{M} \in M$ is defined parametrically by a parameter space $\Theta_{\mathcal{M}}$ and an injective parametrization map $f_{\mathcal{M}}$, i.e. $\mathcal{M} = f_{\mathcal{M}}(\Theta_{\mathcal{M}})$. A *hypothesis space* is a set of hypotheses and, loosely speaking, a *hypothesis* is a proposition that explains the data d. A hypothesis "$g \in \mathcal{M}$" is associated to each $\mathcal{M} \in M$, and a hypothesis "$g = f_{\mathcal{M}}(\theta)$" is associated to each $\theta \in \Theta_{\mathcal{M}}$.

Remark 41. The injectivity of $f_{\mathcal{M}}$ ensures that distinct parameters define distinct hypotheses.

Remark 42. To simplify notations, a model $\mathcal{M}$ and the associated hypothesis "$g \in \mathcal{M}$" are both simply denoted $\mathcal{M}$. Similarly, a parameter $\theta \in \Theta_{\mathcal{M}}$ and the associated hypothesis are both denoted θ.

A hypothesis "explains the data" in the following sense: if $g = f_{\mathcal{M}}(\theta)$, the probability $p(d|\theta, \mathcal{M})$ of observing d is given by

$$p(d|\theta, \mathcal{M}) = \prod_{i=1}^{n} f_{\mathcal{M}}(\theta)(o[i]). \tag{2.1}$$

The probability distribution $p(d|\theta, \mathcal{M})$ is called the *sampling distribution*.

In a Bayesian approach, the *prior distribution* represents our knowledge about the hypotheses before observing the data. A prior distribution may be defined on some σ-field in the hypothesis space as follows. For each $\mathcal{M} \in M$, let $p_{\mathcal{M}}$ be a probability measure such that $p_{\mathcal{M}}(\mathcal{M}') = 0$ if $\mathcal{M}' \neq \mathcal{M}$ and such that there exists a function $p(\cdot|\mathcal{M})$ defined on $\Theta_{\mathcal{M}}$, called the *prior parameter density* of $\mathcal{M}$, satisfying

$$p_{\mathcal{M}}(A) = \int_{f_{\mathcal{M}}^{-1}(A)} p(\theta|\mathcal{M}) d\theta. \tag{2.2}$$

Additionally, for each $\mathcal{M} \in M$, let $\alpha_{\mathcal{M}}$ be a non-negative real, called the *prior model probability* of $\mathcal{M}$, such that $\sum_{\mathcal{M} \in M} \alpha_{\mathcal{M}} = 1$. Then, one can see that the *prior distribution*

$$p = \sum_{\mathcal{M} \in M} \alpha_{\mathcal{M}} p_{\mathcal{M}} \tag{2.3}$$

is a probability measure such that

$$p(\cdot|\mathcal{M}) = p_{\mathcal{M}}(\cdot), \tag{2.4}$$

$$p(\mathcal{M}) = \alpha_{\mathcal{M}}. \tag{2.5}$$

After observing the data, our knowledge about the hypotheses is represented by the *posterior parameter densities*

$$p(\theta|d, \mathcal{M}) = \frac{p(\theta|\mathcal{M})p(d|\theta, \mathcal{M})}{p(d|\mathcal{M})}, \tag{2.6}$$

and the *posterior model probabilities*

$$p(\mathcal{M}|d) = \frac{p(\mathcal{M})p(d|\mathcal{M})}{p(d)}. \tag{2.7}$$

The probability $p(d|\mathcal{M})$ of observing the data given that $g \in \mathcal{M}$ is called the *marginal likelihood* of $\mathcal{M}$, and it is given by

$$p(d|\mathcal{M}) = \int_{\Theta_{\mathcal{M}}} p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})d\theta. \tag{2.8}$$

In the Bayesian approach, a solution to the problem of selecting a good model in M given the dataset d is simply a model $\mathcal{M} \in M$ with maximal posterior probability:

$$\mathcal{M} \in \arg\max_{\mathcal{M} \in M} p(\mathcal{M}|d). \tag{2.9}$$

Estimating a distribution $q \in \mathcal{M} = f_{\mathcal{M}}(\Theta_{\mathcal{M}})$ that approximates g amounts to estimating a suitable parameter $\theta \in \Theta_{\mathcal{M}}$. In the Bayesian approach, the posterior parameter density $p(\theta|d, \mathcal{M})$ represents our knowledge about the parameters after observing d and supposing that $g \in \mathcal{M}$. A solution θ to our learning problem is thus somehow extracted from $p(\theta|\mathcal{M}, d)$. For example, a natural estimate is a parameter maximizing the posterior parameter density:

$$\theta \in \arg\max_{\theta \in \Theta_{\mathcal{M}}} p(\theta|d, \mathcal{M}). \tag{2.10}$$

Other parameter estimates can be obtained from $p(\theta|d, \mathcal{M})$ (see Section 2.4.1).

2.3 Hypothesis Space

This section adapts the Bayesian framework of Section 2.2 to discrete Bayesian network models. Alternatively, an excellent tutorial is given in [Hec98]. For Gaussian Bayesian networks, see [GH94] or [Nea03].

A discrete Bayesian network model $\mathcal{M}$ for a set X of discrete random variables may be defined by

- a DAG $D \in \mathcal{B}(X)$ such that $\mathcal{M} = \mathcal{M}_d(D)$ or
- an essential graph $E \in \mathcal{E}(X)$ such that $\mathcal{M} = \mathcal{M}_d(E)$.

Hence, a set $\mathcal{S}$ of DAGs or essential graphs defines a set M of discrete Bayesian network models.

Remark 43. If D is a DAG, then $\mathcal{M}_d(D)$ is parametrized as $\mathcal{M}_d(D) = f_{d,D}(\Theta_{d,D})$. If E is an essential graph, then $\mathcal{M}_d(E)$ can be parametrized by any $G \in [E]$ as $\mathcal{M}_d(E) = f_{d,G}(\Theta_{d,G})$. To simplify notations, we suppose that a fixed DAG $D \in [G]$ is associated to each essential graph $G \in \mathcal{S}$ and we let $\Theta_{d,G} = \Theta_{d,D}$ and $f_{d,G} = f_{d,D}$. Then, an essential graph or DAG $G \in \mathcal{S}$ defines a discrete Bayesian network model parametrized as $\mathcal{M}_d(G) = f_{d,G}(\Theta_{d,G})$. Furthermore, essential graphs are referred to as structures.

Following the example of Section 2.2, one can associate to each $G \in \mathcal{S}$ the hypothesis $g \in \mathcal{M}_d(G)$ and one can associate to each $\theta \in \Theta_{d,G}$ the hypothesis $g = f_{d,G}(\theta)$. Then, the prior distribution on the hypothesis space is specified by prior structure probabilities $p(G)$ and prior parameter densities $p(\theta|G)$. However, this straithforward choice of hypothesis space leads to technical difficulties discussed in Section 2.3.1. Possible choices for $\mathcal{S}$ are presented in Section 2.3.2.

2.3.1 Technical Difficulties

Unlike essential graphs, distinct DAGs may define the same Bayesian network model and thus the same hypothesis. To specify the prior distribution coherently, the prior structure probabilities should satisfy

$$p(G) = p(H) \tag{2.11}$$

and the prior parameter densities should satisfy

$$\int_{f_{d,G}^{-1}(A)} p(\theta|G)d\theta = \int_{f_{d,H}^{-1}(A)} p(\theta|H)d\theta \tag{2.12}$$

for all DAGs $G, H \in \mathcal{S}$ such that $\mathcal{M}_d(G) = \mathcal{M}_d(H)$, i.e. $G =_I H$ (see Proposition 1.9).

Remark 44. Constraint (2.12) is satisfied if

$$p(\theta|G) = p(f_{d,H}^{-1}(f_{d,G}(\theta))\big|H)|\det J(\theta)| \tag{2.13}$$

where $J(\theta)$ denotes the Jacobian matrix at θ of the transformation $f_{d,H}^{-1} \circ f_{d,G}$.

Remark 45. Despite these constraints, sets of DAGs are still used to specify sets of statistical models because DAGs are more intuitive and easier to manipulate than essential graphs. Although this is not done in this dissertation, the hypothesis associated to a DAG may be redefined to include a causal interpretation of the arrows (see Remark 17). In that case, distinct but Markov equivalent DAGs define hypotheses asserting distinct causal relationships.

The second difficulty, common to sets of DAGs and sets of essential graphs, relates to the constraint $p_{\mathcal{M}}(\mathcal{M}') = 0$ if $\mathcal{M}' \neq \mathcal{M}$. If $\mathcal{M}_d(H) \nsubseteq \mathcal{M}_d(G)$, then

$$p_H(G) = \int_{f_{d,H}^{-1}(\mathcal{M}_d(G))} p(\theta|H)d\theta = 0 \tag{2.14}$$

for any prior parameter density $p(\theta|H)$ (see [HGC95]). However, $\mathcal{M}_d(H) \subseteq \mathcal{M}_d(G)$ implies $p_H(G) \geq p_H(H) = 1$. To avoid this inclusion problem, we associate to each $G \in \mathcal{S}$ the model

$$\mathcal{M}'_d(G) = \mathcal{M}_d(G) \setminus \bigcup_{H \in \mathcal{S} \text{ s.t. } H <_I G} \mathcal{M}_d(H), \tag{2.15}$$

and the hypothesis $g \in \mathcal{M}'_d(G)$.

Remark 46. Technically, the model $\mathcal{M}'_d(G)$ is parametrized by the parameter space $\Theta'_{d,G} = f_{d,G}^{-1}(\mathcal{M}'_d(G))$ and the parametrization map $f_{d,G}$ restricted to $\Theta'_{d,G}$. In practice, these modifications have little consequence because

$$\int_{\Theta_{d,G}} p(\theta|G)d\theta = \int_{\Theta'_{d,G}} p(\theta|G)d\theta. \tag{2.16}$$

Hence, we can specify a prior parameter density $p(\theta|G)$ over $\Theta_{d,G}$ instead of $\Theta'_{d,G}$, and we can use $\Theta_{d,G}$ instead of $\Theta'_{d,G}$ in the computation of the marginal likelihood $p(d|G)$.

2.3.2 Choice of $\mathcal{S}$

In general, the choice of set $\mathcal{S}$ depends on the dataset and prior knowledge. Structures can be classified according to the type of variables they are defined on.

Definition 89. A random variable X is *observable* in a dataset $(o[1], \ldots, o[n])$ if there exists $i \in \{1, \ldots, n\}$ such that $X \in O[i]$.

Definition 90. A random variable X is *hidden* in a dataset $(o[1], \ldots, o[n])$ if there is no $i \in \{1, \ldots, n\}$ such that $X \in O[i]$.

Elements of $\mathcal{S}$ should be defined over the set X of observable variables in d or a superset of X. Indeed, a distribution in a model over a set of variables that does not include X can not generate d.

A natural choice for $\mathcal{S}$ is the set $\mathcal{B}(X)$ of DAGs or the set $\mathcal{E}(X)$ of essential graphs over the observable variables. The latter choice is considered in Section 2.5.3 and Chapter 3. Depending on the application, prior knowledge, or computational constraints, smaller sets, such as $\mathcal{B}_k(X)$, may also be considered.

Learning is often harder when the graphs in $\mathcal{S}$ are defined over both observable and hidden variables. In that case, $\mathcal{S}$ is often restricted to special classes of DAGs. Let us present two such classes: *hierarchical latent class structures* and *Naive Bayes structures*.

Definition 91. A *directed tree* is a DAG such that every vertex has exactly one parent, except a single vertex, called the *root*, that has no parent.

Definition 92. A *leaf* in a directed tree is a vertex without child.

Definition 93. A *hierarchical latent class* (HLC) structure is a directed tree where all the vertices are hidden variables, except the leaves.

Example 28. Figure 1.13 is an HLC structure with hidden variables H_1, H_2, H_3 and observable variables $X_1, \ldots, X_7$.

Learning with HLC structures is discussed in [Zha04] (when the variables are discrete) and [SSGS06] (when the variables are continuous). The following special class of HLC structures is considered in Chapter 4.

Definition 94. A *Naive Bayes* (or *latent class*) structure is a directed tree where the root is the only hidden variable and is the only parent to each observable variable.

Example 29. The structure in Figure 1.12 is a Naive Bayes structure with observable variables X_1, X_2, X_3 and hidden variable H.

2.4 Parameter Learning

This section discusses how to choose a parameter based on its posterior density and how to compute the said density. The posterior parameter density is computed by combining the prior parameter density, sampling distribution, and marginal likelihood as follows:

$$p(\theta|d, G) = \frac{p(\theta|G)p(d|\theta, G)}{p(d|G)}. \tag{2.17}$$

First, several possible parameter estimates are proposed. Then, each term on the right-hand side of (2.17) is considered in turn and assumptions allowing to compute the proposed estimates exactly and efficiently are introduced gradually.

2.4.1 Parameter Estimation

The posterior parameter density provides a wealth of information. Let us present several ways to extract a single parameter that will be the result of our parameter learning procedure. In practice, the choice of estimate depends on the problem at hand.

As mentioned before, an obvious choice of estimate is the most probable parameter given the data.

Definition 95. A *maximum a posteriori* (MP) estimate θ_{MP} is a parameter value that maximizes the posterior parameter density, i.e.

$$\theta_{MP} \in \arg\max_{\theta \in \Theta_{d,G}} p(\theta|d,G). \tag{2.18}$$

In general, the existence and unicity of an MP estimate are not guaranteed. The marginal likelihood may be interpreted as a normalization constant irrelevant to the optimization:

$$\arg\max_{\theta \in \Theta_{d,G}} p(\theta|d,G) = \arg\max_{\theta \in \Theta_{d,G}} p(\theta|G)p(d|\theta,G). \tag{2.19}$$

Typically, the posterior parameter density becomes less dependent on the prior parameter density as the size of the dataset increases. In that case, MP estimates can be approximated by *maximum likelihood* (ML) estimates. For a fixed dataset d, the *likelihood* of a parameter θ is the probability $p(d|\theta,G)$. (The sampling distribution is a function defined on set of datasets for a fixed parameter.)

Definition 96. A *maximum likelihood* (ML) estimate θ_{ML} is a parameter value that maximizes the likelihood, i.e.

$$\theta_{ML} \in \arg\max_{\theta \in \Theta_{d,G}} p(d|\theta,G). \tag{2.20}$$

The existence and unicity of an ML estimate are not guaranteed. Unlike MP estimation, ML estimation does not use the prior parameter density and *no prior needs to be elicited.*

Another natural estimate is the expectation of the parameter θ w.r.t. its posterior density:

$$\langle\theta\rangle = \int_{\Theta_{d,G}} \theta\, p(\theta|d,G)d\theta. \tag{2.21}$$

If it exists, this estimate is unique.

Finally, if the learned parameter θ is to be used to predict future observations, the Bayesian network (G,θ) should approximate in some sense the predictive density

$$p(x|d,G) = \int_{\Theta_{d,G}} p(x|\theta,G)p(\theta|d,G)d\theta. \tag{2.22}$$

2.4.2 Likelihood

Since the observations are independent, the likelihood is given by

$$p(d|\theta, G) = \prod_{i=1}^{N} p_B(o[i]), \tag{2.23}$$

where $p_B = f_{d,G}(\theta)$ (see (2.1)). Depending on the dataset d and G, the complexity of the likelihood varies greatly. It has a simple expression under the *complete dataset* assumption.

Definition 97. An observation $o \in O$ is *complete* for a set of random variables Y if $O = Y$.

Definition 98. A dataset $d = (o[1], \ldots, o[n])$ is *complete* for Y if $O[i] = Y$ for all $i \in \{1, \ldots, n\}$.

Assumption 1 (*Complete dataset*). The dataset d is complete for X and G is defined over X, i.e. there is no hidden variable.

Under Assumption 1, the probability of each observation factorizes recursively according to G. Therefore, we have

$$\begin{aligned} p(d|\theta, G) &= \prod_{i=1}^{n} \prod_{v \in V} p_B(o[i]_v | o[i]_{pa(v)}), & (2.24)\\ &= \prod_{v \in V} \prod_{x_{pa(v)} \in \mathcal{X}_{pa(v)}} \prod_{x_v \in \mathcal{X}_v} (\theta_{x_v}^{X_v, x_{pa(v)}})^{n_{x_v}^{X_v, x_{pa(v)}}}, & (2.25)\end{aligned}$$

where $n_{x_v}^{X_v, x_{pa(v)}}$ denotes the number of observations o in the dataset such that $o_v = x_v$ and $o_{pa(v)} = x_{pa(v)}$. In that case, one can show that the components of a ML estimate satisfy

$$(\theta_{ML})_{x_v}^{X_v, x_{pa(v)}} = \frac{n_{x_v}^{X_v, x_{pa(v)}}}{n^{X_v, x_{pa(v)}}} \tag{2.26}$$

when $n_{x_v}^{X_v, x_{pa(v)}} > 0$ for $x_v \in \mathcal{X}_v$.

If the dataset is incomplete, for example because there are hidden variables, then $p(o[i]|\theta, G)$ may no longer factorize because of the marginalization.

Example 30. Consider a Naive Bayes structure G over $\{H\} \cup \{X_1, X_2, X_3\}$ and a dataset $d = (x[1], \ldots, x[n])$ such that H is hidden and each observation is complete for $\{X_1, X_2, X_3\}$. By Example 20, the probability of an observation $x = (x_1, x_2, x_3)$ is

$$p_B(x_1, x_2, x_3) = \sum_{h \in \mathcal{H}} \theta_h^H \prod_{j=1}^{3} \theta_{x_j}^{X_j, h}. \tag{2.27}$$

Hence, we have

$$p(d|\theta, G) = \prod_{i=1}^{n} \sum_{h \in \mathcal{H}} \theta_h^H \prod_{j=1}^{3} \theta_{x[i]_j}^{X_j, h}. \tag{2.28}$$

In the general case, the computation of ML estimates may be a difficult optimization problem and computationaly expensive methods may be necessary. Gradient-based methods are discussed in [Nea92] and [BKRK97]. Let us also mention the *expectation-maximization algorithm* (see [Lau95] and [NH98]). Although these algorithms only yield local maxima, they are widely used in practice.

2.4.3 Prior Parameter Density

In Bayesian inference, prior knowledge (or the lack thereof) about the parameter θ is represented by a prior parameter density $p(\theta|G)$. The exact choice of $p(\theta|G)$ is not discussed in this dissertation: we only introduce assumptions leading to closed-form parameter estimates and guaranteeing that the constraint (2.13) holds for all independence equivalent structures. These assumptions are thoroughly explained and justified in [HGC95].

The expression of the posterior parameter density is often simple if $p(\theta|G)$ belongs to a conjugate family for the sampling distribution.

Definition 99. Let $\mathcal{F} = f(\Theta)$ denote a family of distributions for a set X of random variables. A family $\mathcal{P}$ of prior densities on Θ is said to be a *conjugate family* for $\mathcal{F}$ if the posterior density $p(\theta|x) \in \mathcal{P}$ for all $x \in \mathcal{X}$ and $p(\theta) \in \mathcal{P}$.

If $\mathcal{P}$ is parametrized as $\mathcal{P} = g(\Lambda)$, an element $\lambda \in \Lambda$ is called a *hyperparameter*. Then, computing the posterior parameter density given a dataset complete for X amounts to updating a hyperparameter.

Example 31. Given $a, b \in \mathbb{R}$ and $\lambda^2 \in \mathbb{R}_{>0}$, let $\mathcal{F} = f(\Theta)$ be the statistical model for a real random variable X defined by

$$\Theta = \mathbb{R}, \tag{2.29}$$

$$f(\theta) = \mathcal{N}(x|a\theta + b, \lambda^2). \tag{2.30}$$

Let us show that the set of Gaussian densities is a conjugate family of prior parameter densities. We have

$$p(\theta|x) \propto p(\theta)p(x|\theta) \tag{2.31}$$

with

$$p(\theta) = \mathcal{N}(x|\mu, \sigma^2), \tag{2.32}$$

$$p(x|\theta) = f(\theta) = \mathcal{N}(x|a\theta + b, \lambda^2). \tag{2.33}$$

It is easy to see that

$$p(\theta|x) \propto e^{-\left(\theta - \frac{a\sigma^2(x-b)+\mu\lambda^2}{\sigma^2a^2+\lambda^2}\right)^2 / \left(2\frac{\sigma^2\lambda^2}{\sigma^2a^2+\lambda^2}\right)}. \tag{2.34}$$

Hence, the posterior $p(\theta|x)$ is the Gaussian density $\mathcal{N}(\theta|\mu', (\sigma')^2)$ with updated hyperparameters

$$\mu' = \frac{a\sigma^2(x-b) + \mu\lambda^2}{\sigma^2 a^2 + \lambda^2}, \tag{2.35}$$

$$(\sigma')^2 = \frac{\sigma^2\lambda^2}{\sigma^2 a^2 + \lambda^2}. \tag{2.36}$$

Using Dirichlet densities (see (A.3) in Appendix A), it is straighforward to design a family of prior parameter densities conjugate for the sampling distribution.

Assumption 2 (*Parameter independence*). The prior parameter density satisfies

$$p(\theta|G) = \prod_{v\in V} p(\theta^{X_v}|G), \tag{2.37}$$

$$p(\theta^{X_v}|G) = \prod_{x_{pa(v)}\in\mathcal{X}_{pa(v)}} p(\theta^{X_v,x_{pa(v)}}|G), \quad \text{for all } v\in V. \tag{2.38}$$

In the above assumption, the components of the parameter $\theta \in \Theta_{d,G}$ as labelled as

$$\theta = \Big(((\theta_{x_v}^{X_v,x_{pa(v)}})_{x_v\in\mathcal{X}_v})_{x_{pa(v)}\in\mathcal{X}_{pa(v)}}\Big)_{v\in V}, \tag{2.39}$$

and we let

$$\theta^{X_v,x_{pa(v)}} = (\theta_{x_v}^{X_v,x_{pa(v)}})_{x_v\in\mathcal{X}_v}, \tag{2.40}$$

$$\theta^{X_v} = (\theta^{X_v,x_{pa(v)}})_{x_{pa(v)}\in\mathcal{X}_{pa(v)}}. \tag{2.41}$$

Assumption 3 (*Dirichlet*). For $v \in V$ and $x_{pa(v)} \in \mathcal{X}_{pa(v)}$, the density $p(\theta^{X_v,x_{pa(v)}}|G)$ is Dirichlet with hyperparameters $(m_{x_v}^{X_v,x_{pa(v)}})_{x_v\in\mathcal{X}_v} \in S^+_{\mathcal{X}_v}$ and $\alpha^{X_v,x_{pa(v)}} > 0$.

Under Assumption 1, the family of prior parameter densities satisfying Assumption 2 and Assumption 3 is conjugate for the sampling distribution. The updated hyperparameters of $p(\theta^{X_v,x_{pa(v)}}|d,G)$ are given by

$$(\alpha^{X_v,x_{pa(v)}})' = \alpha^{X_v,x_{pa(v)}} + n^{X_v,x_{pa(v)}}, \tag{2.42}$$

$$(m_{x_v}^{X_v,x_{pa(v)}})' = \frac{\alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} + n_{x_v}^{X_v,x_{pa(v)}}}{\sum_{x_v\in\mathcal{X}_v} \alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} + n_{x_v}^{X_v,x_{pa(v)}}} \tag{2.43}$$

$$= \frac{\alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} + n_{x_v}^{X_v,x_{pa(v)}}}{\alpha^{X_v,x_{pa(v)}} + n^{X_v,x_{pa(v)}}}, \tag{2.44}$$

where $n^{X_v,x_{pa(v)}} = \sum_{x_v\in\mathcal{X}_v} n_{x_v}^{X_v,x_{pa(v)}}$ is the number of observations o in the dataset such that $o_{pa(v)} = x_{pa(v)}$.

Under Assumptions 1 to 3, the proposed parameter estimates can be computed exactly (see [HGC95]). The components of a MP estimate satisfy

$$(\theta_{MP})_{x_v}^{X_v,x_{pa(v)}} = \frac{n_{x_v}^{X_v,x_{pa(v)}} + \alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} - 1}{n^{X_v,x_{pa(v)}} + \alpha^{X_v,x_{pa(v)}} - |\mathcal{X}_v|} \tag{2.45}$$

if $n_{x_v}^{X_v,x_{pa(v)}} + \alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} > 1$ for all $x_v \in \mathcal{X}_v$. The components of the expected parameter $\langle\theta\rangle$ are given by

$$\langle\theta\rangle_{x_v}^{X_v,x_{pa(v)}} = (m_{x_v}^{X_v,x_{pa(v)}})' \tag{2.46}$$

with $(m_{x_v}^{X_v,x_{pa(v)}})'$ given by (2.43). Finally, we have

$$p(x|d,G) = f(\langle\theta\rangle) \in \mathcal{M}_d(G). \tag{2.47}$$

Hence, $(G,\langle\theta\rangle)$ perfectly represents the predictive density $p(x|d,G)$.

Together with Assumption 2 and Assumption 3, the following assumption ensures that the constraint (2.13) holds for all independence equivalent structures (see [HGC95]).

Assumption 4 (*Prior equivalence*). There exist a strictly positive distribution $q(x)$ and $\alpha > 0$ such that the hyperparameters of $p(\theta^{X_v,x_{pa(v)}}|G)$ satisfy

$$\alpha^{X_v,x_{pa(v)}} = \alpha, \tag{2.48}$$

$$m_{x_v}^{X_v,x_{pa(v)}} = q(x_v, x_{pa(v)}), \tag{2.49}$$

for all $G \in \mathcal{S}$, $v \in V$, $x_{pa(v)} \in \mathcal{X}_{pa(v)}$ and $x_v \in \mathcal{X}_v$.

2.4.4 Marginal Likelihood and Posterior Density

The marginal likelihood may be interpreted as a normalization constant guaranteeing that the posterior density integrates to one:

$$p(d|G) = \int_{\Theta_{d,G}} p(\theta|G)p(d|\theta,G)d\theta. \tag{2.50}$$

If the posterior density does not belong to some well-known family, it may be necessary to compute the marginal likelihood explicitely.

When the exact computation of the marginal likelihood and posterior parameter density are infeasible, it may be simpler to use MP or ML parameter estimates since they do not require this computation, but we may also resort to approximations methods. For example, *Monte Carlo* methods may be used to estimate the marginal likelihood (see e.g. [Mac03] for an introduction). Also, the posterior density can be approximated by a simpler density (see e.g. [Cow98]), in particular for large sample sizes. The *Laplace approximation* is a large-sample approximation where $p(\theta|d,G)$ is approximated by a multivariate Gaussian density. It can be intuitively described as follows (see [CH97] for more details). Suppose the posterior parameter density is sufficiently smooth and attains its maximum over the parameters for a *unique* value $\theta_{MP} \in \Theta_{d,G}$. This parameter also maximizes

$$g(\theta) = \ln(p(d|\theta, G)p(\theta|G)). \tag{2.51}$$

Let us approximate g around θ_{MP} by its second degree Taylor polynomial

$$g(\theta) \approx g(\theta_{MP}) + \frac{1}{2}(\theta - \theta_{MP})^T H(\theta - \theta_{MP}), \tag{2.52}$$

where H is the Hessian matrix of g evaluated at θ_{MP}. Hence, we have

$$p(d|\theta, G)p(\theta|G) \approx p(d|\theta_{MP}, G)p(\theta_{MP}|G)e^{-\frac{1}{2}(\theta-\theta_{MP})^T(-H)(\theta-\theta_{MP})}, \tag{2.53}$$

and $p(\theta|d, G) \propto p(d|\theta, G)p(\theta|G)$ can be approximated by a Gaussian density.

Remark 47. Typically, the quality of this approximations increases with the number of observations. Unfortunately, it also depends on the parametrization chosen for the model (see [Mac98]).

Large-sample approximations for the marginal likelihood are further discussed in Section 2.5.1.

2.5 Structure Learning

The posterior structure distribution $p(G|d)$ with $G \in \mathcal{S}$ represents our uncertainty about the structures after observing the dataset. A structure $G \in \mathcal{S}$ with maximum posterior probability is selected as a solution of our structure learning problem:

$$G \in \arg\max_{G \in \mathcal{S}} p(G|d) \tag{2.54}$$

This learning approach is an example of a *score-based* (or *metric*) approach where a scoring criterion that ranks the structures is defined and a high scoring structure is searched.

Section 2.5.1 discusses the computation of the posterior structure probabilities. In a more general perspective, Section 2.5.2 introduces several properties of scoring criteria that are relevant to learning. Section 2.5.3 discusses the search for a structure that maximizes the chosen score.

2.5.1 Posterior Structure Probability

The posterior probability of a structure is computed by combining its prior probability, the marginal likelihood, and the probability of observing the dataset:

$$p(G|d) = \frac{p(G)p(d|G)}{p(d)}. \tag{2.55}$$

The term $p(d)$ does not depend on the structure G. It is thus irrelevant to the optimisation in (2.54) and we have

$$\arg\max_{G\in\mathcal{S}} p(G|d) = \arg\max_{G\in\mathcal{S}} p(G)p(d|G). \tag{2.56}$$

The elicitation of prior structure probabilities is not discussed in this dissertation. Let us simply recall that structures defining the same hypothesis should have the same prior probability (see (2.11)). As the dataset size increases, it is worth noting that the posterior $p(G|d)$ typically becomes less dependent on the prior and approximately proportional to the marginal likelihood $p(d|G)$.

Under Assumptions 1 to 3, the marginal likelihood can be computed exactly and in closed-form (see [HGC95]):

$$\begin{aligned} p(d|G) = \prod_{v\in V} \prod_{x_{pa(v)}\in\mathcal{X}_{pa(v)}} & \frac{\Gamma(\alpha^{X_v,x_{pa(v)}})}{\Gamma(\alpha^{X_v,x_{pa(v)}} + n^{X_v,x_{pa(v)}})} \\ & \prod_{x_v\in\mathcal{X}_v} \frac{\Gamma(\alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}} + n_{x_v}^{X_v,x_{pa(v)}})}{\Gamma(\alpha^{X_v,x_{pa(v)}} m_{x_v}^{X_v,x_{pa(v)}})}. \end{aligned} \tag{2.57}$$

The following scoring criterion is based on the above expression.

Definition 100. The *Bayesian Dirichlet* (BD) score is defined by

$$\mathrm{BD}(G,d) = p(G)p(d|G), \tag{2.58}$$

with $p(d|G)$ given by (2.57).

Remark 48. If Assumption 4 holds, the Bayesian Dirichlet score is called the *Bayesian Dirichlet likelihood-equivalent* (BDe) score.

As discussed in Section 2.4.4, the exact computation of the marginal likelihood is often intractable in the general case. Let us present some large-sample approximations. Under the complete dataset assumption and some additional regularity assumptions on the priors, we have

$$\ln p(d|G) = \ln p(d|\theta_{ML}, G) - \frac{1}{2} d(\mathcal{M}_d(G)) \ln n + O(1), \tag{2.59}$$

for $n \to \infty$ (see [GHKM01] and [Hau88]). Similarly, if d is complete for X, H is a set of hidden variables, G is a structure over $X \cup H$, and some regularity assumptions hold, we have

$$\ln p(d|G) = \ln p(d|\theta_{ML}, G) - \lambda \ln n + (m-1) \ln\ln n + O(1), \tag{2.60}$$

for $n \to \infty$ (see [RG05]). In this expression, λ is a rational number less than or equal to half the standard dimension of $\mathcal{M}_d(G)$ and m is an integer greater than or equal to 1. In general, both λ and m depend on d and their computation is non-trivial.

Based on these results, several other scoring criteria have been proposed.

Definition 101. The *Bayesian information criterion* (BIC) score is defined by

$$\mathrm{BIC}(G,d) = \ln p(d|\theta_{ML},G) - \frac{1}{2}d(\mathcal{M}_d(G))\ln n. \tag{2.61}$$

The BIC score is often used when G has no hidden variables and the prior parameter density is not conjugate. If $G \in \mathcal{B}(X)$ and d is complete for X, its validity stems from (2.59).

Let d be a dataset such that X is the set of observable variables and H is a set of hidden variables and let $G \in \mathcal{B}(X \cup H)$. In [GHM96], the authors propose to adjust the BIC score to

$$\ln p(d|\theta_{ML},G) - \frac{1}{2}d(\mathcal{M}_H)\ln n, \tag{2.62}$$

where $\mathcal{M}_H$ is the Bayesian network model with hidden variables H associated to $\mathcal{M} = \mathcal{M}_d(G)$ (see Section 1.7.2). Based on (2.60), the authors of [RG03] suggest to approximate $\ln p(d|G)$ by

$$\ln p(d|\theta_{ML},G) - \lambda \ln n + (m-1)\ln\ln n, \tag{2.63}$$

and propose an algorithm computing λ and m. This algorithm is not described fully here, but let us mention a special case: if the preimage $f_G^{-1}(f_G(\theta_{ML}))$ has dimension 0, it returns $\lambda = \frac{1}{2}d(\mathcal{M}_d(G))$ and $m = 1$. In that case, (2.63) coincides with the BIC score.

2.5.2 Properties of Scoring Criteria

This section introduces properties of scoring criteria that influence the characteristics of the search algorithms presented in Section 2.5.3. As we will see, the marginal likelihood satisfies all these properties under appropriate assumptions.

The *equivalence* property ensures that a scoring criterion defined on structures extends to Bayesian network models and essential graphs.

Definition 102 (*equivalence*). A scoring criterion score(G,d) is *equivalent* if it assigns the same value to independence equivalent structures, i.e.

$$G =_I H \Rightarrow \mathrm{score}(G,d) = \mathrm{score}(H,d). \tag{2.64}$$

Hypotheses associated to Markov equivalent structures are equivalent. Hence, the posterior structure probability and the marginal likelihood define score equivalent criteria.

The *decomposability* property affects the computational efficiency of the search algorithms (see Example 34 in Section 2.5.3).

Definition 103 (*decomposability*). A scoring criterion score(G,d) is *decomposable* if it can be decomposed as a sum of functions such that each function depends only on one variable and its parents.

Remark 49. A decomposable scoring criterion is written as

$$\text{score}(G, d) = \sum_{v \in V} f(v, pa_G(v)), \tag{2.65}$$

where the dependence on the dataset d is not shown explicitely.

Under Assumptions 1 to 3, the logarithm of the marginal likelihood is decomposable. If $\ln p(G)$ is decomposable, then $\ln \text{BD}$ is also decomposable.

The *consistency* property describes scoring criteria defined on statistical models. A learning procedure that selects a model maximizing a consistent criterion $\text{score}(\mathcal{M}, d)$ over a set M of statistical models asymptotically returns a parameter optimal model if the generative distribution belongs to at least one model in M.

Definition 104 (*consistency*). Let d be a sequence of n observations from a generative distribution g and let M be a class of statistical models whose dimension is defined. A scoring criterion $\text{score}(\mathcal{M}, d)$ for $\mathcal{M} \in M$ is said to be *(asymptotically) consistent* if

- $g \in \mathcal{M}_1$ and $g \notin \mathcal{M}_2$, or
- $g \in \mathcal{M}_1 \cap \mathcal{M}_2$ and $d(\mathcal{M}_1) < d(\mathcal{M}_2)$

imply that $\text{score}(\mathcal{M}_1, d) > \text{score}(\mathcal{M}_2, d)$ with probability one for $n \to \infty$.

Under Assumption 1 and other mild hypotheses about the prior parameter densities, the marginal likelihood is an asymptotically consistent scoring criterion for a set M of discrete Bayesian network models (see [Hau88] and [GHKM01] for details).

The last property is a local version of the asymptotic consistency property. It is required to guarantee an important property of the UGES structure search algorithm presented in Section 2.5.3.

Definition 105 (*local consistency*). Let d be a sequence of n observations from a strictly positive generative distribution g. Let G and H be DAGs such that H is obtained from G by adding the arrow $X_i \to X_j$. A scoring criterion $\text{score}(G, d)$ for $G \in \mathcal{B}(X)$ is said to be *locally consistent* if

$$X_i \perp X_j | X_{pa_G(j)} \in I(g) \Rightarrow \text{score}(G, d) > \text{score}(H, d) \tag{2.66}$$

$$X_i \perp X_j | X_{pa_G(j)} \notin I(g) \Rightarrow \text{score}(H, d) > \text{score}(G, d) \tag{2.67}$$

with probability one for $n \to \infty$.

Decomposability and consistency imply local consistency.

Proposition 2.1. *If a scoring criterion* $\text{score}(G, d)$ *is decomposable and consistent, then it is locally consistent.*

Proof. This proof is adapted from [Chi02b]. Let d be a sequence of n observations from a strictly positive generative distribution g. Let G and H be DAGs such that H is obtained from G by adding the arrow $X_i \to X_j$.

To begin let us introduce a few preliminary notions. If $X_{j_1}, \ldots, X_{j_k}$ are the parents of X_j in G, let o be a total ordering of X starting with $X_{j_1}, \ldots, X_{j_k}, X_i, X_j$. Let H' be the complete DAG over X such that $X_u \to X_v \in H'$ if X_u precedes X_v in o. Let G' be the DAG obtained from H' by removing $X_i \to X_j$. It is easy to see that the following observations hold:

- $g \in \mathcal{M}_d(H')$ (by completeness of H', see Section 1.5.3);
- $pa_{G'}(X_j) = pa_G(X_j)$;
- $\text{score}(G,d) - \text{score}(H,d) = \text{score}(G',d) - \text{score}(H',d)$ (by decomposability);
- $I(G') = \left\{(X_i \perp X_j | pa_{G'}(X_j)), (X_j \perp X_i | pa_{G'}(X_j))\right\}$.

The proof can be decomposed as follows.

1. Suppose that $X_i \perp X_j | X_{pa_G(j)} \in I(g)$. This implies $I(G') \subseteq I(g)$, and thus $g \in \mathcal{M}_d(G')$. By consistency of the score and the observation $d(\mathcal{M}_d(G') < d(\mathcal{M}_d(H')$, $\text{score}(G',d) - \text{score}(H',d) = \text{score}(G,d) - \text{score}(H,d) > 0$ with probability one for $n \to \infty$.
2. Suppose that $X_i \perp X_j | X_{pa_G(j)} \notin I(g)$. This implies $g \notin \mathcal{M}_d(G')$. By consistency of the score, we have $\text{score}(G',d) - \text{score}(H',d) = \text{score}(G,d) - \text{score}(H,d) < 0$ with probability one for $n \to \infty$. □

2.5.3 Search for an Optimal Structure

This section discusses the search for a structure maximizing the scoring criterion $\text{score}(G,d)$ over a set $\mathcal{S}$ of structures. For simplicity, only structures without hidden variable are considered. For the case of structures with hidden variables, see e.g. [Zha04], [Fri98], [ELFK00] and [EF01]. First, several results highlighting the complexity of the task are presented. Then, greedy search algorithms are introduced.

Some Complexity Results

Even without hidden variable, finding an optimal structure may be a hard problem. Consider the cases where $\mathcal{S} = \mathcal{B}(X)$ or $\mathcal{S} = \mathcal{B}_k(X)$. As shown in Section 1.8, the size of $\mathcal{S}$ increases superexponentially with $|X|$. Consequently, searching for an optimal structure by enumeration is likely to be impractical. For $\mathcal{S} = \mathcal{B}_k(X)$ with $k \geq 2$ and the BDe scoring criterion, the following complexity results hold (see [Chi95]).

Theorem 2.2. *Let d be a dataset complete for a set X of discrete variables. Given $k \geq 2$ and $q \in \mathbb{R}$, the problem of deciding if there exists a structure $G \in \mathcal{B}_k(X)$ such that* $\mathrm{BDe}(G, d) \geq q$ *is NP-complete.*

If d is a dataset complete for a set X of discrete variables and $k \geq 2$, identifying a structure that maximizes the BDe score over $\mathcal{B}_k(X)$ is thus NP-hard.

For $\mathcal{S} = \mathcal{B}_k(X)$ with $k \geq 3$ and a consistent scoring criterion, Theorem 1.18 has the following immediate corollary.

Corollary 2.3. *Let d be a dataset of size n complete for a set X of discrete variables, let* $\mathrm{score}(G, d)$ *be a consistent scoring criterion, and let k be an integer* ≥ 3. *With probability one as $n \to \infty$, the problem of identifying a structure that maximizes* $\mathrm{score}(G, d)$ *over $\mathcal{B}_k(X)$ is NP-hard.*

Greedy Search

Heuristic algorithms are often used to search for optimal structures. Because of their simplicity and because they often serve as a basis for more complicated methods, let us present greedy search algorithms.

Given a structure space $\mathcal{S}$, a *neighborhood* N assign to each structure $G \in \mathcal{S}$ a finite subset of $\mathcal{S}$.

Example 32. For $G \in \mathcal{S} = \mathcal{B}(X)$, $N(G)$ may be defined as the set of DAGs over X that can be obtained from G by adding, removing or inverting a single arrow.

Starting from an initial structure $G_0 \in \mathcal{S}$, a greedy search algorithm explores the structure space iteratively by moving from a current structure G to the highest scoring neighbor $G' \in N(G)$ until a local optimum of the scoring metric is reached.

Algorithm 1 (Greedy search)

```
1. Set $G := G_0$ and $stop := false$.
2. While $stop = false$:
    (a) Find $G' \in \arg\max_{H \in N(G)} \mathrm{score}(H, d)$.
    (b) If $\mathrm{score}(G', d) > \mathrm{score}(G, d)$, set $G := G'$. Otherwise, set $stop := true$.
3. Return $G$.
```
□

Note that Step 2(a) can be performed by enumeration. Also, Algorithm 1 terminates because $\mathrm{score}(G, d)$ strictly increases at each iteration and $\mathcal{S}$ is finite.

The output of Algorithm 1 is a *local* maximum of the scoring criterion in the sense defined by the neighborhood. Algorithm 1 may thus return a suboptimal element of $\mathcal{S}$. In fact, the combination of scoring criterion, structure space, neighborhood, and initial structure determine the properties of the greedy algorithm. The neighborhood N influences whether the scoring criterion has local maxima or plateaux that may trap the greedy algorithm, but also the computational complexity of the algorithm.

Example 33. If $\mathcal{S} = \mathcal{B}(X)$, let $N(G) = \mathcal{B}(X) \setminus \{G\}$. With this neighborhood, local maxima are also global. Hence, a greedy search will return an optimal solution. However, Step 2(a) of Algorithm 1 is equivalent to our original problem.

Example 34. Suppose that N is the same as in Example 32. Step 2(a) of Algorithm 1 can be performed by enumeration. If the scoring criterion is decomposable, the computation of the score of a neighbor can be performed incrementally, i.e. the difference $\text{score}(H, d) - \text{score}(G, d)$ for $H \in N(G)$ can be evaluated easily. For example, if $H \in N(G)$ is obtained by adding $X_u \rightarrow X_v$ to G, then we have

$$\text{score}(H, d) - \text{score}(G, d) = f(v, pa_H(v)) - f(v, pa_G(v)), \tag{2.68}$$

where $pa_H(v) = pa_G(v) \cup \{u\}$.

Neighborhoods may also be defined over sets of essential graphs. Similarly to the neighborhood on structures defined in Example 32, the set of neighbors $N(E)$ of an essential graph $E \in \mathcal{E}(X)$ can be defined by local edges modifications, but these modifications are not as straightforward (see [Chi02a]). Indeed, the addition or removal of an edge or the inversion of an arrow in E does not always produce another essential graph. Alternatively, a neighborhood may be defined in terms of the independence models represented by essential graphs as follows (see [KC01]).

Definition 106. An essential graph G belongs to the *inclusion boundary* $\text{IB}(E)$ of an essential graph $E \in \mathcal{E}(X)$ if, and only if, $G \in \mathcal{E}(X)$ and one of the following conditions holds:

(a) $G <_I E$ and there is no $H \in \mathcal{E}(X)$ such that $G <_I H <_I E$

(b) $E <_I G$ and there is no $H \in \mathcal{E}(X)$ such that $E <_I H <_I G$.

Example 35. The inclusion boundary of an essential graph over $X = \{X_1, X_2, X_3\}$ can be read off Figure 1.10. For example, the inclusion boundary of $X_1 - X_3 \quad X_2$ is composed of $X_1 \quad X_2 \quad X_3$, $X_3 \rightarrow X_1 \leftarrow X_2$, $X_3 - X_1 - X_2$ and $X_1 \rightarrow X_3 \leftarrow X_2$.

The efficient computation of the inclusion boundary $\text{IB}(E)$ and the score of its elements is the topic of Chapter 3.

The *unrestricted greedy equivalence search* (UGES) algorithm is the greedy search algorithm over $\mathcal{E}(X)$ defined by $N(E) = \text{IB}(E)$. It has the following property (from [CM02]).

Theorem 2.4. *Let d be a sequence of n observations from a generative distribution g that satisfies the composition property* (1.44) *and let M be the set of all discrete BN models for X. With probability one as $n \rightarrow \infty$, UGES starting from any $E \in \mathcal{E}(X)$ and using a score equivalent and locally consistent scoring criterion returns an essential graph E such that $\mathcal{M}_d(E)$ is inclusion optimal w.r.t. M and g.*

Suppose that a distribution p is faithfull to $D \in \mathcal{B}(X)$. Then, p satisfies the composition property (see Section 1.5.2) and faithfulness, parameter optimality, and inclusion optimality are equivalent by Proposition 1.16. Hence, Theorem 2.4 has the following corollary.

Corollary 2.5. *Let d be a sequence of n observations from a generative distribution g faithfull to $E \in \mathcal{E}(X)$. With probability one for $n \to \infty$, UGES starting from any essential graph in $\mathcal{E}(X)$ and using a score equivalent and locally consistent scoring criterion returns E.*

Variants of UGES and further details can be found in [CM02], [Chi02b], [CK03], and [NKP03].

Chapter 3

Computation of the Inclusion Boundary

3.1 Introduction

The inclusion boundary is the neighborhood used in the unrestricted greedy equivalence search algorithm (see Section 2.5.3). By Definition 106, recall that an essential graph G belongs to the inclusion boundary $\mathrm{IB}(E)$ of $E \in \mathcal{E}(V)$ if, and only if, $G \in \mathcal{E}(V)$ and one of the following conditions holds:

(a) $G <_I E$ and there is no $H \in \mathcal{E}(V)$ such that $G <_I H <_I E$

(b) $E <_I G$ and there is no $H \in \mathcal{E}(V)$ such that $E <_I H <_I G$.

Unfortunately, it is not easy to enumerate $\mathrm{IB}(E)$ with the above description. Using the results presented in Section 1.6, this chapter introduces algorithms that efficiently compute the inclusion boundary. The boundary is computed with a divide and conquer approach: it is partitioned until each element of the partition is sufficiently simple to be enumerated. The boundary is first partitioned according to the skeleton of its elements and then according to their v-structures. This contribution was published in an earlier form in [AW02]. Independently of our original work, the computation of the inclusion boundary is also carried out in [Chi02b]. Additionally, the inclusion boundary is characterized in [Stu05].

With a greedy algorithm, it is important to efficiently compute the difference in score between essential graphs adjacent in the search space. This chapter also demonstrates that the decomposability property of the scoring criterion can be exploited to that end.

Section 3.2 introduces notions necessary to manipulate essential graphs. Section 3.3 tackles the computation of the boundary.

3.2 From DAGs to Essential Graphs and Vice-Versa

Although elegant, the definition of essential graphs given in Section 1.6.2 is not very practical: a naive computation of the essential graph associated to a Markov equivalence class of structures C requires the knowledge of all the elements of C. This is problematic because the cardinality of C may be very large. Also, C may only be specified by one of its elements and the computation of the complete class C may not be easy. Finally, the definition of essential graphs does not make their graphical properties apparent. Consequently, it may be difficult to determine whether a given graph is essential.

These problems are addressed in the following sections. Section 3.2.1 provides a graphical characterization of essential graphs. Section 3.2.2 describes the elements of a Markov equivalence class represented by its essential graph. Section 3.2.3 presents an algorithm to compute the essential graph associated to a Markov equivalence class represented by one of its elements.

3.2.1 Graphical Characterization of Essential Graphs

This section introduces graphical properties that characterize essential graphs. This characterization and the notions introduced here are used throughout this chapter. Hence, it is important for the reader to become familiar with them.

Preliminary Notions

Definition 107. The *subgraph of $G = (V, E)$ induced by a non-empty set $A \subseteq V$* is the graph $G_A = (A, E \cap (A \times A))$.

EXAMPLE 36. If G is the graph given in Figure 3.1, then $G_{\{a,b,c,d\}}$ and $G_{\{f,i,j\}}$ are given in Figure 3.2.

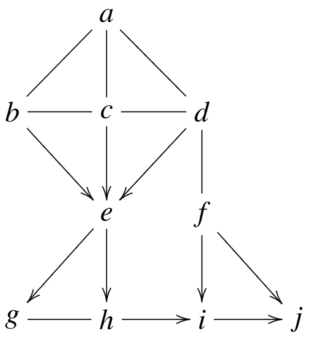
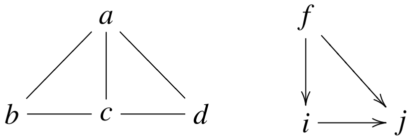


Figure 3.1: A graph G

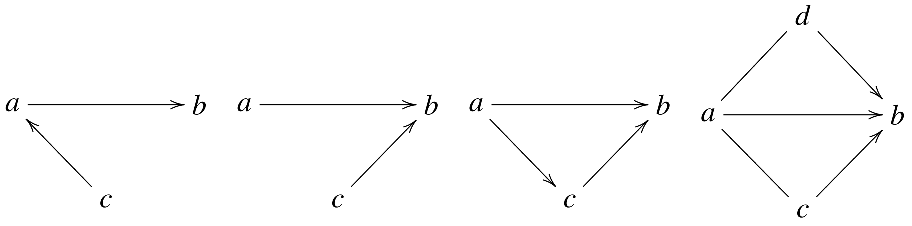


(a) $G_{\{a,b,c,d\}}$ (b) $G_{\{f,i,j\}}$

Figure 3.2: Subgraphs of G

Definition 108. An arrow $a \to b \in G$ is *strongly protected* if

- there exist vertices c, d such that $G_{\{a,b,c,d\}}$ is given in Figure 3.3(d) or
- there exists a vertex c such that $G_{\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c).



(a) (b) (c) (d)

Figure 3.3: Subgraphs that strongly protect $a \to b \in G$

Example 37. If G is the graph given in Figure 3.1, then

- $h \to i$ is strongly protected by $G_{\{e,h,i\}}$ and $G_{\{f,h,i\}}$
- $b \to e$ and $d \to e$ are strongly protected by $G_{\{b,d,e\}}$
- $f \to j$ is strongly protected by $G_{\{f,i,j\}}$
- $c \to e$ is strongly protected by $G_{\{b,c,d,e\}}$.

The following class of graphs encompasses DAGs and undirected graphs.

Definition 109. A graph G is a *chain graph* if it has no directed cycle.

Example 38. The graph G given in Figure 3.1 is a chain graph that is neither undirected nor a DAG. The graph H given in Figure 3.4 is not a chain graph because of the directed cycle a, b, c, a.

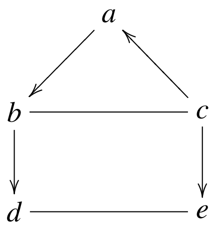


Figure 3.4: A graph H that is not a chain graph

Definition 110. If G is a graph, let $\rightleftharpoons$ be the relation between vertices defined by $a \rightleftharpoons b$ if, and only if, $a = b$ or G has a path from a to b and a path from b to a.

Remark 50. The relation $\rightleftharpoons$ is an equivalence relation.

Example 39. The equivalence classes induced by $\rightleftharpoons$ on the vertex set of the graph of Figure 3.1 are $\{a, b, c, d, f\}$, $\{e\}$, $\{g, h\}$, $\{i\}$, and $\{j\}$. The equivalence classes induced by $\rightleftharpoons$ on the vertex set of the graph of Figure 3.4 are $\{a, b, c\}$ and $\{d, e\}$.

Definition 111. The *chain components* of a chain graph G are the equivalence classes of vertices induced by $\rightleftharpoons$.

Remark 51. A graph G is a chain graph if, and only if, the subgraphs of G induced by the equivalence classes induced by $\rightleftharpoons$ are undirected.

Example 40. The graph G given in Figure 3.1 is a chain graph because $G_{\{a,b,c,d,f\}}$, $G_{\{e\}}$, $G_{\{g,h\}}$, $G_{\{i\}}$ and $G_{\{j\}}$ are undirected. The graph H given by Figure 3.4 is not a chain graph because $H_{\{a,b,c\}}$ is not undirected.

Definition 112. An undirected graph is *chordal* if every cycle of length $n \geq 4$ has a chord, i.e. a line between two non-consecutive vertices in the cycle.

Example 41. Let G be given in Figure 3.1. The only cycle of $G_{\{a,b,c,d,f\}}$ of length ≥ 4 is a, b, c, d, a, and it has the chord $c - d$. Hence, $G_{\{a,b,c,d,f\}}$ is chordal. The subgraph $G_{\{g,h\}}$ is also chordal since it has no cycle of length ≥ 4. The graph given in Figure 3.5 is not chordal since the cycle a, b, d, c, a has no chord.

Characterization

The following theorem characterizes graphically essential graphs (see [AMP97]).

Theorem 3.1. *A graph $G = (V, E)$ is an essential graph if, and only if, the following propositions hold:*

(a) G is a chain graph

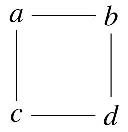


Figure 3.5: A non-chordal graph

(b) for every chain component τ of G, G_τ is chordal

(c) there is no $\{a, b, c\} \subseteq V$ such that $G_{\{a,b,c\}}$ is the subgraph $a \rightarrow b - c$

(d) every arrow in G is strongly protected.

Example 42. The graph given in Figure 3.1 is essential.

3.2.2 Markov Equivalence Class Associated to an Essential Graph

A Markov equivalence class of structures may be represented by its essential graph. This section describes the elements of such an equivalence class using the notions of consistent extension and perfect ordering. Also, it introduces the maximum cardinality search algorithm. This algorithm is used in Section 3.3.2 and Section 3.3.3 to compute structures in the equivalence class. In particular, it is used as a subroutine of Algorithms 4 to 7.

Description of $[E]$

Definition 113. A DAG D is a *consistent extension* of a graph G if they have the same skeleton and v-structures and every arrow in G is also in D.

Example 43. The graphs given in Figure 3.6 are consistent extensions of the graph $G_{\{a,b,c,d\}}$ given in Figure 3.2(a). The graphs given in Figure 3.4 and Figure 3.5 have

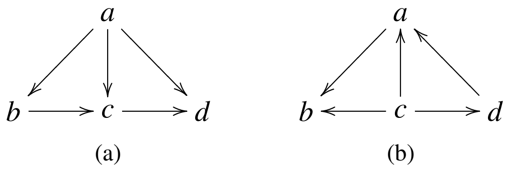


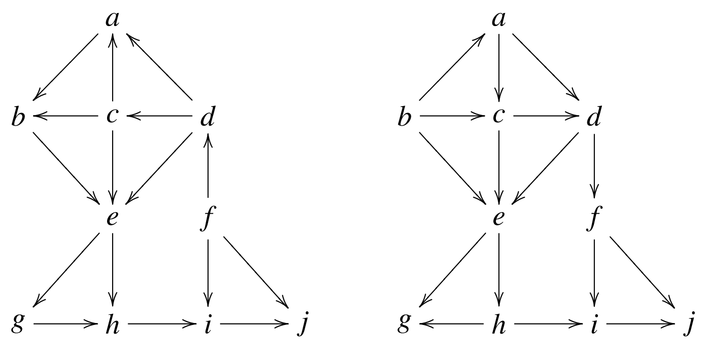
Figure 3.6: Consistent extensions of $G_{\{a,b,c,d\}}$

no consistent extension.

An algorithm that checks whether a graph has a consistent extension, and provides one if it does, is given in [DT92].

If $D \in \mathcal{B}(V)$ and $E \in \mathcal{E}(V)$, then $D \in [E]$ if, and only if, D is a consistent extension of E. Hence, $[E]$ can be described with the notion of consistent extension.

Example 44. The graphs given in Figure 3.7 are consistent extensions of the essential graph G given in Figure 3.1. Hence, they belong to $[G]$.



(a) (b)

Figure 3.7: Consistent extensions of G

The consistent extensions of an undirected graph can be described with the notion of perfect ordering.

Definition 114. Let $G = (V, E)$ be an undirected graph, let o be a total ordering of V, and let D be the DAG with vertex set V and edges such that $a \rightarrow b \in D$ if $a - b \in G$ and a precedes b in o. The ordering o is a *perfect ordering* of G if D is a consistent extension of G.

Example 45. If $G_{\{a,b,c,d\}}$ is given in Figure 3.2(a), the perfect ordering a, b, c, d leads to the consistent extension of $G_{\{a,b,c,d\}}$ given in Figure 3.6(a) and the perfect ordering c, d, a, b leads to the consistent extension of $G_{\{a,b,c,d\}}$ given in Figure 3.6(b).

Definition 115. If $D = (V, E)$ is a DAG, a total ordering o of V is an *ancestral ordering of D* if the existence of a path in D from u to v implies that u precedes v.

Example 46. The ordering a, b, c, d is an ancestral ordering of the DAG given in Figure 3.6(a). The ordering $f, d, c, a, b, e, g, h, i, j$ is an ancestral ordering of the DAG given in Figure 3.7(a).

It is worth noting that a DAG admits at least one ancestral ordering. Also, an ancestral ordering of a consistent extension of a graph G is a perfect ordering of G.

The notions of perfect ordering and chordality are connected by the following theorem (see [CDLS99]).

Theorem 3.2. *An undirected graph is chordal if, and only if, it admits at least one perfect ordering.*

Example 47. The graph given in Figure 3.5 has no consistent extension because it is not chordal. The graph given in Figure 3.2(a) admits a perfect ordering (see Example 45) and is thus chordal.

The elements of an equivalence class of structures specified by an essential graph are described by the following theorem (see [AMP97]).

Theorem 3.3. *If $D \in \mathcal{B}(V)$ and $E \in \mathcal{E}(V)$, then $D \in [E]$ if, and only if, D is obtained from E by orienting the lines of every subgraph induced by a chain component of E according to a perfect ordering.*

Remark 52. Theorem 3.3 is rather easy to prove with the following observations. Let D be obtained from E by orienting the lines of every subgraph induced by a chain component of E according to a perfect ordering. First, the orientation of the lines does not create any v-structure in D that was not in E because the orderings used are perfect and E does not induce a subgraph of the type $a \to b - c$. Second, no cycle is created inside a chain component because the orderings are perfect or across chain components because E is a chain graph.

Computation of Some Elements of $[E]$

In practice, a perfect ordering of a chordal undirect graph G can be obtained with the *maximum cardinality search* (MCS) algorithm (see [CDLS99]). To present the MCS algorithm, let us introduce the following definitions.

Definition 116. If $G = (V, E)$ is a graph, the set $ne_G(a)$ of *neighbors* of $a \in V$ is

$$ne_G(a) = \{b \in V \big| a - b \in G\}. \tag{3.1}$$

Definition 117. A set of vertices $c \neq \emptyset$ is *complete in a graph G* if G_c is complete.

Remark 53. A set of vertices with one element is complete.

The MCS algorithm takes for input an undirected graph G and determines whether G is chordal by searching for a perfect ordering.

Algorithm 2 (Maximum cardinality search)

```
1. Set $output$ := '$G$ is chordal', counter $i := 0$, $L := \phi$, and $c(v) := 0$ for all $v \in V$.
2. While $L \neq V$ and $output$ = '$G$ is chordal':
    (a) Set $U := V \setminus L$ and $i := i + 1$.
    (b) Select a vertex $v_i$ maximizing $c(v)$ over $v \in U$.
    (c) If $ne(v_i) \cap L$ is not complete in $G$, set $output$ := '$G$ is not chordal'.
        Otherwise, set $c(v) := c(v) + 1$ for each $v \in ne(v_i) \cap U$.
    (d) Set $L := L \cup \{v_i\}$.
3. Return $output$ and $v_1, \ldots, v_i$.
```

□

The following theorem holds (see [CDLS99]).

Theorem 3.4. *If G is chordal, then MCS terminates with $output$ = 'G is chordal' and $v_1, \ldots, v_{|V|}$ is a perfect ordering of G. If G is not chordal, then MCS terminates with $output$ = 'G is not chordal'.*

The MCS algorithm is non-deterministic as any vertex maximizing $c(v)$ over $v \in U$ can be chosen at Step 2(b). This non-determinism has a useful consequence.

Lemma 3.5. *If $G = (V, E)$ is a chordal undirected graph and $c \subseteq V$ is complete in G, there exists a perfect ordering of G starting with any permutation of c.*

Proof. Let $c_1, \ldots, c_k$ be a permutation of c. By Theorem 3.4, it is sufficient to show there exists an execution of Algorithm 2 such that $v_i = c_i$ for $i = 1, \ldots, k$.

Since $c(v) = 0$ for all $v \in V$, we can choose v_1 arbitrarily, and we select $v_1 = c_1$. At the current step of the Algorithm, suppose that we have to select v_j with $j \leq k$ and that $v_i = c_i$ for all $1 \leq i < j$. Because G_c is complete, it is easy to see that $c(c_j) = j - 1 = \max_{v \in U} c(v)$. Hence, we can choose $v_j = c_j$. □

3.2.3 Computation of the Essential Graph Associated to the Markov Equivalence Class of a DAG

This section presents an algorithm that computes the essential graph associated to a Markov equivalence class of structures given by one of its elements. Then, the applicability of the algorithm is extended to allow for the incorporation of partial knowledge about the essential graph. The algorithm is used in Section 3.3.2 and Section 3.3.3.

Definition 118. If $D \in \mathcal{B}(V)$, let $E(D) \in \mathcal{E}(V)$ be such that $D \in [E(D)]$.

The following algorithm can be used to compute $E(D)$ (see [AMP97]). It takes for input a graph G and returns a graph G_i.

Algorithm 3

```
1. Set $G_0 := G$, $i := 0$, and $stop := false$.
2. While $stop = false$:
    (a) Set $G_{i+1}$ to the graph obtained from $G_i$ by converting every arrow that
        is not strongly protected into a line.
```

(b) If $G_i = G_{i+1}$, set *stop* := *true*. Otherwise, set $i := i + 1$. □

The following theorem holds (see [AMP97]).

Theorem 3.6. *Algorithm 3 applied to $G \in \mathcal{B}(V)$ returns $E(G)$.*

Remark 54. An alternative algorithm to compute $E(D)$ is given in [Chi02b].

The following proposition extends the applicability of Algorithm 3.

Proposition 3.7. *If $D = (V, E_D)$ is a DAG and $G = (V, E_G)$ is a graph such that*

(a) G does not induce $a \to b - c$ and

(b) $E_D \subseteq E_G \subseteq E_E$ where $(V, E_E) = E(D)$,

then Algorithm 3 applied to G returns $E(D)$.

The following lemma is used to prove Proposition 3.7.

Lemma 3.8. *Let $L = (V, E_V)$ be a graph whose arrows are strongly protected and let $S = (V, E_S)$ be a graph such that L and S have the same skeleton and, for each $a \to b \in L$,*

(a) $a \to b \in S$ and

(b) if there is no vertex $c \in V$ such that $L_{\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c), then there exist vertices $c, d \in V$ such that $L_{\{a,b,c,d\}}$ is given in Figure 3.3(d) and $S_{\{a,b,c,d\}}$ is one of the graphs given in Figure 3.8.

If S' is the graph obtained from S by converting every non-strongly protected arrow into a line, then L and S' have the same skeleton and, for each $a \to b \in L$, the above propositions (a) and (b) also hold when S' replaces S.

Proof. Converting arrows into line preserves the skeleton. Hence, S, S' and L have the same skeleton. Given $a \to b \in L$, let us show that the propositions (a) and (b) hold when S' replaces S. By hypothesis, one of the two following possibilities holds.

1. There exists c such that $L_{\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c). Trivially, (b) holds. Because L and S have the same skeleton and every arrow in L is also in S, we have $S_{\{a,b,c\}} = L_{\{a,b,c\}}$. Hence, $a \to b$ is strongly protected in S, and thus $a \to b \in S'$.

2. Suppose that there is no vertex c such that $L_{\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c). By hypothesis, there exist vertices c, d such that $L_{\{a,b,c,d\}}$ is given in Figure 3.3(d) and $S_{\{a,b,c,d\}}$ is one of the graphs given in Figure 3.8. Let us now discuss the possible subgraphs $S_{\{a,b,c,d\}}$ and show that each case leads to the conclusion that $a \to b \in S'$ and $S'_{\{a,b,c,d\}}$ is one of the graphs given in Figure 3.8. First, note that $d \to b \in S$ and $c \to b \in S$ are strongly protected by $S_{\{b,c,d\}}$ in all the case. Hence, we have $d \to b \in S'$ and $c \to b \in S'$.

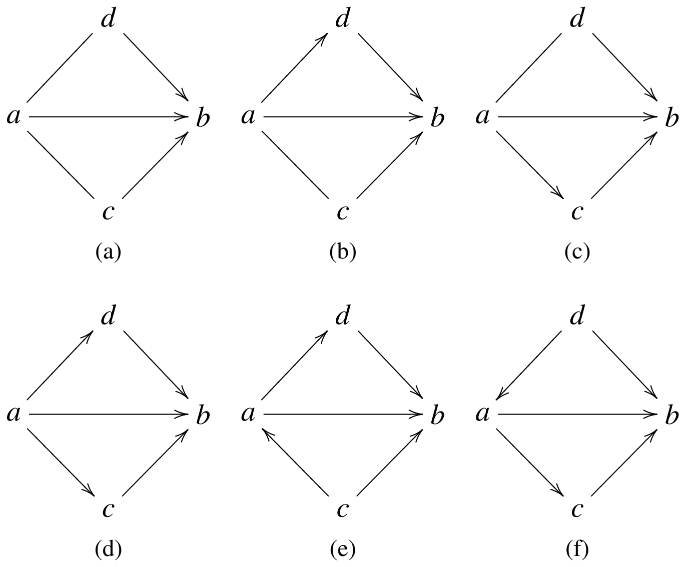


Figure 3.8: Possible subgraphs $S_{\{a,b,c,d\}}$.

(a) Suppose $S_{\{a,b,c,d\}}$ is given in Figure 3.8(a). Then, $a \to b \in S$ is strongly protected by $S_{\{a,b,c,d\}}$. Hence, $a \to b \in S'$ and $S'_{\{a,b,c,d\}}$ is given in Figure 3.8(a).

(b) Suppose that $S_{\{a,b,c,d\}}$ is given in Figure 3.8(b). Then, $a \to b \in S$ is strongly protected by $S_{\{a,b,d\}}$. Hence, $a \to b \in S'$ and $S'_{\{a,b,c,d\}}$ is given in Figure 3.8(a) or 3.8(b).

(c) The case where $S_{\{a,b,c,d\}}$ is given in Figure 3.8(c) is similar to the previous case.

(d) Suppose that $S_{\{a,b,c,d\}}$ is given in Figure 3.8(d). Then, $a \to b \in S$ is strongly protected by $S_{\{a,b,d\}}$. Hence, $a \to b \in S'$ and $S'_{\{a,b,c,d\}}$ is given in Figure 3.8(a), 3.8(b), or 3.8(c).

(e) Suppose that $S_{\{a,b,c,d\}}$ is given in Figure 3.8(e). Then, $a \to b \in S$ is strongly protected by $S_{\{a,b,d\}}$ and $a \to d \in S$ is strongly protected by $S_{\{a,c,d\}}$. Hence, $a \to b \in S'$ and $S'_{\{a,b,c,d\}}$ is given in Figure 3.8(b) or 3.8(e).

(f) The case where $S_{\{a,b,c,d\}}$ is given in Figure 3.8(f) is similar to the previous case. □

Proof (Proposition 3.7.). Let $G_0, \dots, G_k$ with $G_0 = G$ be the sequence of graphs produced by Algorithm 3. Let us show that G_k and $E(D)$ have the same skeleton and arrows.

1. Let us show that G_k and $E(D)$ have the same skeleton. First, the graphs $G_0, \dots, G_k$ have the same skeleton because each step of Algorithm 3 preserves the skeleton. Second, $G_0 = G$ and $E(D)$ have the same skeleton because D and $E(D)$ have the same skeleton and $E_D \subseteq E_G \subseteq E_E$.

2. Let us show that every arrow in $E(D)$ is also in G_k. If the hypotheses of Lemma 3.8 with $L = E(D)$ and $S = G = G_0$ are satisfied, then they are also satisfied with $L = E(D)$ and $S = G_i$, $i = 0, \dots, k$. In particular, every arrow in $L = E(D)$ is in $S = G_k$.

   As shown before, $E(D)$ and G have the same skeleton. Moreover, every arrow in $E(D)$ is also in G because $E_G \subseteq E_E$. By Theorem 3.1, every arrow in $E(D)$ is strongly protected. Consider $a \to b \in E(D)$ and suppose that there is no c such that $E(D)_{\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c). It is easy to see that G induces a subgraph given in Figure 3.8 because G and $E(D)$ have the same skeleton, every arrow in $E(D)$ is also in G, and G does not induce $d \to a - c$ or $c \to a - d$.

3. Let us show that every arrow in G_k is also in $E(D)$. Let $D_0, \dots, D_l$ with $D_0 = D$ and $D_l = E(D)$ be the sequence of graphs produced by Algorithm 3. If the hypotheses of Lemma 3.8 with $L = G_k$ and $S = D$ are satisfied, then they are also satisfied with $L = G_k$ and $S = D_i$, $i = 0, \dots, l$. In particular, every arrow in $L = G_k$ is in $S = D_l = E(D)$.

   As shown before, G_k and D have the same skeleton. We have $E_D \subseteq E_G \subseteq E_{G_K}$, and thus every arrow in G_k is in D. By definition of Algorithm 3, all the arrows in G_k are strongly protected. Consider $a \to b \in G_k$ and suppose that there is no c such that $G_{k\{a,b,c\}}$ is given in Figure 3.3(a), 3.3(b), or 3.3(c). It is easy to see that D induces a subgraph of Figure 3.8(d), 3.8(e), or 3.8(f) because G_k and D have the same skeleton, every arrow in G_k is also in D, and D is directed. □

3.3 Computation of the Inclusion Boundary IB(E)

Given an essential graph $E \in \mathcal{E}(V)$, this section discusses the computation of the inclusion boundary IB(E). The case $|V| = 1$ being trivial, we suppose that $|V| \geq 2$.

The inclusion boundary may be partitioned by the skeleton of its elements, i.e.

$$\left\{S^{-1}(t) \cap \mathrm{IB}(E) \middle| t \in S(\mathrm{IB}(E))\right\} \tag{3.2}$$

is a partition of IB(E). Subsequently, each element $S^{-1}(t) \cap \mathrm{IB}(E)$ of the partition by skeleton may be partitioned by the v-structures of its elements, i.e.

$$\left\{v^{-1}(s) \cap S^{-1}(t) \cap \mathrm{IB}(E) \middle| s \in v(S^{-1}(t) \cap \mathrm{IB}(E))\right\} \tag{3.3}$$

is a partition of $S^{-1}(t) \cap \mathrm{IB}(E)$. By Theorem 1.10, note that each set $v^{-1}(s) \cap S^{-1}(t) \cap \mathrm{IB}(E)$ is a singleton. Then, the computation of $\mathrm{IB}(E)$ proceed as follows:

- The partition by skeleton given by (3.2) is enumerated.
- For each set $S^{-1}(t) \cap \mathrm{IB}(E)$, the set $v(S^{-1}(t) \cap \mathrm{IB}(E))$ is enumerated.
- For each $s \in v(S^{-1}(t) \cap \mathrm{IB}(E))$, the set $v^{-1}(s) \cap S^{-1}(t) \cap \mathrm{IB}(E)$ is computed.

Section 3.3.1 describes the partition by skeleton graphically to simplify its manipulation and enumeration. Section 3.3.2 and Section 3.3.3 discuss the enumeration of $v(S^{-1}(t) \cap \mathrm{IB}(E))$ and the computation of $v^{-1}(s) \cap S^{-1}(t) \cap \mathrm{IB}(E)$ for each $s \in v(S^{-1}(t) \cap \mathrm{IB}(E))$.

3.3.1 Graphical Characterization of the Partition by Skeleton

The following theorem characterizes the inclusion boundary graphically.

Theorem 3.9. *If $E \in \mathcal{E}(V)$, then $G \in \mathrm{IB}(E)$ if, and only if, $G \in \mathcal{E}(V)$ and there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by adding or removing an arrow.*

Proof. By Theorem 1.11 and Theorem 1.13, the assertions $G, H \in \mathcal{B}(V)$ and $G <_I H$ imply that there exists a sequence of $x \geq 0$ covered arrow reversals and $y \geq 1$ legal arrow additions turning G into H.

1. Suppose $G \in \mathcal{E}(V)$ and there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by adding or removing one arrow. Let us show that $G \in \mathrm{IB}(E)$.

   (a) Suppose L is obtained from K by adding one arrow. This implies that $K <_I L$, and thus $E <_I G$. Let us show by contradiction there is no $H \in \mathcal{E}(V)$ such that $E <_I H <_I G$. Suppose that such an H exists. If $M \in [H]$, we have $K <_I M <_I L$. As noted before, there thus exists a sequence of x covered arrow reversals and $y \geq 1$ legal arrow additions turning K into M, and there exists a sequence of x' covered arrow reversals and $y' \geq 1$ legal arrow additions turning M into L. This contradicts the assertion that L is obtained from K by adding one single arrow. Hence, $G \in \mathrm{IB}(E)$.

   (b) Suppose L is obtained from K by removing one arrow. This implies that $L <_I K$ and $G <_I E$. Let us show by contradiction there is no $H \in \mathcal{E}(V)$ such that $G <_I H <_I E$. Suppose that such an H exists. If $M \in [H]$, we have $L <_I M <_I K$. There thus exists a sequence of x covered arrow reversals and $y \geq 1$ legal arrow additions turning L into M, and there exists a sequence of x' covered arrow reversals and $y' \geq 1$ legal arrow additions turning M into K. This contradicts the assertion that L is obtained from K by removing one single arrow. Hence, $G \in \mathrm{IB}(E)$.

2. Suppose that $G \in \mathrm{IB}(E)$. Let us discuss the case where $E <_I G$ and the case where $G <_I E$ separately.

   (a) Suppose that $E <_I G$ and there is no $H \in \mathcal{E}(V)$ such that $E <_I H <_I G$. Consider $K \in [E]$ and $L \in [G]$. We have $K <_I L$ and there thus exists a sequence s of x covered arrow reversals and $y \geq 1$ legal arrow additions turning K into L. Suppose that $y = 1$ and consider the sequence of DAGs obtained by applying the sequence s to K. Because $y = 1$, this sequence of DAGs can be written $K = K_0, \ldots, K_a, L_0, \ldots, L_b = L$ where

      - L_0 is obtained from K_a by adding one arrow;
      - for $i \in \{0, \ldots, a-1\}$, K_{i+1} is obtained from K_i by reversing a covered arrow
      - for $i \in \{0, \ldots, b-1\}$, L_{i+1} is obtained from L_i by reversing a covered arrow.

      By Theorem 1.11, $K =_I K_a$ and $L_0 =_I L$. Hence, we have $K_a \in [E]$, $L_0 \in [G]$ such that L_0 is obtained from K_a by adding one arrow. Let us now prove that $y \leq 1$ by contradiction. Together with $y \geq 1$, this will prove that $y = 1$. Suppose that $y \geq 2$ and consider the sequence of DAGs obtained by applying s to K. Let M be the DAG of this sequence obtained after adding the first arrow. We have $K <_I M <_I L$. Hence, there exists $H = E(M) \in \mathcal{E}(V)$ such that $E <_I H <_I G$. This contradicts the assertion that $G \in \mathrm{IB}(E)$, and we have $y \leq 1$.

   (b) The proof of the case where $G <_I E$ and there is no $H \in \mathcal{E}(V)$ such that $G <_I H <_I E$ is similar. □

Remark 55. The difference in score between neighboring essential graphs can be computed incrementally if the scoring criterion score is decomposable. By Theorem 3.9, $G \in \mathrm{IB}(E)$ implies that there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by adding or removing an arrow. Hence, we have

$$\begin{aligned} \mathrm{score}(G) - \mathrm{score}(E) &= \mathrm{score}(L) - \mathrm{score}(K) && (3.4)\\ &= f(u, pa_L(u)) - f(u, pa_K(u)) && (3.5) \end{aligned}$$

where u is the destination of the arrow added or removed.

The partition of IB(E) by skeleton can be described graphically.

Definition 119. If $E \in \mathcal{E}(V)$ and $\{a, b\} \subseteq V$, let $\mathrm{IB}_{\{a,b\}}$ be defined by $G \in \mathrm{IB}_{\{a,b\}}$ if, and only if, $G \in \mathcal{E}(V)$ and there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by adding or removing an arrow between a and b.

Remark 56. Like $\mathrm{IB}_{\{a,b\}}$, the objects defined in this chapter depend on E. However, the essential graph E is fixed. To simplify notations, the dependence on E will not appear explicitely.

Proposition 3.10. *If* $E \in \mathcal{E}(V)$*, then* $\{\mathrm{IB}_{\{a,b\}} \big| \{a,b\} \subseteq V\}$ *and* $\{S^{-1}(s) \cap \mathrm{IB}(E) \big| s \in S(\mathrm{IB}(E))\}$ *are equivalent partitions of* $\mathrm{IB}(E)$*.*

Proof. If $G \in \mathrm{IB}_{\{a,b\}}$ and $a \cdots b \in E$, then $S(G)$ is the undirected graph obtained from $S(E)$ by removing $a - b$. If $G \in \mathrm{IB}_{\{a,b\}}$ and $a \cdots b \notin E$, then $S(G)$ is the undirected graph obtained from $S(E)$ by adding $a - b$. For each $\{a,b\} \subseteq V$, S is thus constant over $\mathrm{IB}_{\{a,b\}}$. Moreover, if $\{a,b\} \subseteq V$ and $\{c,d\} \subseteq V$ are distinct, $G_1 \in \mathrm{IB}_{\{a,b\}}$ and $G_2 \in \mathrm{IB}_{\{c,d\}}$, then $S(G_1) \neq S(G_2)$. To conclude the proof, let us show that $\{\mathrm{IB}_{\{a,b\}} \big| \{a,b\} \subseteq V\}$ is a partition of $\mathrm{IB}(E)$.

1. By Theorem 3.9, $\mathrm{IB}(E)$ is the union of the sets in $\{\mathrm{IB}_{\{a,b\}} | \{a,b\} \subseteq V\}$.

2. If $\{a,b\} \subseteq V$ and $\{c,d\} \subseteq V$ are distinct, then $\mathrm{IB}_{\{a,b\}}$ and $\mathrm{IB}_{\{c,d\}}$ do not intersect because elements of $\mathrm{IB}_{\{a,b\}}$ and elements of $\mathrm{IB}_{\{c,d\}}$ have distinct skeleton.

3. Let us show that each $\mathrm{IB}_{\{a,b\}}$ is non-empty. There exists a DAG $K \in [E]$.

   (a) Suppose that $a \cdots b \in E$. We have $a \cdots b \in K$, and the graph L obtained from K by removing the arrow between a and b is a DAG. Hence, $E(L) \in \mathrm{IB}_{\{a,b\}}$.

   (b) Suppose $a \cdots b \notin E$. We have $a \cdots b \notin K$, and we let L_1 (resp. L_2) be the graph obtained from K by adding $a \rightarrow b$ (resp. $b \rightarrow a$). By acyclity, K does not have simultaneously a path from a to b and a path from b to a. If K does not have a path from a to b, then L_2 is a acyclic and $E(L_2) \in \mathrm{IB}_{\{a,b\}}$. If K does not have a path from b to a, then L_1 is a acyclic and $E(L_1) \in \mathrm{IB}_{\{a,b\}}$. □

3.3.2 Computation of $\mathrm{IB}_{\{a,b\}}$ when $a \cdots b \in E$

This section describes the image $\nu(\mathrm{IB}_{\{a,b\}})$ when $a \cdots b \in E$ so that it can be enumerated, and introduces an algorithm to compute $\nu^{-1}(s) \cap \mathrm{IB}_{\{a,b\}}$ for $s \in \nu(\mathrm{IB}_{\{a,b\}})$. If the scoring criterion score is decomposable, the score of each element in $\mathrm{IB}_{\{a,b\}}$ is also computed incrementally.

Preliminary Notions

Let us introduce notions necessary to describe $\nu(\mathrm{IB}_{\{a,b\}})$.

Definition 120. If $G = (V, E)$ is a graph, the set $ch_G(a)$ of *children* of $a \in V$ is

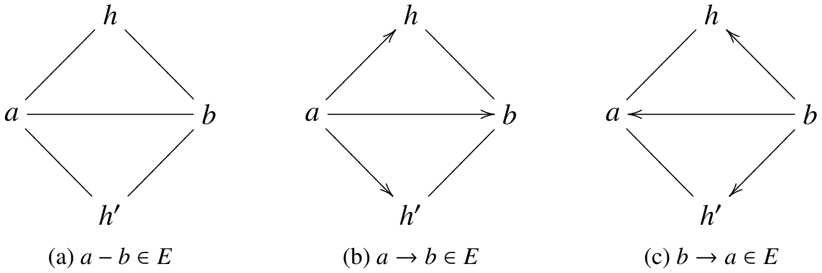
$$ch_G(a) = \{b \in V \big| a \rightarrow b \in G\}. \tag{3.6}$$

Definition 121. If $E \in \mathcal{E}(V)$ and $a \cdots b \in E$, let $H^-_{\{a,b\}}$ be defined by

$$H^-_{\{a,b\}} = (ne_E(a) \cap ne_E(b)) \cup (ch_E(a) \cap ne_E(b)) \cup (ch_E(b) \cap ne_E(a)). \tag{3.7}$$

Remark 57. Since E is a chain graph and thus has no directed cycle, at most one of the sets $ne_E(a) \cap ne_E(b)$, $ch_E(a) \cap ne_E(b)$ and $ch_E(b) \cap ne_E(a)$ is non-empty:

- if $a - b \in E$, then $H^-_{\{a,b\}} = ne_E(a) \cap ne_E(b)$ (see Figure 3.9(a))
- if $a \to b \in E$, then $H^-_{\{a,b\}} = ch_E(a) \cap ne_E(b)$ (see Figure 3.9(b))
- if $b \to a \in E$, then $H^-_{\{a,b\}} = ch_E(b) \cap ne_E(a)$ (see Figure 3.9(c)).



(a) $a - b \in E$ (b) $a \to b \in E$ (c) $b \to a \in E$

Figure 3.9: $E_{\{a,b,h,h'\}}$ for $h, h' \in H^-_{\{a,b\}}$

Example 48. If E is given in Figure 3.1, then

$$H^-_{\{a,c\}} = \{b, c, d\} \cap \{a, b, d\} = \{b, d\}, \tag{3.8}$$

$$H^-_{\{e,g\}} = \{g, h\} \cap \{h\} = \{h\}, \tag{3.9}$$

$$H^-_{\{i,j\}} = \{j\} \cap \emptyset = \emptyset. \tag{3.10}$$

Remark 58. The chain component of E that includes

- $\{a, b\}$ if $a - b \in E$,
- $\{b\}$ if $a \to b \in E$, or
- $\{a\}$ if $b \to a \in E$

also includes $H^-_{\{a,b\}}$. This observation will be used in Algorithm 4.

Definition 122. If $E \in \mathcal{E}(V)$ and $a \cdots b \in E$, the set $S^-_{\{a,b\}}$ is defined by

$$S^-_{\{a,b\}} = \Big\{c \subseteq H^-_{\{a,b\}} \Big| c \text{ is empty or complete in } E\Big\}. \tag{3.11}$$

Example 49. If E is given in Figure 3.1, then

$$S^-_{\{a,c\}} = \{\emptyset, \{b\}, \{d\}\}, \tag{3.12}$$

$$S^-_{\{e,g\}} = \{\emptyset, \{h\}\}, \tag{3.13}$$

$$S^-_{\{i,j\}} = \{\emptyset\}. \tag{3.14}$$

Remark 59. If $c \in S^-_{\{a,b\}}$, the set

- $c \cup \{a, b\}$ if $a - b \in E$,
- $c \cup \{b\}$ if $a \to b \in E$, or
- $c \cup \{a\}$ if $b \to a \in E$

is complete in E. This observation will be used in Algorithm 4.

Definition 123. If $E \in \mathcal{E}(V)$ and $a \cdots b \in E$, the function $f^-_{\{a,b\}}$ is defined on $S^-_{\{a,b\}}$ by

$$f^-_{\{a,b\}}(c) = \Big\{(h, \{a, b\})\Big| h \in (H^-_{\{a,b\}} \setminus c) \cup (ch_E(a) \cap ch_E(b))\Big\} \bigcup \Big(v(E) \setminus (\{(b, \{a, v\})|v \in V\} \cup \{(a, \{b, v\})|v \in V\})\Big). \quad (3.15)$$

Remark 60. The function $f^-_{\{a,b\}}$ is injective. Hence, $\left|S^-_{\{a,b\}}\right| = \left|f^-_{\{a,b\}}(S^-_{\{a,b\}})\right|$.

The Image $v(\mathrm{IB}_{\{a,b\}})$

Let us prove that $f^-_{\{a,b\}}(S^-_{\{a,b\}}) = v(\mathrm{IB}_{\{a,b\}})$. First, Proposition 3.11 states that $v(\mathrm{IB}_{\{a,b\}}) \subseteq f^-_{\{a,b\}}(S^-_{\{a,b\}})$. Then, Corollary 3.13 states that $f^-_{\{a,b\}}(S^-_{\{a,b\}}) \subseteq v(\mathrm{IB}_{\{a,b\}})$.

Proposition 3.11. *If $E \in \mathcal{E}(V)$ and $a \cdots b \in E$, then $v(\mathrm{IB}_{\{a,b\}}) \subseteq f^-_{\{a,b\}}(S^-_{\{a,b\}})$.*

Proof. Given $s \in v(\mathrm{IB}_{\{a,b\}})$, we show there exists $c \in S^-_{\{a,b\}}$ such that $s = f^-_{\{a,b\}}(c)$.

1. Let us define a candidate c. First, $s \in v(\mathrm{IB}_{\{a,b\}})$ implies there exists $G \in \mathrm{IB}_{\{a,b\}}$ such that $s = v(G)$. By Definition 119, there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by removing the arrow between a and b. If $x = (ch_K(a) \setminus ch_E(a))$ and $y = (ch_K(b) \setminus ch_E(b))$, let

$$c = H^-_{\{a,b\}} \setminus \Big((x \cap y) \cup (ch_E(a) \cap y) \cup (ch_E(b) \cap x)\Big). \quad (3.16)$$

2. Let us show that $c \in S^-_{\{a,b\}}$. We have $c \subseteq H^-_{\{a,b\}}$. Without loss of generality, we suppose that $a \to b \in K$ (the case $b \to a \in K$ is similar). First, we show that $c \subseteq pa_k(b)$. It is easy to see that c is the union of $(ne_E(a) \cap ne_E(b)) \setminus (x \cap y)$, $(ch_E(a) \cap ne_E(b)) \setminus (ch_E(a) \cap y)$ and $(ch_E(b) \cap ne_E(a)) \setminus (ch_E(b) \cap x)$.

   (a) Suppose that $a - b \in E$. By Remark 57, $c = (ne_E(a) \cap ne_E(b)) \setminus (x \cap y)$. One can see that $c = ne_E(a) \cap ne_E(b) \cap \Big((ch_K(a) \cap pa_K(b)) \cup (ch_K(b) \cap pa_K(a))\Big)$. By acyclicity of K, $a \to b \in K$ implies $ch_K(b) \cap pa_K(a) = \emptyset$. Hence, $c = ne_E(a) \cap ne_E(b) \cap ch_K(a) \cap pa_K(b) \subseteq pa_K(b)$.

   (b) Suppose that $a \to b \in E$. By Remark 57, $c = (ch_E(a) \cap ne_E(b)) \setminus (ch_E(a) \cap y)$. One can see that $c = ch_E(a) \cap ne_E(b) \cap ch_K(a) \cap pa_K(b)$. Hence, $c \subseteq pa_K(b)$.

Let us show by contradiction that $c \subseteq pa_k(b)$ implies that E_c is complete. There exists a chain component of E such that $c \subseteq H^-_{\{a,b\}} \subseteq \delta$. By Theorem 3.3, $K \in [E]$ implies that K_δ and thus also K_c have no v-structure. If $h_i, h_j \in c$ are distinct and $h_i \cdots h_j \notin K_c$, then $(b, \{h_i, h_j\}) \in v(K_c)$. Hence, for distinct $h_i, h_j \in c$, we have $h_i \cdots h_j \in K_c$ and thus $h_i - h_j \in E_c$.

3. Let us show that $s = v(G) = f^-_{\{a,b\}}(c)$. We have

$$v(G) = (v(G) \setminus v(E)) \cup (v(G) \cap v(E)) = (v(L) \setminus v(K)) \cup (v(L) \cap v(K)). \quad (3.17)$$

(a) Let us compute $v(L) \setminus v(K)$. It is easy to see that

$$v(L) \setminus v(K) = \Big\{(h, \{a, b\})\Big| h \in ch_K(a) \cap ch_K(b)\Big\}. \quad (3.18)$$

Moreover, $ch_K(a) = ch_E(a) \cup x$ and $ch_K(b) = ch_E(b) \cup y$. Hence, $ch_K(a) \cap ch_K(b)$ is equal to

$$(ch_E(a) \cap ch_E(b)) \cup (x \cap y) \cup (ch_E(a) \cap y) \cup (ch_E(b) \cap x). \quad (3.19)$$

It is easy to see that $c \subseteq H^-_{\{a,b\}}$. Hence, $H^-_{\{a,b\}} \setminus c = (x \cap y) \cup (ch_E(a) \cap y) \cup (ch_E(b) \cap x)$ and $v(L) \setminus v(K)$ is equal to

$$\Big\{(h, \{a, b\})\Big| h \in (H^-_{\{a,b\}} \setminus c) \cup (ch_E(a) \cap ch_E(b))\Big\}. \quad (3.20)$$

(b) It is easy to see that $v(K) \cap v(L)$ is equal to

$$v(K) \setminus \Big(\{(b, \{a, v\})\big| v \in V\} \cup \{(a, \{b, v\})\big| v \in V\}\Big), \quad (3.21)$$

and thus to

$$v(E) \setminus \Big(\{(b, \{a, v\})\big| v \in V\} \cup \{(a, \{b, v\})\big| v \in V\}\Big). \quad (3.22)$$

□

Let us prove constructively that $f^-_{\{a,b\}}(S^-_{\{a,b\}}) \subseteq v(\mathrm{IB}_{\{a,b\}})$ with Algorithm 4 and Proposition 3.12. Algorithm 4 takes for input $c \in S^-_{\{a,b\}}$, and returns a graph D.

Algorithm 4

1. Set p to a permutation of c.
2. Set δ to the chain component of E including
   - $\{a, b\}$ if $a - b \in E$,
   - $\{b\}$ if $a \to b \in E$, or
   - $\{a\}$ if $b \to a \in E$.
3. Set o to a perfect ordering of E_δ such that o starts with
   - apb if $a - b \in E$,

- pb if $a \to b \in E$, or
- pa if $b \to a \in E$.

4. Set D to the graph obtained from E by

   (a) Orienting the lines in E_δ according to o and

   (b) Orienting the lines in the subgraphs induced by the other chain components of E according to perfect orderings.

5. Remove the arrow between a and b from D.

6. Return D. □

Remark 61. By Remark 58, Remark 59 and Lemma 3.5, the perfect ordering o defined at Step 3 exists.

Remark 62. Algorithm 4 is non-deterministic because of the freedom in the choice of perfect orderings and permutation of c.

Example 50. Let E be the essential graph given in Figure 3.1. For each element $c' \in S^-_{\{a,b\}} = \{\emptyset, \{b\}, \{d\}\}$, let us describe possible executions of Algorithm 4. The chain component δ including $\{a, b\}$ is $\{a, b, c, d, f\}$, and we pick the perfect ordering

- $o = acbdf$ when $c' = \emptyset$,
- $o = abcdf$ when $c' = \{b\}$, and
- $o = adcbf$ when $c' = \{d\}$.

The only chain component $\tau \neq \delta$ of E such that E_τ contains lines is $\{g, h\}$, and we pick the perfect ordering hg. With these choices of perfect orderings, Algorithm 4 returns the graphs of Figure 3.10.

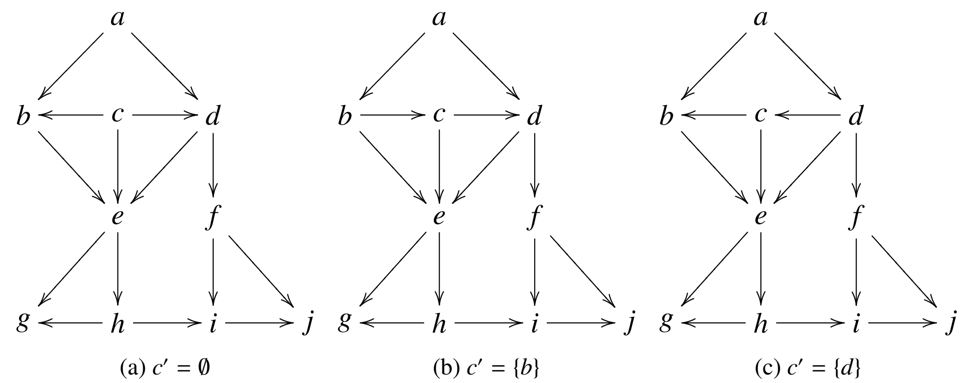
Proposition 3.12. *If D is the result of Algorithm 4 applied to $c \in S^-_{\{a,b\}}$, then $E(D) \in \mathrm{IB}_{\{a,b\}}$ and $v(E(D)) = f^-_{\{a,b\}}(c)$.*

Proof.

1. Let us show that $E(D) \in \mathrm{IB}_{\{a,b\}}$. By Theorem 3.3, the graph K obtained after Step 4 satisfies $K \in [E]$. Hence, D is a DAG, and $E(D) \in \mathrm{IB}_{\{a,b\}}$ by Definition 119.

2. Let us show that $v(E(D)) = f^-_{\{a,b\}}(c)$. We have

$$v(E(D)) = (v(E(D)) \setminus v(E)) \cup (v(E(D)) \cap v(E)) = (v(D) \setminus v(K)) \cup (v(D) \cap v(K)). \tag{3.23}$$



(a) $c' = \emptyset$ (b) $c' = \{b\}$ (c) $c' = \{d\}$

Figure 3.10: Graphs returned by Algorithm 4

(a) Let us compute $v(D) \setminus v(K)$. We have

$$v(D) \setminus v(K) = \Big\{(h, \{a, b\})\Big| h \in ch_K(a) \cap ch_K(b)\Big\}. \tag{3.24}$$

It is easy to see that $ch_K(a) \cap ch_K(b) = c \cup (ch_E(a) \cap ch_E(b))$. Hence, $v(D) \setminus v(K)$ is equal to

$$\Big\{(h, \{a, b\})\Big| h \in c \cup (ch_E(a) \cap ch_E(b))\Big\}. \tag{3.25}$$

(b) It is easy to see that $v(K) \cap v(L)$ is equal to

$$v(K) \setminus \Big(\{(b, \{a, v\})\big| v \in V\} \cup \{(a, \{b, v\})\big| v \in V\}\Big), \tag{3.26}$$

and thus to

$$v(E) \setminus \Big(\{(b, \{a, v\})\big| v \in V\} \cup \{(a, \{b, v\})\big| v \in V\}\Big). \tag{3.27}$$

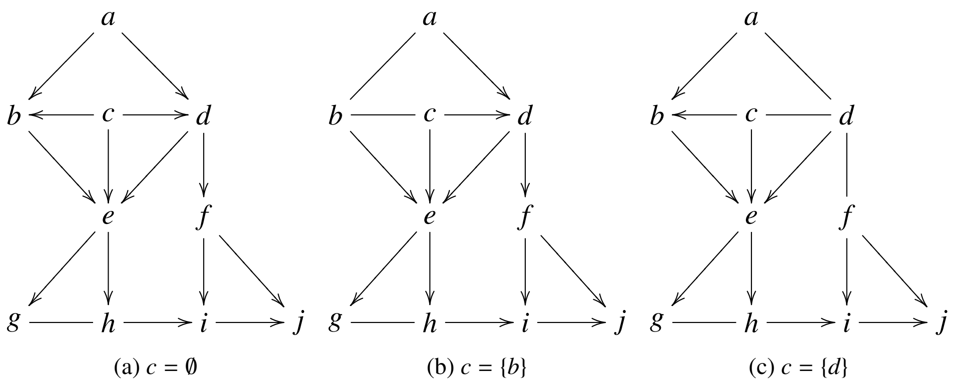
□

Corollary 3.13. *If $E \in \mathcal{E}(V)$ and $a \cdots b \in E$, then $f^-_{\{a,b\}}(S^-_{\{a,b\}}) \subseteq v(\mathrm{IB}_{\{a,b\}})$.*

By Proposition 3.12, to enumerate $\mathrm{IB}_{\{a,b\}}$, it is sufficient to enumerate $S^-_{\{a,b\}}$ and apply successively Algorithm 4 and Algorithm 3 to each $c \in S^-_{\{a,b\}}$. In particular, it is not necessary to compute and enumerate $v(\mathrm{IB}_{\{a,b\}})$.

Example 51. Let E be the essential graph given in Figure 3.1. By Example 50, the graphs given in Figure 3.10 are obtained by applying Algorithm 4 to the elements of $S^-_{\{a,c\}}$. Applying Algorithm 3 to these graphs, we obtain $\mathrm{IB}_{\{a,c\}}$ (see Figure 3.11).

As the following example illustrates, $\mathrm{IB}_{\{a,b\}}$ may contain a large number of elements.



(a) $c = \emptyset$ (b) $c = \{b\}$ (c) $c = \{d\}$

Figure 3.11: $\mathrm{IB}_{\{a,c\}}$

Example 52. Suppose that the essential graph $E \in \mathcal{E}(V)$ is complete and undirected. For each $\{a, b\} \subseteq V$, the set $H^-_{\{a,b\}} = V \setminus \{a, b\}$ is complete. Hence, every non-empty subset of $H^-_{\{a,b\}}$ is also complete, and

$$\left|\mathrm{IB}_{\{a,b\}}\right| = \left|\nu(\mathrm{IB}_{\{a,b\}})\right| = \left|f^-_{\{a,b\}}(S^-_{\{a,b\}})\right| = \left|S^-_{\{a,b\}}\right| = 2^{|V|-2}. \tag{3.28}$$

Incremental Computation of the Neighbors

For each $c \in S^-_{\{a,b\}}$, the neighbor $G = \nu^{-1}(f^-_{\{a,b\}}(c))$ may be obtained with Algorithm 4 and Algorithm 3 as described above. However, Proposition 3.7 allows us to exploit the non-determinism of Algorithm 4 to speed up the computation of G.

The following algorithm takes for input $c \in S^-_{\{a,b\}}$, and returns a graph G.

Algorithm 5

```
1. Set $p$ to a permutation of $c$.

2. Set $\delta$ to the chain component of $E$ including

   • $\{a, b\}$ if $a - b \in E$,
   • $\{b\}$ if $a \to b \in E$, or
   • $\{a\}$ if $b \to a \in E$.

3. Set $o$ to a perfect ordering of $E_\delta$ such that $o$ starts with

   • $apb$ if $a - b \in E$,
   • $pb$ if $a \to b \in E$, or
   • $pa$ if $b \to a \in E$.

4. Set $G := E$.

5. Orient according to $o$ each $u - v \in G$ such that $\{u, v\} \subseteq \delta$ and $\{u, v\} \not\subseteq c$.
```

6. Remove the edge between a and b from G.
7. Apply Algorithm 3 to G and return the result. □

Proposition 3.14. *If G is the result of Algorithm 5 applied to $c \in S^-_{\{a,b\}}$, then $G \in \mathrm{IB}_{\{a,b\}}$ and $v(G) = f^-_{\{a,b\}}(c)$.*

Proof. Let $G = (V, E_G)$ be the graph obtained before Step 7, let $D = (V, E_V)$ be the graph obtained by applying Algorithm 4 to c and using the same perfect ordering o of E_δ, and let $(V, E_E) = E(D)$. We apply Proposition 3.7 to prove that $E(D)$ is the result of Algorithm 3 applied to G. Let us show that the hypotheses are satisfied.

1. Let us show that $E_D \subseteq E_G \subseteq E_E$.
   (a) It is easy to see that D, G, and $E(D)$ share the same skeleton.
   (b) Because Algorithm 5 and Algorithm 4 use the same perfect ordering o, we have $E_D \subseteq E_G$.
   (c) Let us show that $E_G \subseteq E_E$. Let D' be the result of Algorithm 4 applied to c. By Lemma 3.5, we can choose the perfect orderings used to obtain D' so that $u \to v \in D'$ for each $\{u, v\}$ in a chain component of E distinct from δ, and for each $\{u, v\} \subseteq c$. Hence, $u - v \in E(D)$ for each such vertices $\{u, v\}$. This implies that $E_G \subseteq E_E$.
2. Let us show by contradiction that G does not induce $x \to y - z$. Suppose that G induces $x \to y - z$ for some vertices x, y, and z. In that case, $y - z \in E$ and either $x \to y \in E$ or $x - y \in E$.
   (a) Suppose $x \to y \in E$. Then, E induces $x \to y - z$, which contradicts Theorem 3.1.
   (b) Suppose that $x - y \in E$. Then, $x, y, z \in \delta$ with $x \notin c$ and $y, z \in c$. Hence, y precedes x in o, and thus $x \to y \notin G$. □

Incremental Computation of the Score

Proposition 3.15. *Suppose the scoring criterion* score *is decomposable and suppose $a \to b \in E$ or $a - b \in E$. If $G = v^{-1}(f^-_{\{a,b\}}(c)) \in \mathrm{IB}_{\{a,b\}}$, then*

$$\mathrm{score}(G) - \mathrm{score}(E) = f(b, (pa_E(b) \setminus \{a\}) \cup c) - f(b, pa_E(b) \cup \{a\} \cup c). \quad (3.29)$$

Proof. Let L be the result of Algorithm 4 applied to c, and let K be the DAG obtained from L by adding $a \to b$. We have $\mathrm{score}(G) - \mathrm{score}(E) = \mathrm{score}(L) - \mathrm{score}(K) = f(b, pa_L(b)) - f(b, pa_K(b))$. It is easy to see that $pa_L(b) = (pa_E(b) \setminus \{a\}) \cup c$ and $pa_K(b) = pa_E(b) \cup \{a\} \cup c$. □

To compute the increment in score, the knowledge of c is sufficient. In particular, it is not necessary to compute G. This is advantageous for a greedy search using this neighborhood since we only need to compute the neighbor that most increases the score.

3.3.3 Computation of $\mathrm{IB}_{\{a,b\}}$ when $a \cdots b \notin E$

This section describes the image $v(\mathrm{IB}_{\{a,b\}})$ when $a \cdots b \notin E$ so that it can be enumerated, and introduces an algorithm to compute $v^{-1}(s) \cap \mathrm{IB}_{\{a,b\}}$ for $s \in v(\mathrm{IB}_{\{a,b\}})$. If the scoring criterion score is decomposable, the score of each element in $\mathrm{IB}_{\{a,b\}}$ is also computed incrementally.

Preliminary Notions

Let us introduce notions necessary to describe $v(\mathrm{IB}_{\{a,b\}})$.

Definition 124. If $G = (V, E)$ is a graph, the set $ad_G(a)$ of vertices *adjacent* to $a \in V$ is

$$ad_G(a) = \{b \in V \big| a \cdots b \in G\}. \tag{3.30}$$

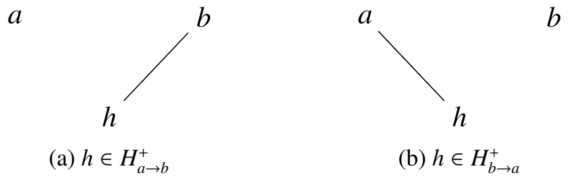
Definition 125. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, let

$$H^+_{a \to b} = ne_E(b) \setminus ad_E(a). \tag{3.31}$$

Definition 126. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, let

$$H^+_{b \to a} = ne_E(a) \setminus ad_E(b). \tag{3.32}$$

Remark 63. It is easy to see that $h \in H^+_{a \to b}$ if, and only if, $E_{\{a,b,h\}}$ is given in Figure 3.12(a). Similarly, $h \in H^+_{b \to a}$ if, and only if, $E_{\{a,b,h\}}$ is given in Figure 3.12(b).



(a) $h \in H^+_{a \to b}$ (b) $h \in H^+_{b \to a}$

Figure 3.12: $E_{\{a,b,h\}}$

Example 53. If E is given in Figure 3.1, then

$$H^+_{g \to b} = \{a, c\} \setminus \{e, h\} = \{a, c\}, \tag{3.33}$$

$$H^+_{b \to g} = \{h\} \setminus \{a, c, e\} = \{h\}, \tag{3.34}$$

$$H^+_{f \to c} = \{a, b, d\} \setminus \{d, i, j\} = \{a, b\} \tag{3.35}$$

$$H^+_{f \to e} = \emptyset \setminus \{d, i, j\} = \emptyset. \tag{3.36}$$

Remark 64. The chain component of E that includes $\{b\}$ (resp. $\{a\}$) also includes $H^+_{a \to b}$ (resp. $H^+_{b \to a}$).

Definition 127. If $c \subseteq H^{+}_{a\to b}$, let $g_{a\to b}(c)$ be the graph obtained from E by

(a) orienting $h - b$ as $h \to b$ for each $h \in c$ and

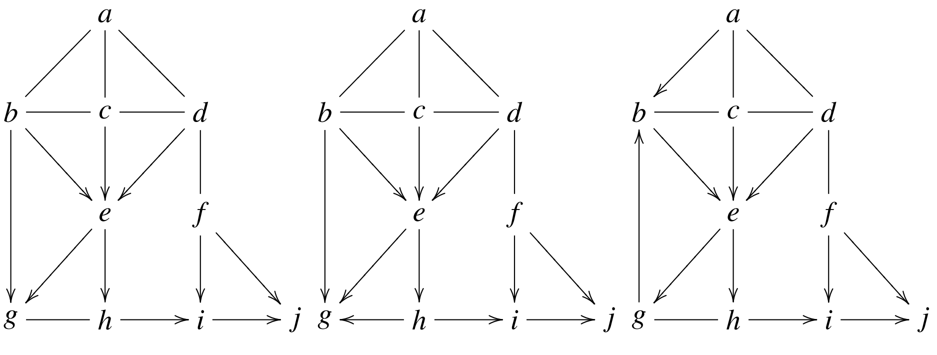
(b) adding $a \to b$.

Definition 128. If $c \subseteq H^{+}_{b\to a}$, let $g_{b\to a}(c)$ be the graph obtained from E by

(a) orienting $h - a$ as $h \to a$ for each $h \in c$ and

(b) adding $b \to a$.

Example 54. If E is given in Figure 3.1, then $g_{b\to g}(\emptyset)$, $g_{b\to g}(\{h\})$ and $g_{g\to b}(\{a\})$ are given in Figure 3.13.



(a) $g_{b\to g}(\emptyset)$ (b) $g_{b\to g}(\{h\})$ (c) $g_{g\to b}(\{a\})$

Figure 3.13: $g_{b\to g}(\emptyset)$, $g_{b\to g}(\{h\})$, and $g_{g\to b}(\{a\})$

Definition 129. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, the set $S^{+}_{a\to b}$ is defined by

$$S^{+}_{a\to b} = \{c \subseteq H^{+}_{a\to b} | c \text{ is empty or complete in } E, \text{ and } g_{a\to b}(c) \text{ has a consistent extension}\}. \quad (3.37)$$

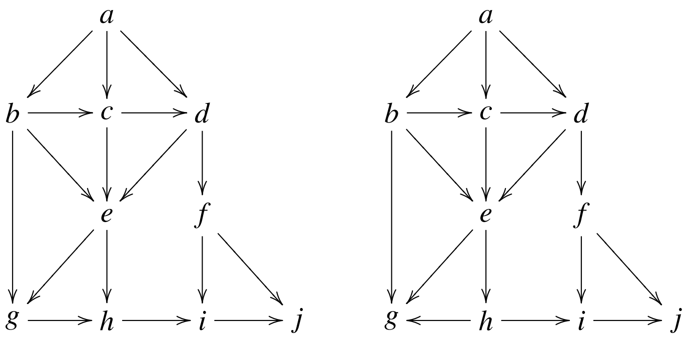
Definition 130. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, the set $S^{+}_{b\to a}$ is defined by

$$S^{+}_{b\to a} = \{c \subseteq H^{+}_{b\to a} | c \text{ is empty or complete in } E, \text{ and } g_{b\to a}(c) \text{ has a consistent extension}\}. \quad (3.38)$$

As noted in Section 3.2.2, there exists an algorithm that checks if a graph has a consistent extension. It is thus possible to enumerate $S^{+}_{a\to b}$ and $S^{+}_{b\to a}$.

Example 55. Suppose that E is given in Figure 3.1. Let us compute $S^{+}_{g\to b}$ and $S^{+}_{b\to g}$. For $c \subseteq H^{+}_{g\to b}$, the graph $g_{g\to b}(c)$ has the cycle b, e, g, b (e.g. see the graph given in Figure 3.13(c)) and thus no consistent extension. Hence, we have

$$S^{+}_{g\to b} = \emptyset. \quad (3.39)$$



(a) (b)

Figure 3.14: Consistent extensions of $g_{b\to g}(\emptyset)$ and $g_{b\to g}(\{h\})$

By Example 53, we have $H^+_{b\to g} = \{h\}$. The graph given in Figure 3.14(a) (resp. Figure 3.14(b)) is a consistent extension of $g_{b\to g}(\emptyset)$ (resp. $g_{b\to g}(\{h\})$). Moreover, $E_{\{h\}}$ is complete. Hence, we have

$$S^+_{b\to g} = \{\emptyset, \{h\}\}. \tag{3.40}$$

Definition 131. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, the function $f^+_{a\to b}$ is defined on $S^+_{a\to b}$ by

$$f^+_{a\to b}(c) = \Big\{(b, \{a, h\})\Big| h \in c \cup (pa_E(b) \backslash ad_E(a))\Big\} \bigcup \Big(v(E) \backslash \{(v, \{a, b\}) \big| v \in V\}\Big). \tag{3.41}$$

Definition 132. If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, let $f^+_{b\to a}$ be the function defined on $S^+_{b\to a}$ by

$$f^+_{b\to a}(c) = \Big\{(a, \{b, h\})\Big| h \in c \cup (pa_E(a) \backslash ad_E(b))\Big\} \bigcup \Big(v(E) \backslash \{(v, \{a, b\}) \big| v \in V\}\Big). \tag{3.42}$$

Remark 65. The functions $f^+_{a\to b}$ and $f^+_{b\to a}$ are injective.

The Image $v(\mathrm{IB}_{\{a,b\}})$

Let us prove that $v(\mathrm{IB}_{\{a,b\}}) = f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a})$. First, Proposition 3.16 states that $v(\mathrm{IB}_{\{a,b\}}) \subseteq f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a})$. Then, Corollary 3.19 states that $f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a}) \subseteq v(\mathrm{IB}_{\{a,b\}})$.

Proposition 3.16. *If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, then*

$$v(\mathrm{IB}_{\{a,b\}}) \subseteq f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a}). \tag{3.43}$$

Proof. Given $s \in v(\mathrm{IB}_{\{a,b\}})$, we show there exists $c \in S^+_{a\to b}$ such that $s = f^+_{a\to b}(c)$ or $c \in S^+_{b\to a}$ such that $s = f^+_{b\to a}(c)$.

1. Let us define a candidate c. First, $s \in v(\mathrm{IB}_{\{a,b\}})$ implies there exists $G \in \mathrm{IB}_{\{a,b\}}$ such that $s = v(G)$. By Definition 119, there exist $K \in [E]$ and $L \in [G]$ such that L is obtained from K by adding an arrow between a and b. Without loss of generality, suppose that $a \to b \in L$ and let

$$c = (pa_K(b) \setminus pa_E(b)) \setminus ad_E(a). \qquad (3.44)$$

2. Let us show that $c \in S^+_{a\to b}$.
   (a) We have $c \subseteq H^+_{a\to b}$ because $pa_K(b) \setminus pa_E(b) \subseteq ne_E(b)$.
   (b) Let us show by contradiction that E_c is complete. We have $c \subseteq pa_K(b)$. There exists a chain component of E such that $c \subseteq H^+_{a\to b} \subseteq \delta$. By Theorem 3.3, $K \in [E]$ implies that K_δ and thus also K_c has no v-structure. If $h_i, h_j \in c$ are distinct and $h_i \cdots h_j \notin K_c$, then $(b, \{h_i, h_j\}) \in v(K_c)$. Hence, for distinct $h_i, h_j \in c$, we have $h_i \cdots h_j \in K_c$, and thus h_i-$- h_j \in E$.
   (c) It is easy to see that L is a consistent extension of $g_{a\to b}(c)$ because $c \subseteq pa_L(b)$ and K is a consistent extension of E.

3. Let us show that $s = v(G) = f^+_{a\to b}(c)$. We have

$$v(G) = (v(G) \setminus v(E)) \cup (v(G) \cap v(E)) = (v(L) \setminus v(K)) \cup (v(L) \cap v(K)). \quad (3.45)$$

   (a) Let us compute $v(L) \setminus v(K)$. It is easy to see that

$$v(L) \setminus v(K) = \left\{(b, \{a, h\})\middle| h \in pa_K(b) \setminus ad_K(a)\right\}. \qquad (3.46)$$

   We have $ad_K(a) = ad_E(a)$ and $pa_K(b) = pa_E(b) \cup (pa_K(b) \setminus pa_E(b))$. Hence, we have $pa_K(b) \setminus ad_K(a) = pa_E(b) \setminus ad_E(a) \cup (pa_K(b) \setminus pa_E(b)) \setminus ad_E(a)$. Hence,

$$v(L) \setminus v(K) = \left\{(b, \{a, h\})\middle| h \in c \cup (pa_E(a) \setminus ad_E(b))\right\}. \qquad (3.47)$$

   (b) It is easy to see that

$$v(L) \cap v(K) = v(K) \setminus \{(v, \{a, b\})|v \in V\} = v(E) \setminus \{(v, \{a, b\})|v \in V\}. \qquad (3.48)$$

□

Let us prove constructively that $f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a}) \subseteq v(\mathrm{IB}_{\{a,b\}})$ with Proposition 3.17 and Proposition 3.18.

Proposition 3.17. *If $c \in S^+_{a\to b}$ and D is a consistent extension of $g_{a\to b}(c)$, then $E(D) \in \mathrm{IB}_{\{a,b\}}$ and $v(E(D)) = f^+_{a\to b}(c)$.*

Proof.

1. Let us show that $E(D) \in \mathrm{IB}_{\{a,b\}}$. It is easy to see that the graph K obtained from D by removing $a \to b$ is a consistent extension of E, i.e. $K \in [E]$. By Definition 119, we thus have $E(D) \in \mathrm{IB}_{\{a,b\}}$.

2. Let us show that $v(E(D)) = f^+_{a\to b}(c)$. We have

$$\begin{aligned} v(E(D)) &= (v(E(D)) \setminus v(E)) \cup (v(E(D)) \cap v(E)) && (3.49)\\ &= (v(D) \setminus v(K)) \cup (v(D) \cap v(K)). && (3.50) \end{aligned}$$

   (a) We have

$$v(D) \setminus v(K) = \Big\{(b, \{a, h\})\Big| h \in c \cup (pa_E(b) \setminus ad_E(a))\Big\}. \tag{3.51}$$

   (b) We have

$$\begin{aligned} v(D) \cap v(K) &= v(K) \setminus \{(v, \{a, b\})\big| v \in V\} && (3.52)\\ &= v(E) \setminus \{(v, \{a, b\})\big| v \in V\}. && (3.53) \end{aligned}$$

□

Proposition 3.18. *If $c \in S^+_{b\to a}$ and D is a consistent extension of $g_{a\to b}(c)$, then $E(D) \in \mathrm{IB}_{\{a,b\}}$ and $v(E(D)) = f^+_{b\to a}(c)$.*

Proof. The proof is similar to the proof of Proposition 3.17. □

Corollary 3.19. *If $E \in \mathcal{E}(V)$ and $a \cdots b \notin E$, then*

$$f^+_{a\to b}(S^+_{a\to b}) \cup f^+_{b\to a}(S^+_{b\to a}) \subseteq v(\mathrm{IB}_{\{a,b\}}). \tag{3.54}$$

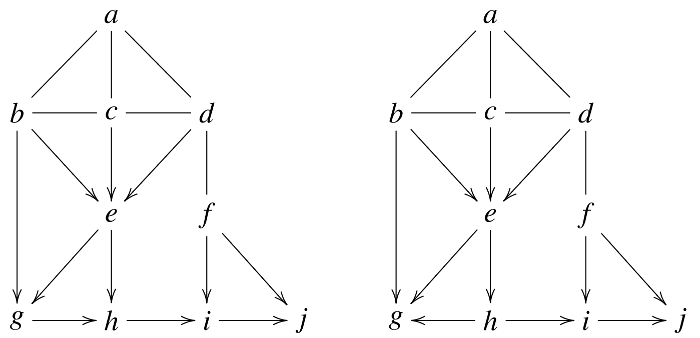
To obtain $\mathrm{IB}_{\{a,b\}}$, Algorithm 3 is applied to the consistent extensions of the graphs in the set

$$\{g_{a\to b}(c)\big| c \in S^+_{a\to b}\} \cup \{g_{b\to a}(c)\big| c \in S^+_{b\to a}\}. \tag{3.55}$$

Example 56. Suppose that E is given in Figure 3.1. By Example 55, $S^+_{g\to b} = \emptyset$ and $S^+_{b\to g} = \{\emptyset, \{h\}\}$. Moreover, consistent extensions of $g_{b\to g}(\emptyset)$ and $g_{b\to g}(\{h\})$ are given in Figure 3.14. Applying Algorithm 3 to these graphs, we obtain $\mathrm{IB}_{\{b,g\}}$ (see Figure 3.15).

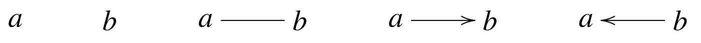
The intersection of $f^+_{a\to b}(S^+_{a\to b})$ and $f^+_{b\to a}(S^+_{b\to a})$ may not be empty. However, the overlap is limited to a most one element since the only possible element in both sets is $v(E) \setminus \{(v, \{a, b\})\big| v \in V\}$.

Example 57. Let E, G, D and D' be given in Figure 3.16. The graphs $g_{a\to b}(\emptyset) = D$ and $g_{b\to a}(\emptyset) = D'$ are DAGs, and thus consistent extensions of themselves. Moreover, $\emptyset = f^+_{a\to b}(\emptyset) = f^+_{b\to a}(\emptyset)$. Note that $E(D) = E(D') = G$.



(a) (b)

Figure 3.15: $\mathrm{IB}_{\{b,g\}}$



(a) E (b) G (c) D (d) D'

Figure 3.16: Non-empty intersection of $f^+_{a\to b}(S^+_{a\to b})$ and $f^+_{b\to a}(S^+_{b\to a})$

An Alternative Description of $S^+_{a\to b}$ and $S^+_{a\to b}$

The following theorem from [Chi02b] allows us to describe $S^+_{a\to b}$ and $S^+_{b\to a}$ without explicitely checking for the existence of a consistent extension.

Theorem 3.20. *Given $c \subseteq H^+_{a\to b}$, the graph $g_{a\to b}(c)$ has a consistent extension if, and only if,*

(a) $c \cup (ne_E(b) \cap ad_E(a))$ is empty or complete in E and

(b) every path in E from b to a has a vertex in $c \cup (ne_E(b) \cap ad_E(a))$.

Corollary 3.21. *We have $c \in S^+_{a\to b}$ if, and only if, $c \subseteq H^+_{a\to b}$, $c \cup (ne_E(b) \cap ad_E(a))$ is empty or complete in E and every path in E from b to a has a vertex in $c \cup (ne_E(b) \cap ad_E(a))$.*

Corollary 3.22. *We have $c \in S^+_{b\to a}$ if, and only if, $c \subseteq H^+_{b\to a}$, $c \cup (ne_E(a) \cap ad_E(b))$ is empty or complete in E and every path in E from a to b has a vertex in $c \cup (ne_E(a) \cap ad_E(b))$.*

Incremental Computation of the Neighbors

Using Corollaries 3.21 and 3.22, it is possible to avoid the computation of consistent extensions and use the MCS algorithm instead.

The following algorithm takes for input $c \in S^+_{a\to b}$, and returns a graph D.

Algorithm 6

```
1. Set $\delta$ to the chain component of $E$ containing $b$.
2. Set $p$ to a permutation of $c \cup (ne_E(b) \cap ad_E(a))$.
3. Set $o$ to a perfect ordering of $E_\delta$ starting with $pb$.
4. Set $D$ to the graph obtained from $E$ by
    (a) Orienting the lines in $E_\delta$ according to $o$ and
    (b) Orienting the lines in the subgraphs induced by the other chain components of $E$ according to perfect orderings.
5. Add $a \to b$ to $D$.
6. Return $D$.
```

□

Remark 66. Algorithm 6 is non-deterministic because of the freedom in the choice of perfect orderings and permutation p.

Proposition 3.23. *If D is the result of Algorithm 6 applied to $c \in S^+_{a \to b}$, then $E(D) \in \mathrm{IB}_{\{a,b\}}$ and $v(E(D)) = f^+_{a \to b}(c)$.*

Proof. Let us show that D is a consistent extension of $g_{a \to b}(c)$ and conclude by Proposition 3.17.

Let K be the graph obtained from D by removing $a \to b$. We have $K \in [E]$, i.e. K is a consistent extension of E. To show that L is a consistent extension of $g_{a \to b}(\{(b, \{a, h\}) | h \in c\})$, it is sufficient to show that L is acyclic, $h \to b \in L$ for $h \in c$, and $b \to h \in L$ for $h \in H^+_{a \to b} \setminus c$.

1. If $h \in c$, then h precedes b in o, and thus $h \to b \in L$.
2. If $h \in H^+_{a \to b} \setminus c$, then b precedes h in o, and thus $b \to h \in L$.
3. Let us show that L is acyclic. Because K is acyclic, L will have a cyle only if it has a path $\pi = b, \dots, a$. Let h be the first vertex in π such that $h \in c \cup (ne_E(b) \cap ad_E(a))$, and let d be the vertex immediately preceding it. If $d \to h \in E$, then E contains the directed cycle $b, \dots, d, h, b$. Hence, $d - h \in E$ and $d \in \delta \setminus (c \cup (ne_E(b) \cap ad_E(a)))$. But d follows h in o, and thus $h \to d \in K$ and $h \to d \in L$. This contradicts the assumption that π is a path, and thus L is acyclic. □

Given $c \in S^+_{a \to b}$, Proposition 3.7 allows us to exploit the non-determinism of Algorithm 6 to speed up the computation of $v^{-1}(f^+_{a \to b}(c))$.

The following algorithm takes for input $c \in S^+_{a \to b}$, and returns a graph G.

Algorithm 7

```
1. Set $\delta$ to the chain component of $E$ containing $b$.
```

2. Set p to a permutation of $c \cup (ne_E(b) \cap ad_E(a))$.
3. Set o to a perfect ordering of E_δ starting with pb.
4. Set $G := E$.
5. Orient according to o each $u - v \in G$ such that $\{u, v\} \subseteq \delta$ and $\{u, v\} \not\subseteq c \cup (ne_E(b) \cap ad_E(a))$.
6. Add $a \to b$ to G.
7. Apply Algorithm 3 to G and return the result. □

Proposition 3.24. *If G is the result of Algorithm 7 applied to $c \in S^+_{a\to b}$ then $G \in$ $\mathrm{IB}_{\{a,b\}}$ and $v(G) = f^+_{a\to b}(c)$.*

Proof. Let $G = (V, E_G)$ be the graph obtained before Step 7, let $D = (V, E_D)$ be the DAG obtained by applying Algorithm 6 to c and using the same perfect ordering o of E_δ, and let $E(D) = (V, E_E)$. We apply Proposition 3.7 to prove that $E(D)$ is the result of Algorithm 3 applied to G. Let us show that the hypotheses are satisfied.

1. Let us show that $E_D \subseteq E_G \subseteq E_E$.
   (a) It is easy to see that D, G, and $E(D)$ share the same skeleton.
   (b) Because Algorithm 6 and Algorithm 7 use the same perfect ordering o, we have $E_D \subseteq E_G$.
   (c) Let us show that $E_G \subseteq E_E$. Let D' be the result of Algorithm 6 applied to c. By Lemma 3.5, we can choose the perfect orderings used to obtain D' so that $u \to v \in D'$ for each $\{u, v\}$ such that $\{u, v\} \subseteq c \cup (ne_E(b) \cap ad_E(a))$ or such that $\{u, v\}$ is included in a chain component of E distinct from δ. Hence, $u - v \in E(D)$ for each such $\{u, v\}$. This implies that $E_G \subseteq E_E$.
2. Let us show by contradiction that G does not induce $x \to y - z$. Suppose that G induces $x \to y - z$ for some vertices x, y, z. We have $y - z \in E$ and either $x \to y \in E$ or $x - y \in E$.
   (a) Suppose $x \to y \in E$. Then, E induces $x \to y - z$ which contradicts Theorem 3.1.
   (b) Suppose that $x - y \in E$. Then, we have $x, y, z \in \delta$ with $x \notin c \cup (ne_E(b) \cap ad_E(a))$ and $y, z \in c \cup (ne_E(b) \cap ad_E(a))$. Hence, y precedes x in o and thus $x \to y \notin G$. □

Example 58. Suppose that E is given in Figure 3.1. If Algorithm 7 is applied to $\emptyset \in S^+_{b\to g}$ and $\{h\} \in S^+_{b\to g}$, the graphs of Figure 3.16 are obtained. In this particular case, it is not even necessary to apply Algorithm 3.

Incremental Computation of the Score

Proposition 3.25. *Suppose the scoring criterion* score *is decomposable and that* $a \cdots b \notin E$. *If* $G = v^{-1}(f^{+}_{a \to b}(c)$, *then*

$$\begin{aligned}\text{score}(G) - \text{score}(E) = f(&b, pa_E(b) \cup c \cup (ne_E(b) \cap ad_E(a)))- \\ & f(b, pa_E(b) \cup c \cup (ne_E(b) \cap ad_E(a)) \cup \{a\}). \quad (3.56)\end{aligned}$$

Proof. Let L be the result of Algorithm 6 applied to c, and let K be the graph obtained from L by removing $a \to b$. We have $\text{score}(G) - \text{score}(E) = \text{score}(L) - \text{score}(K) = f(b, pa_L(b)) - f(b, pa_K(b))$. It is easy to see that $pa_L(b) = pa_E(b) \cup c \cup (ne_E(b) \cap ad_E(a))$ and $pa_K(b) = pa_E(b) \cup c \cup (ne_E(b) \cap ad_E(a)) \cup \{a\}$. □

Again, to compute the increment in score, the knowledge of c is sufficient and it is not necessary to compute G.

3.4 Conclusion

This chapter presented algorithms that efficiently compute the inclusion boundary neighborhood of an arbitrary essential graph as follows. First, elements of the boundary are identified by their skeleton and their set of v-structures. Admissible skeletons and sets of v-structures, i.e corresponding to neighbors, are obtained by performing graphical tests. Using Chickering's results, these tests take an especially simple form (see Corollary 3.21 and Corollary 3.22). Once a neighbor has been identified, its score can be computed incrementally if the scoring criterion is decomposable. The neighbor itself can also be computed incrementally with Algorithm 5 or Algorithm 7.

As discussed in Section 2.5.3, the inclusion boundary can be used as a neighborhood for greedy structure learning algorithms such as UGES. Note that the UGES algorithm is implemented in the *WinMine* toolkit from Microsoft[1] and the *Structure Learning Package* of the BayesNet Matlab toolbox[2]. Depending on the topology of the essential graph, the cardinality of the inclusion boundary can become very large. An interesting question is whether a greedy search will actually encounter a very large inclusion boundary if the initial graph is sparse (say empty) and the data generating distribution admits a sparse inclusion-optimal graph. More generally, the link between the maximal size of the inclusion boundary computed in a greedy search, the independence relations holding in the data generating distribution, and the initial graph should be investigated. Note that another approach to the issue of the size of the boundary in a greedy search can be found in [NKP03].

The identification of a neighbor only requires elementary and very intuitive graphical tests. However, its actual computation is more involved because Algorithm 3 is applied. Consequently, it is difficult to interpret in simple graphical terms

[1] see http://research.microsoft.com/~dmax/winmine/tooldoc.htm

[2] see http://bnt.insa-rouen.fr/

the relation between an essential graph and its inclusion boundary. It is not clear if the complexity of the computation of neighbors is unavoidable. Searching for a more explicit description of neighbors may be an interesting research direction.

Chapter 4

Learning Parameters in Discrete Naive Bayes Models by Computing Fibers of the Parametrization Map

4.1 Introduction

Consider a set $\mathcal{M} = f(\Theta)$ of probability distributions over a set X of discrete random variables, for instance a discrete Bayesian network model or a discrete Naive Bayes model with hidden class variable. Given a sequence of n independent observations, a maximum likelihood estimate is a parameter $\theta \in \Theta$ satisfying

$$\theta \in \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} (f(\theta))(o[i]). \tag{4.1}$$

Finding a maximum likelihood estimate is an optimisation problem that has a simple solution when there is no constraint and the observations are complete for X. Given a sequence of n observations complete for X, let $\hat{p}$ be the distribution of relative frequencies, i.e. the distribution over X such that

$$\hat{p}(x) = \frac{n_x}{n}, \quad x \in \mathcal{X}, \tag{4.2}$$

where n_x is the number of observations o in d such that $o = x$. As shown in Appendix C, we have

$$\arg\max_{p} \prod_{i=1}^{n} p(o[i]) = \{\hat{p}\}. \tag{4.3}$$

If the observations are complete, it is well-known that $\theta \in \Theta$ maximizes the likelihood if, and only if,

$$\theta \in \arg\min_{\theta\in\Theta} D(\hat{p} \parallel f(\theta)), \tag{4.4}$$

where D measures the Kullback-Leibler distance (see Appendix C). In other words, the distribution $f(\theta)$ associated to a maximum likelihood estimate θ can be interpreted as a projection of $\hat{p}$ onto $\mathcal{M}$ that minimizes the Kullback-Leibler distance.

In this light, a function $\pi : \Lambda \to \Theta$ defined on a set Λ of distributions over X specifies a parameter learning algorithm taking for input a dataset such that $\hat{p} \in \Lambda$ and returning the learned parameter $\pi(\hat{p})$ (see Figure 4.1). As discussed above, such a function is optimal for parameter learning if $(f \circ \pi)(p)$ minimizes the Kullback-Leibler distance from p to $\mathcal{M}$. To design a function $\pi : \Lambda \to \Theta$ in practice, the following constraints may be imposed:

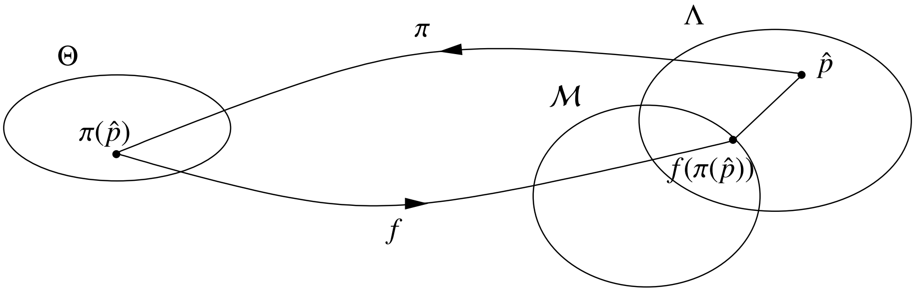


Figure 4.1: Parameter learning in terms of projections. The distance on paper between $p \in \Lambda$ and $q \in \mathcal{M}$ is assumed to be proportional to $D(p \parallel q)$.

- $(f \circ \pi)(p) = p$ for $p \in \Lambda \cap \mathcal{M}$,
- $\Lambda \cap \mathcal{M}$ is included in the interior of Λ, and
- $f \circ \pi$ is continuous at all $p \in \Lambda \cap \mathcal{M}$.

In Section 4.2, the set of all probability distributions on X is represented by a set $S_X \subseteq \mathbb{R}^{|\mathcal{X}|}$. This representation allows us to transpose the euclidian distance to distributions and use the topology induced from the euclidian topology of $\mathbb{R}^{|\mathcal{X}|}$. The first constraint guarantees that π is optimal when $\hat{p} \in \Lambda \cap \mathcal{M}$. The second guarantees that π is defined on distributions sufficiently close to $\Lambda \cap \mathcal{M}$. Together, these three constraints lead to the following property:

$$\begin{aligned}&(\forall p \in \Lambda \cap \mathcal{M})(\forall \epsilon > 0)\Big(\exists \delta > 0 \text{ s.t.}\\ &\qquad (\hat{p} \in S_X \text{ and } |p - \hat{p}| < \delta) \Rightarrow (\hat{p} \in \Lambda \text{ and } |p - (f \circ \pi)(\hat{p})| < \epsilon)\Big),\end{aligned} \tag{4.5}$$

which translates into a consistency property by the strong law of large numbers.

Example 59. If $\mathcal{M}$ is a discrete Bayesian network model $\mathcal{M}_d(G) = f_{d,G}(\Theta_{d,G})$, the constraint $(f_{d,G} \circ \pi)(p) = p$ for $p \in \Lambda \cap \mathcal{M}_d(G)$ is an excellent starting point to obtain a suitable π. As discussed in Section 1.4.2, the parametrization map $f_{d,G}$ is injective and we have

$$f_{d,G}^{-1}(p) = \Big(\big((\theta_{x_v}^{X_v, x_{pa(v)}})_{x_v \in \mathcal{X}_v}\big)_{x_{pa(v)} \in \mathcal{X}_{pa(v)}}\Big)_{X_v \in X} \tag{4.6}$$

for $p \in \mathcal{M}_d(G)$, with

$$\theta_{x_v}^{X_v, x_{pa(v)}} = \frac{p(x_v, x_{pa(v)})}{p(x_{pa(v)})}. \tag{4.7}$$

Equation (4.7) defines a parameter whenever p is strictly positive. Hence, an obvious choice is to define π on the set of strictly positive distributions for X by

$$\pi(p) = \Big(((\theta_{x_v}^{X_v, x_{pa(v)}})_{x_v \in \mathcal{X}_v})_{x_{pa(v)} \in \mathcal{X}_{pa(v)}}\Big)_{X_v \in X} \tag{4.8}$$

with $\theta_{x_v}^{X_v, x_{pa(v)}}$ given by (4.7). In fact, it turns out that this choice for π is optimal:

$$\{\pi(\hat{p})\} = \arg \min_{\theta \in \Theta_{d,G}} D(\hat{p} \parallel f_{d,G}(\theta)) \tag{4.9}$$

if $\hat{p}$ is strictly positive by (2.26).

This chapter presents a family of functions π satisfying the above requirements for the class of discrete Naive Bayes models with hidden class variable. The above ideas have already been applied in [Pea88] for the special case of discrete Naive Bayes models with two classes and binary variables (see also [GHKM01] for the computation of fibers), and we do not claim that they are novel. The interest of this work lies in their actual implementation for the larger class of models considered and in the theorems developped to that end. The resulting parameter learning algorithms are preliminary and should be seen as a proof of concept. They suggest that the constraints on π may lead to interesting parameter learning algorithms for discrete Naive Bayes models with hidden class variable. However, we do not claim that our algorithms will be successful in practice. In particular, their computational complexity is very large and they were not tested extensively (no experimental result is given). Parts of this chapter were published in [AGW06].

Section 4.2 introduces discrete Naive Bayes models with hidden class variable. Section 4.3 presents definitions and technical results constituting the main contribution of this chapter. With those preliminary results, Section 4.4 introduces two algorithms to compute fibers of the parametrization map. Section 4.5 derives projection functions that satisfy our requirements. Section 4.6 concludes. To lighten the presentation, some proofs are gathered in Section 4.7.

4.2 The Discrete Naive Bayes Model with hidden class

This section defines discrete Naive Bayes models with hidden class variable and presents some elementary properties.

4.2.1 Parametric Definition

If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables, a distribution $p(x)$ is represented by the vector $(p_x)_{x \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}$ such that $p_x = p(x)$ for $x \in \mathcal{X}$. Also, if $S \subseteq X$ and $s \in \mathcal{S}$, then p_s denotes the marginal probability $p(s)$. The set of all probability distributions over X is represented as follows.

Definition 133. If $X = \{X_1, \dots, X_n\}$ is a set of discrete random variables, let

$$S_X = \Big\{(p_x)_{x\in\mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|} \Big| \sum_{x\in\mathcal{X}} p_x = 1, (\forall x \in \mathcal{X} : p_x \geq 0)\Big\}. \tag{4.10}$$

Remark 67. The set S_X is a semi-algebraic subset of $\mathbb{R}^{|\mathcal{X}|}$ of dimension

$$d(S_X) = |\mathcal{X}| - 1. \tag{4.11}$$

Without loss of generality, the random variables in X are assumed to have pairwise disjoint sets of possible values. Then, a random variable is uniquely identified by one of its values and the set of all values $\cup_{X_i\in X}\mathcal{X}_i$ is in bijection with $\cup_{X_i\in X}\cup_{x_i\in\mathcal{X}_i}(x_i, X_i)$. The parameter space of a discrete Naive Bayes model with m hidden classes over X is defined as follows.

Definition 134. If $X = \{X_1, \dots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the set $\Theta_{m,X}$ is defined by

$$(\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^n\mathcal{X}_i})_{t=1}^m \in \Theta_{m,X} \tag{4.12}$$

if, and only if,

$$\omega_t > 0, \qquad \theta_{t,x_i} > 0, \tag{4.13}$$

$$\sum_{t=1}^m \omega_t = 1, \qquad \sum_{x_i\in\mathcal{X}_i} \theta_{t,x_i} = 1 \tag{4.14}$$

for $t \in \{1, \dots, m\}$, $X_i \in X$, and $x_i \in \mathcal{X}_i$.

Remark 68. The set $\Theta_{m,X}$ is a semi-algebraic subset of $\mathbb{R}^{m+m\sum_{X_i\in X}|\mathcal{X}_i|}$ (and a smooth manifold) of dimension $d(\Theta_{m,X})$ given by

$$d(\Theta_{m,X}) = (m-1) + m\sum_{X_i\in X}(|\mathcal{X}_i| - 1). \tag{4.15}$$

Remark 69. The components of a parameter in $\Theta_{m,X}$ are strictly positive. The results presented in this chapter do not fundamentally depend on this restriction, but it does simplify some proofs and statements.

The parametrization map is defined as follows.

Definition 135. If $X = \{X_1, \dots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the function $f_{m,X} : \Theta_{m,X} \to S_X$ is defined by

$$f_{m,X}\Big((\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^n\mathcal{X}_i})_{t=1}^m\Big) = (p_{(x_1,\dots,x_n)})_{(x_1,\dots,x_n)\in\mathcal{X}}, \tag{4.16}$$

where

$$p_{(x_1,\dots,x_n)} = \sum_{t=1}^m \omega_t \prod_{i=1}^n \theta_{t,x_i}. \tag{4.17}$$

The discrete Naive Bayes model (with hidden class variable) over X is defined parametrically as follows.

Definition 136 (*Discrete Naive Bayes model*). If $X = \{X_1, \dots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the *discrete Naive Bayes model* $\mathcal{NB}_{m,X}$ with m classes is defined by

$$\mathcal{NB}_{m,X} = f_{m,X}(\Theta_{m,X}). \tag{4.18}$$

Remark 70. A discrete Naive Bayes model with hidden class variable is a Bayesian network model with hidden variables: if $X = \{X_1, \dots, X_n\}$ is a set of discrete random variables, H is a discrete random variable with $m = |\mathcal{H}|$ values, and G is the structure over $X \cup \{H\}$ given in Figure 4.2, then $\mathcal{NB}_{m,X}$ is the Bayesian network model with hidden variable H obtained from $\mathcal{M}_d(G)$.

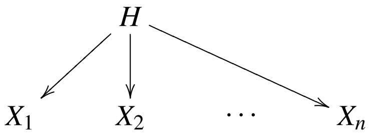


Figure 4.2: Naive Bayes structure G

Remark 71. Distributions in $\mathcal{NB}_{m,X}$ can be interpreted as mixtures. Indeed, consider a distribution

$$p = (p_{(x_1,\dots,x_n)})_{(x_1,\dots,x_n)\in\mathcal{X}} = f_{m,X}\Big(\big(\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^{n}\mathcal{X}_i}\big)_{t=1}^{m}\Big). \tag{4.19}$$

For $t \in \{1, \dots, m\}$, let $p^t = (p^t_{(x_1,\dots,x_n)})_{(x_1,\dots,x_n)\in\mathcal{X}} \in S_X$ with

$$p^t_{(x_1,\dots,x_n)} = \prod_{i=1}^{n} \theta_{t,x_i}. \tag{4.20}$$

Each p^t satisfies $p^t \in \mathcal{M}_d(D_e)$, where D_e is the empty DAG over X, and we have

$$p = \sum_{t=1}^{m} \omega_t p^t. \tag{4.21}$$

Hence, p is a mixture of m distributions $p^1, \dots, p^m \in \mathcal{M}_d(D_e)$ with mixing coefficients $\omega_1, \dots, \omega_m$.

Figure 4.3 summarizes the relationships between the objects defined in this section and a potential projection function π. It is a special case of Figure 4.1.

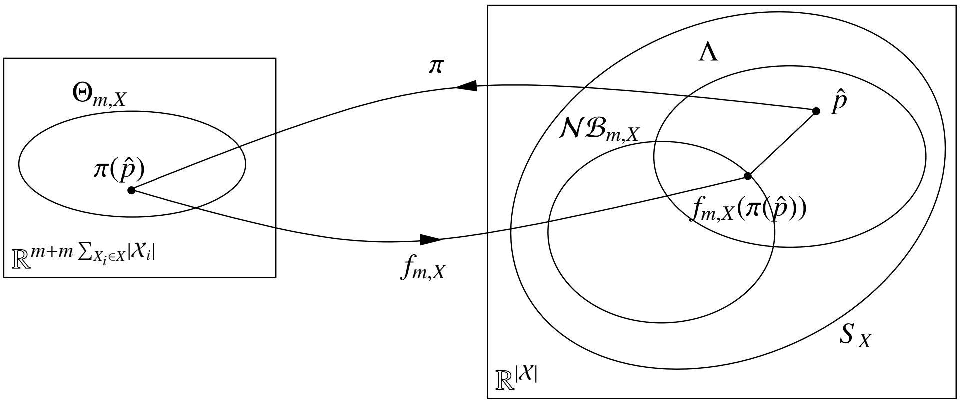


Figure 4.3: Parameter learning for $\mathcal{NB}_{m,X}$ in terms of projections.

4.2.2 Elementary Properties

Discrete Naive Bayes models are nested.

Proposition 4.1. $\mathcal{NB}_{m,X} \subseteq \mathcal{NB}_{m+1,X}$.

Proof. Consider $p = f_{m,X}(\theta)$ with

$$\theta = (\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^n\mathcal{X}_i})_{t=1}^m \in \Theta_{m,X}. \tag{4.22}$$

There exists an element

$$\theta' = (\omega_t', (\theta_{t,x_i}')_{x_i\in\cup_{i=1}^n\mathcal{X}_i})_{t=1}^{m+1} \in \Theta_{m+1,X} \tag{4.23}$$

such that

$$\begin{aligned}
\omega_t' &= \omega_t && \text{for } t \in \{1,\dots,m-1\}, && (4.24)\\
\omega_m' + \omega_{m+1}' &= \omega_m, && && (4.25)\\
\theta_{t,x_i}' &= \theta_{t,x_i} && \text{for } t \in \{1,\dots,m\} \text{ and } x_i \in \cup_{i=1}^n\mathcal{X}_i, && (4.26)\\
\theta_{m+1,x_i}' &= \theta_{m,x_i} && \text{for } x_i \in \cup_{i=1}^n\mathcal{X}_i. && (4.27)
\end{aligned}$$

For such an element θ', we have $f_{m+1,X}(\theta') = p \in \mathcal{NB}_{m+1,X}$. Hence, $\mathcal{NB}_{m,X} \subseteq \mathcal{NB}_{m+1,X}$. □

Despite their simplicity, discrete Naive Bayes models are versatile: any distribution on X can be approximated arbitrarily closely by a distribution in $\mathcal{NB}_{m,X}$ for sufficiently large m. The closure of a set A is denoted $\overline{A}$.

Proposition 4.2. *If* $m \geq |\mathcal{X}|/(\max_{X_i\in X}|\mathcal{X}_i|)$, *then* $\overline{\mathcal{NB}_{m,X}} = S_X$.

Proof. The proof is adapted from [KZ02].

1. By definition of $f_{m,X}$, we have $\mathcal{NB}_{m,X} \subseteq S_X$. Also, we have $\overline{S_X} = S_X$. Hence, $\overline{\mathcal{NB}_{m,X}} \subseteq S_X$.

2. Without loss of generality, suppose that X_n has maximum cardinality and let $m = |\mathcal{X}|/|\mathcal{X}_n| = \prod_{i=1}^{n-1}|\mathcal{X}_i|$. Since $\mathcal{NB}_{k,X} \subseteq \mathcal{NB}_{k+1,X}$, it is sufficient to prove that $S_X \subseteq \overline{\mathcal{NB}_{m,X}}$. Consider $p = (p_{x_1,\ldots,x_n})_{(x_1,\ldots,x_n)\in\mathcal{X}} \in S_X$. For $\epsilon > 0$, let

$$\theta_\epsilon = (\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i})_{t\in\mathcal{X}_1\times\ldots\mathcal{X}_{n-1}} \tag{4.28}$$

be such that

$$\omega_{(x_1,\ldots,x_{n-1})} = p_{x_1,\ldots,x_{n-1}}, \tag{4.29}$$

$$\theta_{(x_1,\ldots,x_{n-1}),x_i'} = \begin{cases} \epsilon & \text{if } x_i' \neq x_i, \\ 1 - \epsilon(|\mathcal{X}_i| - 1) & \text{if } x_i' = x_i, \end{cases} \tag{4.30}$$

$$\theta_{(x_1,\ldots,x_{n-1}),x_n} = p_{x_1,\ldots,x_n}/p_{x_1,\ldots,x_{n-1}}. \tag{4.31}$$

for $i \in \{1,\ldots,n-1\}$, $x_n \in \mathcal{X}_n$ and $(x_1,\ldots,x_{n-1}) \in \mathcal{X}_1 \times \ldots \mathcal{X}_{n-1}$. If ϵ is sufficiently small, it is easy to see that $\theta_\epsilon \in \Theta_{m,X}$. Moreover, we have $\lim_{\epsilon\to 0^+} f_{m,X}(\theta_\epsilon) = p$ and thus $p \in \overline{\mathcal{NB}_{m,X}}$. Therefore $S_X \subseteq \overline{\mathcal{NB}_{m,X}}$. □

From the outset, one can identify two reasons why the parametrization map $f_{m,X}$ is non-injective: *aliasing* and the inclusion relation $\mathcal{NB}_{m,X} \subseteq \mathcal{NB}_{m+1,X}$. The *aliasing* phenomenon is a consequence of the commutativity of the sum in (4.17). It can be described as follows: if σ is a permutation of the set $\{1,\ldots,m\}$ and

$$\theta = (\omega_t, (\theta_{t,x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i})_{t=1}^m \in \Theta_{m,X}, \tag{4.32}$$

it is easy to see that

$$\theta' = (\omega_{\sigma(t)}, (\theta_{\sigma(t),x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i})_{t=1}^m \in \Theta_{m,X}, \tag{4.33}$$

and $f_{m,X}(\theta) = f_{m,X}(\theta')$. In other words, the mixture components can be freely permuted. As shown in the proof of Proposition 4.1, there exists infinitely many ways to parametrize in $\mathcal{NB}_{m+1,X}$ a distribution that belongs to $\mathcal{NB}_{m,X}$. If $p \in \mathcal{NB}_{m,X}$, the preimage $f_{m+1,X}^{-1}(p)$ is thus not finite.

4.3 Preliminaries

This chapter contains material necessary to develop our algorithms. First, an alternative parametrization of $\mathcal{NB}_{m,X}$ is introduced. Then, theorems that lie at the core of this chapter and its algorithms are presented. Despite the alternative parametrization, these results are not easily formulated as they require the definition of several families of functions. To simplify the presentation, the proofs of our core results are deferred until Section 4.7.

4.3.1 Alternative Parametrization of $\mathcal{NB}_{m,X}$

First, a new parameter space $\Pi_{m,X}$ in bijection with $\Theta_{m,X}$ is provided. Then, probability distributions are described by elements of a set R_X in bijection with S_X. Finally, the new parametrization map $h_{m,X}$ is described.

Alternative Parameter Space

The idea behind the new parameter space $\Pi_{m,X}$ is very simple. A new parameter in $\Pi_{m,X}$ corresponding to $\theta = (\omega_t, (\theta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m \in \Theta_{m,X}$ keeps the components ω_t, but, instead of θ_{t,x_i}, it has the components $\delta_{t,x_i} = \theta_{t,x_i} - \sum_{t=1}^m \omega_t \theta_{t,x_i}$ and $\lambda_{x_i} = \sum_{t=1}^m \omega_t \theta_{t,x_i}$. An important consequence of this change of variables is that $\sum_{t=1}^m \omega_t \delta_{t,x_i} = 0$. Formally, the new parameter space is defined as follows.

Definition 137. If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the set $\Pi_{m,X}$ is defined by

$$\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \in \Pi_{m,X} \tag{4.34}$$

if, and only if,

$$\omega_t > 0, \qquad \delta_{t,x_i} + \lambda_{x_i} > 0, \qquad \sum_{x_i \in \mathcal{X}_i} \lambda_{x_i} = 1, \tag{4.35}$$

$$\sum_{t=1}^m \omega_t = 1, \qquad \sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = 0, \qquad \sum_{t=1}^m \omega_t \delta_{t,x_i} = 0 \tag{4.36}$$

for $t \in \{1, \ldots, m\}$, $X_i \in X$ and $x_i \in \mathcal{X}_i$.

The correspondence between $\Pi_{m,X}$ and $\Theta_{m,X}$ is defined by the following function.

Definition 138. If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the function $\phi : \Pi_{m,X} \to \Theta_{m,X}$ is defined by

$$\phi\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) = (\omega_t, (\theta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m \tag{4.37}$$

where

$$\theta_{t,x_i} = \delta_{t,x_i} + \lambda_{x_i}. \tag{4.38}$$

Finally, we make sure that $\Pi_{m,X}$ and $\Theta_{m,X}$ can be exchanged.

Proposition 4.3. *The function ϕ is a bijection with inverse given by*

$$\phi^{-1}\Big((\omega_t, (\theta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m\Big) = \Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \tag{4.39}$$

where

$$\lambda_{x_i} = \sum_{t=1}^m \omega_t \theta_{t,x_i}, \tag{4.40}$$

$$\delta_{t,x_i} = \theta_{t,x_i} - \sum_{t=1}^m \omega_t \theta_{t,x_i}. \tag{4.41}$$

Proof. It is straighforward to prove that ϕ is injective, and that the function ϕ^{-1} is its inverse. □

Remark 72. The bijection ϕ and its inverse are polynomial. As a consequence, $\Pi_{m,X}$ is a semi-algebraic set with the same dimension as $\Theta_{m,X}$.

Alternative Parametrization of Distributions

To complement the new parameter space, an alternative description of distributions over X is introduced. It can be seen as an extension to non-binary variables of a similar description in [GHKM01]. As shown in the next section, the new description results in a new parametrization map very similar to $f_{m,X}$.

Definition 139. If $X = \{X_1, \ldots, X_n\}$ is a set of random variables, $S \subseteq X$, and $x = (x_1, \ldots, x_n) \in \mathcal{X}$, let x_S denote the value $(x_i)_{X_i \in S} \in \mathcal{S}$.

Definition 140. If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables, the set R_X is defined by

$$\left((\lambda_{x_i})_{x_i \in \cup_{i=1}^{n} \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\right) \in R_X \tag{4.42}$$

if, and only if,

$$q_{()} = 1 \tag{4.43}$$

$$q_{x_i} = 0 \quad \text{for } x_i \in \mathcal{X}_i \text{ and } X_i \in X \tag{4.44}$$

$$\sum_{x_i \in \mathcal{X}_i} \lambda_{x_i} = 1 \quad \text{for } X_i \in X \tag{4.45}$$

$$\sum_{s_i \in \mathcal{S}_i} q_{(s_1, \ldots, s_i, \ldots, s_{|S|})} = 0 \quad \text{for } S \subseteq X, S_i \in S, \text{ and, with } Y = S \setminus \{S_i\}, \; (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_{|S|}) \in \mathcal{Y} \tag{4.46}$$

$$\sum_{S \subseteq X} q_{x_S} \prod_{i \in X \setminus S} \lambda_{x_{\{i\}}} \geq 0 \quad \text{for } x \in \mathcal{X} \tag{4.47}$$

$$\sum_{x \in \mathcal{X}} \sum_{S \subseteq X} q_{x_S} \prod_{i \in X \setminus S} \lambda_{x_{\{i\}}} = 1. \tag{4.48}$$

Elements of S_X and R_X are connected by the following function.

Definition 141. If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables, the function $\psi : S_X \to R_X$ is defined recursively by

$$\psi((p_x)_{x \in \mathcal{X}}) = \left((\lambda_{x_i})_{x_i \in \cup_{i=1}^{n} \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\right) \tag{4.49}$$

where

$$\lambda_{x_i} = p_{x_i} \tag{4.50}$$

$$q_s = p_s - \sum_{P \subsetneq S} q_{s_P} \prod_{i \in S \setminus P} p_{s_{\{i\}}}. \tag{4.51}$$

Remark 73. In the above definition, we adopt the convention that $p_{()} = 1$. It is straighforward to check that $\psi(S_X) \subseteq R_X$.

Example 60. If X_i, X_j and X_k are distinct variables in X and

$$\psi((p_x)_{x\in\mathcal{X}}) = \Big((\lambda_{x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big), \tag{4.52}$$

then

$$q_{(x_i,x_j)} = p_{(x_i,x_j)} - p_{x_i}p_{x_j}, \tag{4.53}$$

$$\begin{aligned} q_{(x_i,x_j,x_k)} = {} & p_{(x_i,x_j,x_k)} - p_{x_i}p_{(x_j,x_k)} - p_{x_j}p_{(x_i,x_k)} \\ & - p_{x_k}p_{(x_i,x_j)} + 2p_{x_i}p_{x_j}p_{x_k}. \end{aligned} \tag{4.54}$$

Proposition 4.4. *The function ψ is a bijection with inverse given by*

$$\psi^{-1}\Big((\lambda_{x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big) = (p_x)_{x\in\mathcal{X}} \tag{4.55}$$

where

$$p_x = \sum_{S\subseteq X} q_{x_S} \prod_{i\in X\setminus S} \lambda_{x_{\{i\}}}. \tag{4.56}$$

Proof. It is straighforward to prove that ψ is injective, and that the function ψ^{-1} is its inverse. □

Remark 74. The bijection ψ and its inverse are polynomial. As a consequence, R_X is a semi-algebraic set with the same dimension as S_X.

Marginal independence in a distribution represented by $q \in R_X$ has a natural expression. Note that the λ_{x_i} components are not necessary to express marginal independence.

Proposition 4.5. *Let X be a set of discrete random variables, let A and B be non-empty and disjoint subsets of X, and let p be a distribution for X with*

$$\psi(p) = \Big((\lambda_{x_i})_{x_i\in\cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big). \tag{4.57}$$

The sets of random variables A and B are independent in p if, and only if, for all $C \subseteq A$, $c \in C$, $D \subseteq B$ and $d \in \mathcal{D}$, we have

$$q_{(c,d)} = q_c q_d. \tag{4.58}$$

Proof. Let us prove the proposition by induction.

1. Suppose that $|A \cup B| = 2$. We have $A \perp B$ if, and only if, $p_{(a,b)} - p_a p_b = 0$ for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$. By definition of ψ, we have

$$q_{(a,b)} = p_{(a,b)} - p_a p_b, \quad q_a = q_b = 0. \tag{4.59}$$

Hence, $p_{(a,b)} - p_a p_b = 0$ if, and only if, $q_{(a,b)} = q_a q_b$.

2. Suppose that $|A \cup B| > 2$. We have $A \perp B$ if, and only if, $p_{(a,b)} - p_a p_b = 0$ for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$. On the other hand, $A \perp B$ implies that $C \perp D$ for all $C \subseteq A$ and $D \subseteq B$. By inductive hypothesis, $A \perp B$ if and only if,

   - $p_{(a,b)} - p_a p_b = 0$ for all $a \in \mathcal{A}$ and $b \in \mathcal{B}$,
   - $q_{(c,d)} = q_c q_d$ for all $c \in \mathcal{C}$ and $d \in \mathcal{D}$ such that $C \subseteq A$, $D \subseteq B$ and $|C \cup D| < |A \cup B|$.

   Hence, it is sufficient to prove that $p_{(a,b)} - p_a p_b = q_{(a,b)} - q_a q_b$ under the assumption that $q_{(c,d)} = q_c q_d$ for all $c \in \mathcal{C}$ and $d \in \mathcal{D}$ such that $C \subseteq A$, $D \subseteq B$ and $|C \cup D| < |A \cup B|$. By Proposition 4.4, we have

$$p_{(a,b)} = \sum_{P \subseteq A \cup B} q_{(a,b)_P} \prod_{i \in (A \cup B) \setminus P} \lambda_{(a,b)_{\{i\}}} \tag{4.60}$$

$$p_a p_b = \Big(\sum_{P \subseteq A} q_{a_P} \prod_{i \in A \setminus P} \lambda_{a_{\{i\}}}\Big)\Big(\sum_{P \subseteq B} q_{b_P} \prod_{i \in B \setminus P} \lambda_{b_{\{i\}}}\Big). \tag{4.61}$$

   Hence, we have

$$p_{(a,b)} = q_{(a,b)} + \sum_{P \subsetneq A \cup B} q_{(a,b)_P} \prod_{i \in (A \cup B) \setminus P} \lambda_{(a,b)_{\{i\}}} \tag{4.62}$$

$$= q_{(a,b)} + \sum_{P \subsetneq A \cup B} q_{a_{P \cap A}} q_{b_{P \cap B}} \prod_{i \in (A \cup B) \setminus P} \lambda_{(a,b)_{\{i\}}}, \tag{4.63}$$

   where the latter equality holds by inductive hypothesis. Also, we have

$$p_a p_b = \sum_{P \subseteq A \cup B} q_{a_{P \cap A}} q_{b_{P \cap B}} \prod_{i \in (A \cup B) \setminus P} \lambda_{(a,b)_{\{i\}}} \tag{4.64}$$

$$= q_a q_b + \sum_{P \subsetneq A \cup B} q_{a_{P \cap A}} q_{b_{P \cap B}} \prod_{i \in (A \cup B) \setminus P} \lambda_{(a,b)_{\{i\}}}. \tag{4.65}$$

   It is now easy to see that $p_{(a,b)} - p_a p_b = q_{(a,b)} - q_a q_b$. □

Example 61. If $p \in \mathcal{NB}_{1,X}$ and

$$\psi(p) = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big), \tag{4.66}$$

then, for $S \subseteq X$, $S \neq \emptyset$ and $s \in \mathcal{S}$, we have

$$q_s = \prod_{i \in S} q_{s_{\{i\}}} = 0, \tag{4.67}$$

since all the variables are independent.

Alternative Parametrization Map

The new parametrization map $h_{m,X}$ is defined as follows. Note how (4.69) is similar to (4.17).

Definition 142. If $X = \{X_1, \ldots, X_n\}$ is a set of discrete random variables and m is an integer ≥ 1, the function $h_{m,X} : \Pi_{m,X} \to R_X$ is defined by

$$h_{m,X}\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big), \tag{4.68}$$

where

$$q_s = \sum_{t=1}^m \omega_t \prod_{i \in S} \delta_{t,s_{\{i\}}}. \tag{4.69}$$

Proposition 4.6. *We have $h_{m,X} = \psi \circ f_{m,X} \circ \phi$.*

Proof. Given $\pi = \Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big)$, let

$$\Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) = h_{m,X}(\pi), \tag{4.70}$$

$$\Big((\lambda'_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q'_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) = (\psi \circ f_{m,X} \circ \phi)(\pi). \tag{4.71}$$

Let us show that $h_{m,X}(\pi) = (\psi \circ f_{m,X} \circ \phi)(\pi)$. We have

$$\phi(\pi) = (\omega_t, (\delta_{t,x_i} + \lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m. \tag{4.72}$$

Letting $((p_x)_{x \in \mathcal{X}}) = (f_{m,X} \circ \phi)(\pi)$, we thus have

$$p_x = \sum_{t=1}^m \omega_t \prod_{i \in X} \left(\delta_{t,x_{\{i\}}} + \lambda_{x_{\{i\}}}\right). \tag{4.73}$$

Hence, we have

$$\lambda'_{x_i} = p_{x_i} = \sum_{t=1}^m \omega_t \left(\delta_{t,x_i} + \lambda_{x_i}\right) = \lambda_{x_i}, \tag{4.74}$$

and

$$q'_s = p_s - \sum_{P \subsetneq S} q'_{s_P} \prod_{i \in S \setminus P} p_{s_{\{i\}}}. \tag{4.75}$$

To conclude the proof, let us show inductively that $q'_s = q_s$.

1. For $S = \emptyset$, we have

$$q'_{()} = 1 = \sum_{t=1}^m \omega_t = q_{()}. \tag{4.76}$$

2. For $|S| \geq 1$, by inductive hypothesis and because $p_{s_{\{i\}}} = \lambda_{s_{\{i\}}}$, we have

$$q'_s = p_s - \sum_{P \subsetneq S} q_{s_P} \prod_{i \in S \setminus P} \lambda_{s_{\{i\}}} \tag{4.77}$$

$$= p_s - \sum_{P \subsetneq S} \Big(\sum_{t=1}^{m} \omega_t \prod_{i \in P} \delta_{t,s_{\{i\}}} \Big) \Big(\prod_{i \in S \setminus P} \lambda_{s_{\{i\}}} \Big). \tag{4.78}$$

On the other hand, by (4.73), we have

$$p_s = \sum_{t=1}^{m} \omega_t \prod_{i \in S} (\delta_{t,s_{\{i\}}} + \lambda_{s_{\{i\}}}) \tag{4.79}$$

$$= \sum_{t=1}^{m} \omega_t \sum_{P \subseteq S} \Big(\prod_{i \in P} \delta_{t,s_{\{i\}}} \Big) \Big(\prod_{i \in S \setminus P} \lambda_{s_{\{i\}}} \Big), \tag{4.80}$$

$$= \sum_{P \subseteq S} \Big(\sum_{t=1}^{m} \omega_t \prod_{i \in P} \delta_{t,s_{\{i\}}} \Big) \Big(\prod_{i \in S \setminus P} \lambda_{s_{\{i\}}} \Big). \tag{4.81}$$

Hence, we have $q'_s = q_s$. □

Corollary 4.7. *We have $f_{m,X}^{-1} = \phi \circ h_{m,X}^{-1} \circ \psi$.*

Figure 4.4 summarizes the link between the old and new parametrization. Note that $h_{m,X}(\Pi_{m,X}) = \psi(\mathcal{NB}_{m,X})$.

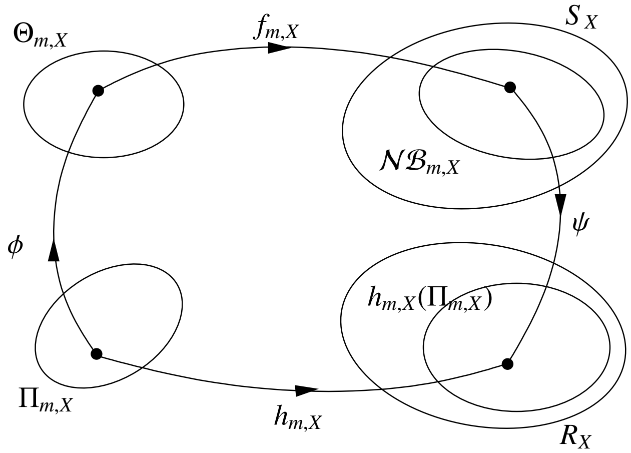


Figure 4.4: New parametrization of $\mathcal{NB}_{m,X}$

4.3.2 Notations

This section introduces notations to manipulate vectors and matrices. In matrix operations, vectors of $\mathbb{R}^m$ are considered row vectors, i.e. elements of $\mathbb{R}^{1\times m}$.

Definition 143. If A is a $m \times k$ matrix and $p \subsetneq \{1, \dots, m\}$, let $A^{\hat{p}}$ be the $(m - |p|) \times k$ matrix obtained from A by removing the rows with indices in p.

Definition 144. If A is a $m \times k$ matrix and $p \subsetneq \{1, \dots, k\}$, let $A_{\hat{p}}$ be the $m \times (k - |p|)$ matrix obtained from A by removing the columns with indices in p.

Definition 145. Let A be a $m \times k$ matrix given by

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mk} \end{pmatrix}. \tag{4.82}$$

If $p = (p_1, \dots, p_l) \in \{1, \dots, m\}^l$, let A^p denote the $l \times k$ matrix given by

$$A^p = \begin{pmatrix} a_{p_1 1} & \cdots & a_{p_1 k} \\ \vdots & & \vdots \\ a_{p_l 1} & \cdots & a_{p_l k} \end{pmatrix}. \tag{4.83}$$

If $p = (p_1, \dots, p_l) \in \{1, \dots, k\}^l$, let A_p denote the $m \times l$ matrix given by

$$A_p = \begin{pmatrix} a_{1p_1} & \cdots & a_{1p_l} \\ \vdots & & \vdots \\ a_{mp_1} & \cdots & a_{mp_l} \end{pmatrix}. \tag{4.84}$$

Remark 75. To simplify the notation of the above operations, a vector p with a single component is denoted by its component and a set p with a single element is denoted by its element. Hence, the element of a matrix A at the intersection of the ith row and jth column is denoted A^i_j. The ith element of a row vector ω is denoted ω_i.

4.3.3 Core Results

The constraints $\sum_{t=1}^m \omega_t \delta_{t,x_i} = 0$ on the components of an element of $\Pi_{m,X}$ have the following consequence.

Theorem 4.8. *Let $w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$. If*

$$\left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\right) \in \Pi_{m,X}, \tag{4.85}$$

and $t \in \{1, \dots, m\}$, then

$$(-1)^t \det A^{\hat{t}} = \omega_t \sum_{j=1}^m (-1)^j \det A^{\hat{j}}, \tag{4.86}$$

where $A \in \mathbb{R}^{m \times (m-1)}$ is such that $A^i_j = \delta_{i,w_j}$.

Proof. Theorem 4.8 is a special case of Theorem 4.32 (see Section 4.7). □

Theorem 4.8 will be used to express the ω_t's in terms of the δ_{t,x_i}'s:

$$\omega_t = \frac{(-1)^t \det A^{\hat{t}}}{\sum_{j=1}^m (-1)^j \det A^{\hat{j}}}, \tag{4.87}$$

if $\sum_{j=1}^m (-1)^j \det A^{\hat{j}} \neq 0$.

To formulate subsequent results and assumptions, let us define a family $\alpha^m_{\{u,v\}}$ of real-valued functions defined on R_X.

Definition 146. If $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$ for all $i, j \in \{1, \dots, m-1\}$, the function $\alpha^m_{\{u,v\}} : R_X \to \mathbb{R}$ is defined by

$$\alpha^m_{\{u,v\}}\Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) = \det B \tag{4.88}$$

where $B \in \mathbb{R}^{(m-1)\times(m-1)}$ is such that $B^i_j = q_{(u_i, v_j)}$.

Remark 76. In the above definition, each assumption $U_i \neq V_j$ ensures that $q_{(u_i,v_j)} = p_{(u_i,v_j)} - p_{u_i} p_{v_i}$ is well-defined.

Remark 77. Different choices of parameters u and v for the family $\alpha^m_{\{u,v\}}$ may lead to the same function. For example, as implied by our notation, we have $\alpha^m_{\{u,v\}} = \alpha^m_{\{v,u\}}$.

Example 62. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$ and $m = 2$, then

$$\alpha^2_{\{u,v\}}(q) = \det\Big(q_{(u_1,v_1)}\Big) = q_{(u_1,v_1)}. \tag{4.89}$$

Example 63. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$ and $m = 3$, then

$$\alpha^3_{\{u,v\}}(q) = \det\begin{pmatrix} q_{(u_1,v_1)} & q_{(u_1,v_2)} \\ q_{(u_2,v_1)} & q_{(u_2,v_2)} \end{pmatrix}. \tag{4.90}$$

Applying a function $\alpha^m_{\{u,v\}}$ to an element of $h_{m,X}(\Pi_{m,X})$ returns a noteworthy value.

Theorem 4.9. *Let $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$ for all $i, j \in \{1, \dots, m-1\}$. If*

$$q = h_{m,X}\Big(\big(\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big)_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big), \tag{4.91}$$

then

$$\alpha^m_{\{u,v\}}(q) = \Big(\prod_{j=1}^m \omega_j\Big)\Big(\sum_{j=1}^m (-1)^j \det A_u^{\hat{j}}\Big)\Big(\sum_{j=1}^m (-1)^j \det A_v^{\hat{j}}\Big), \tag{4.92}$$

where $A_u, A_v \in \mathbb{R}^{m\times(m-1)}$ are such that $(A_u)^i_j = \delta_{i,u_j}$ and $(A_v)^i_j = \delta_{i,v_j}$.

Proof. Theorem 4.9 is a special case of Theorem 4.33 (see Section 4.7). □

Theorem 4.9 has the following corollary.

Corollary 4.10. *Let $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^m$ satisfy $U_i \neq V_j$ for all $i, j \in \{1, \ldots, m\}$. If $q \in h_{m,X}(\Pi_{m,X})$, then*

$$\alpha^{(m+1)}_{\{u,v\}}(q) = 0. \tag{4.93}$$

Proof. Corollary 4.10 is a special case of Corollary 4.34 (see Section 4.7). □

Remark 78. Let $q \in h_{m,X}(\Pi_{m,X})$. The function $\alpha^{(m+1)}_{\{u,v\}}$ is applied to q in the above corollary, while $\alpha^m_{\{u,v\}}$ is applied in Theorem 4.9.

Remark 79. The expression $\alpha^m_{\{u,v\}}(\psi(p))$ is a polynomial in the components of p. Hence, Corollary 4.10 may be useful to derive a semi-algebraic description of $\mathcal{NB}_{m,X}$.

Theorem 4.11. *Let $u \in (\cup_{X_i \in X} \mathcal{X}_i)^m$ and $v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$ for all $i \in \{1, \ldots, m\}$ and all $j \in \{1, \ldots, m-1\}$. If*

$$q = h_{m,X}\Big(\big(\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big)_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \tag{4.94}$$

then, for all $t \in \{1, \ldots, m\}$,

$$\sum_{j=1}^m (-1)^j \delta_{t,u_j} \alpha^m_{\{u_{\hat{j}},v\}}(q) = 0. \tag{4.95}$$

Proof. Theorem 4.11 is a special case of Theorem 4.35 (see Section 4.7). □

If $\alpha^m_{\{u_{\hat{m}},v\}} \neq 0$, Theorem 4.11 can be used to express δ_{t,u_m} as a function of q and $\delta_{t,u_1}, \ldots, \delta_{t,u_{m-1}}$:

$$\delta_{t,u_m} = -\frac{(-1)^m}{\alpha^m_{\{u_{\hat{m}},v\}}} \sum_{j=1}^{m-1} (-1)^j \delta_{t,u_j} \alpha^m_{\{u_{\hat{j}},v\}}(q). \tag{4.96}$$

Let us define two families $\beta^m_{w,\{u,v\},p}$ and $\gamma^m_{w,\{u,v\},p}$ of functions on R_X.

Definition 147. If $w \in \cup_{X_i \in X} \mathcal{X}_i$ and $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $W \neq U_i$, $W \neq V_j$, and $U_i \neq V_j$ for all $i, j \in \{1, \ldots, m-1\}$ and if p is an integer such that $1 \leq p \leq m$, the function $\beta^m_{w,\{u,v\},p} : R_X \to \mathbb{R}$ is defined by

$$\begin{aligned} &\beta^m_{w,\{u,v\},p}\Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) \\ &\qquad = \begin{cases} \sum_{(P_1,P_2) \in P_{m,p}} \det B_{P_1,P_2} & \text{if } 1 \leq p \leq m-1, \\ 0 & \text{if } p = m, \end{cases} \end{aligned} \tag{4.97}$$

where $P_{m,p}$ is the set of pairs (P_1, P_2) such that $\{P_1, P_2\}$ is a partition of $\{1, \ldots, m-1\}$, $|P_1| = m - 1 - p$ and $|P_2| = p$, and $B_{P_1,P_2} \in \mathbb{R}^{(m-1)\times(m-1)}$ is such that

$$(B_{P_1,P_2})^i_j = \begin{cases} q_{(u_i,v_j)} & \text{if } i \in P_1, \\ q_{(w,u_i,v_j)} & \text{if } i \in P_2. \end{cases} \tag{4.98}$$

Example 64. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big)$ and $m = 2$, then

$$\beta^2_{w,\{u,v\},1}(q) = \det\Big(q_{w,u_1,v_1}\Big) \tag{4.99}$$

$$\beta^2_{w,\{u,v\},2}(q) = 0. \tag{4.100}$$

Example 65. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big)$ and $m = 3$, then

$$\beta^3_{w,\{u,v\},1}(q) = \det\begin{pmatrix} q_{(u_1,v_1)} & q_{(u_1,v_2)} \\ q_{(w,u_2,v_1)} & q_{(w,u_2,v_2)} \end{pmatrix} + \det\begin{pmatrix} q_{(w,u_1,v_1)} & q_{(w,u_1,v_2)} \\ q_{(u_2,v_1)} & q_{(u_2,v_2)} \end{pmatrix} \tag{4.101}$$

$$\beta^3_{w,\{u,v\},2}(q) = \det\begin{pmatrix} q_{(w,u_1,v_1)} & q_{(w,u_1,v_2)} \\ q_{(w,u_2,v_1)} & q_{(w,u_2,v_2)} \end{pmatrix} \tag{4.102}$$

$$\beta^3_{w,\{u,v\},3}(q) = 0. \tag{4.103}$$

Definition 148. If $w \in \cup_{X_i\in X}\mathcal{X}_i$ and $u, v \in \left(\cup_{X_i\in X}\mathcal{X}_i\right)^{m-1}$ satisfy $W \neq U_i$, $W \neq V_j$, and $U_i \neq V_j$ for all $i, j \in \{1, \ldots, m-1\}$ and if p is an integer such that $1 \leq p \leq m$, the function $\gamma^m_{w,\{u,v\},p} : R_X \to \mathbb{R}$ is defined by

$$\gamma^m_{w,\{u,v\},p}\Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big) = \begin{cases} 0 & \text{if } p = 1, \\ \sum_{(P_1,P_2,P_3)\in P'_{m,p}} \det B_{P_1,P_2,P_3} & \text{if } 2 \leq p \leq m, \end{cases} \tag{4.104}$$

where $P'_{m,p}$ is the set of triples (P_1, P_2, P_3) such that $\{P_1, P_2, P_3\}$ is a partition of $\{1, \ldots, m-1\}$, $|P_1| = m - p$, $|P_2| = p - 2$ and $|P_3| = 1$, and $B_{P_1,P_2,P_3} \in \mathbb{R}^{(m-1)\times(m-1)}$ is such that

$$(B_{P_1,P_2,P_3})^i_j = \begin{cases} q_{(u_i,v_j)} & \text{if } i \in P_1, \\ q_{(w,u_i,v_j)} & \text{if } i \in P_2, \\ q_{(w,u_i)}q_{(w,v_j)} & \text{if } i \in P_3. \end{cases} \tag{4.105}$$

Example 66. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big)$ and $m = 2$, then

$$\gamma^2_{w,\{u,v\},1}(q) = 0 \tag{4.106}$$

$$\gamma^2_{w,\{u,v\},2}(q) = \det\Big(q_{(w,u_1)}q_{(w,v_1)}\Big). \tag{4.107}$$

Example 67. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s\in\mathcal{S}})_{S\subseteq X}\Big)$ and $m = 3$, then

$$\gamma^3_{w,\{u,v\},1}(q) = 0 \tag{4.108}$$

$$\gamma^3_{w,\{u,v\},2}(q) = \det\begin{pmatrix} q_{(u_1,v_1)} & q_{(u_1,v_2)} \\ q_{(w,u_2)}q_{(w,v_1)} & q_{(w,u_2)}q_{(w,v_2)} \end{pmatrix} + \det\begin{pmatrix} q_{(w,u_1)}q_{(w,v_1)} & q_{(w,u_1)}q_{(w,v_2)} \\ q_{(u_2,v_1)} & q_{(u_2,v_2)} \end{pmatrix} \tag{4.109}$$

$$\gamma^3_{w,\{u,v\},3}(q) = \det\begin{pmatrix} q_{(w,u_1,v_1)} & q_{(w,u_1,v_2)} \\ q_{(w,u_2)}q_{(w,v_1)} & q_{(w,u_2)}q_{(w,v_2)} \end{pmatrix} + \det\begin{pmatrix} q_{(w,u_1)}q_{(w,v_1)} & q_{(w,u_1)}q_{(w,v_2)} \\ q_{(w,u_2,v_1)} & q_{(w,u_2,v_2)} \end{pmatrix}, \tag{4.110}$$

Using the families $\beta^m_{w,\{u,v\},p}$ and $\gamma^m_{w,\{u,v\},p}$, let us define a family $\nu^m_{w,\{u,v\}}$ of functions on R_X returning a polynomial with real coefficients.

Definition 149. If $w \in \cup_{X_i \in X} \mathcal{X}_i$ and $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $W \neq U_i$, $W \neq V_j$, and $U_i \neq V_j$ for all $i, j \in \{1, \dots, m-1\}$, let $\nu^m_{w,\{u,v\}}$ be the function defined on R_X that returns the polynomial in s given by

$$\nu^m_{w,\{u,v\}}(q) = s^m \alpha^m_{\{u,v\}}(q) + \sum_{p=1}^{m} s^{m-p} (\beta^m_{w,\{u,v\},p}(q) - \gamma^m_{w,\{u,v\},p}(q)). \tag{4.111}$$

Example 68. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$ and $m = 2$, then

$$\nu^2_{w,\{u,v\}}(q) = s^2 q_{u_1,v_1} + s q_{w,u_1,v_1} - q_{w,u_1} q_{w,v_1}. \tag{4.112}$$

Example 69. If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$ and $m = 3$, then

$$\begin{aligned}\nu^3_{w,\{u,v\}}(q) = {} & s^3\big(q_{(u_1,v_1)} q_{(u_2,v_2)} - q_{(u_2,v_1)} q_{(u_1,v_2)}\big) + s^2\big(q_{(u_1,v_1)} q_{(w,u_2,v_2)} - q_{(w,u_2,v_1)} q_{(u_1,v_2)} \\ & + q_{(w,u_1,v_1)} q_{(u_2,v_2)} - q_{(u_2,v_1)} q_{(w,u_1,v_2)}\big) + s\big(q_{(w,u_1,v_1)} q_{(w,u_2,v_2)} - q_{(w,u_2,v_1)} q_{(w,u_1,v_2)} \\ & - q_{(u_1,v_1)} q_{(w,u_2)} q_{(w,v_2)} + q_{(w,u_2)} q_{(w,v_1)} q_{(u_1,v_2)} - q_{(w,u_1)} q_{(w,v_1)} q_{(u_2,v_2)} \\ & + q_{(u_2,v_1)} q_{(w,u_1)} q_{(w,v_2)}\big) + \big(-q_{(w,u_1,v_1)} q_{(w,u_2)} q_{(w,v_2)} + q_{(w,u_2)} q_{(w,v_1)} q_{(w,u_1,v_2)} \\ & - q_{(w,u_1)} q_{(w,v_1)} q_{(w,u_2,v_2)} + q_{(w,u_2,v_1)} q_{(w,u_1)} q_{(w,v_2)}\big).\end{aligned} \tag{4.113}$$

The roots of the polynomial $\nu^m_{w,\{u,v\}}(q)$ are of interest.

Theorem 4.12. *Let $w \in \cup_{X_i \in X} \mathcal{X}_i$ and $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $W \neq U_i$, $W \neq V_j$, and $U_i \neq V_j$ for all $i, j \in \{1, \dots, m-1\}$. If*

$$q = h_{m,X}\Big(\big(\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big)_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big), \tag{4.114}$$

then

$$\nu^m_{w,\{u,v\}}(q) = \alpha^m_{\{u,v\}}(q) \prod_{j=1}^{m} (s + \delta_{j,w}). \tag{4.115}$$

Proof. Theorem 4.12 is a special case of Theorem 4.36 (see Section 4.7). □

By Theorem 4.12, if $q \in h_{m,X}(\Pi_{m,X})$ and $\alpha^m_{\{u,v\}}(q) \neq 0$, the set of roots of the polynomial $\nu^m_{w,\{u,v\}}(q)$ is the set $\{-\delta_{1,w}, \dots, -\delta_{m,w}\}$.

Let us define a last family $\zeta^m_{t,\{u,v\}}$ of functions on R_X.

Definition 150. If $T \subseteq X$, $t \in \mathcal{T}$ and $u, v \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \notin T$, $V_j \notin T$, and $U_i \neq V_j$ for all $i, j \in \{1, \dots, m-1\}$, the function $\zeta^m_{t,\{u,v\}} : R_X \to \mathbb{R}$ is defined by

$$\zeta^m_{t,\{u,v\}}\Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) = \sum_{p=1}^{m-1} \det B(p), \tag{4.116}$$

where $B(p) \in \mathbb{R}^{(m-1)\times(m-1)}$ is such that

$$B(p)^i_j = \begin{cases} q_{(t,u_i,v_j)} & \text{for } i = p, \\ q_{(u_i,v_j)} & \text{for } i \neq p. \end{cases} \tag{4.117}$$

The following theorem holds.

Theorem 4.13. *Let $S \subseteq X$, $s \in \mathcal{S}$ and $u, v \in (\cup_{X_i \in X}\mathcal{X}_i)^{m-1}$ satisfy $U_i \notin S$, $V_j \notin S$ and $U_i \neq V_j$ for all $i, j \in \{1, \ldots, m-1\}$. If*

$$q = h_{m,X}\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big), \tag{4.118}$$

then

$$\alpha^m_{\{u,v\}}(q)(\sum_{i=1}^m \prod_{j \in S} \delta_{i,s_{\{j\}}}) = \alpha^m_{\{u,v\}}(q)q_s + \zeta^m_{s,\{u,v\}}(q). \tag{4.119}$$

Proof. Theorem 4.13 is a special case of Theorem 4.39 (see Section 4.7). □

4.4 Computation of Fibers of h_m

This section discusses the application of the results of Section 4.3.3 to compute the preimage of a distribution $q \in h_{m,X}(\Pi_{m,X})$, and it proposes two algorithms.

4.4.1 A First Algorithm

Using Theorem 4.8 and Theorem 4.11, let us show that a parameter

$$\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \in \Pi_{m,X} \tag{4.120}$$

can be obtained from a subset of its components and $q = h_{m,X}(\pi)$ under appropriate assumptions.

Suppose that there exist $u, v, w \in (\cup_{X_i \in X}\mathcal{X}_i)^{m-1}$ such that $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \ldots, m-1\}$. This assumption ensures that the functions $\alpha^m_{\{w,v\}}$, $\alpha^m_{\{w,u\}}$, $\alpha^m_{\{(w_{\hat{j}},x_i),v\}}$ and $\alpha^m_{\{(w_{\hat{j}},x_i),u\}}$ used below are defined. Also, suppose we are given $q = h_{m,X}(\pi)$ and the components of π that are elements of the matrix

$$A = \begin{pmatrix} \delta_{1,w_1} & \cdots & \delta_{1,w_{m-1}} \\ \vdots & & \vdots \\ \delta_{m,w_1} & \cdots & \delta_{m,w_{m-1}} \end{pmatrix}. \tag{4.121}$$

By Theorem 4.8, if $\sum_{j=1}^m (-1)^j \det A^{\hat{j}} \neq 0$, then

$$\omega_t = \frac{(-1)^t \det A^{\hat{t}}}{\sum_{j=1}^m (-1)^j \det A^{\hat{j}}}. \tag{4.122}$$

If $\alpha^m_{\{w,v\}}(q) \neq 0$, $X_i \in X \setminus \cup_{i=1}^{m-1}\{V_i\}$, and $x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\}$, then

$$\delta_{t,x_i} = \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j M^t_j \alpha^m_{\{(w_{\hat{j}}, x_i), v\}}(q) \tag{4.123}$$

by Theorem 4.11. Similarly, if $\alpha^m_{\{w,u\}}(q) \neq 0$, $X_i \in \cup_{i=1}^{m-1}\{V_i\}$, and $x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\}$, then

$$\delta_{t,x_i} = \frac{(-1)^{m+1}}{\alpha^m_{\{w,u\}}(q)} \sum_{j=1}^{m-1} (-1)^j M^t_j \alpha^m_{\{(w_{\hat{j}}, x_i), u\}}(q). \tag{4.124}$$

To describe these equations and assumptions concisely, a family $f_{u,v,w}$ of functions is defined as follows.

Definition 151. If $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \dots, m-1\}$, the set $A_{u,v,w}$ is defined by

$$(M, q) \in A_{u,v,w}, \tag{4.125}$$

if, and only if, $M \in \mathbb{R}^{m \times (m-1)}$, $q \in R_X$, and

$$\alpha^m_{\{u,w\}}(q) \neq 0 \tag{4.126}$$

$$\alpha^m_{\{v,w\}}(q) \neq 0 \tag{4.127}$$

$$\sum_{j=1} (-1)^j \det M^{\hat{j}} \neq 0. \tag{4.128}$$

Definition 152. If $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \dots, m-1\}$, the function $f_{u,v,w}$ is defined on $A_{u,v,w}$ by

$$f_{u,v,w}\Big(M, q\Big) = \Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \tag{4.129}$$

where, if $q = \Big((\lambda'_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$,

$$\lambda_{x_i} = \lambda'_{x_i} \tag{4.130}$$

$$\omega_t = \frac{(-1)^t \det M^{\hat{t}}}{\sum_{j=1}^m (-1)^j \det M^{\hat{j}}} \tag{4.131}$$

$$\delta_{t,x_i} = \begin{cases} M^t_i & \text{if } x_i \in \{w_1, \dots, w_{m-1}\} \\ \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j M^t_j \alpha^m_{\{(w_{\hat{j}}, x_i), v\}}(q) & \text{if } x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\} \\ & \text{and } X_i \in X \setminus \cup_{i=1}^{m-1}\{V_i\} \\ \frac{(-1)^{m+1}}{\alpha^m_{\{w,u\}}(q)} \sum_{j=1}^{m-1} (-1)^j M^t_j \alpha^m_{\{(w_{\hat{j}}, x_i), u\}}(q) & \text{if } x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\} \\ & \text{and } X_i \in \cup_{i=1}^{m-1}\{V_i\}. \end{cases} \tag{4.132}$$

As discussed above, the following lemma holds.

Lemma 4.14. *Let $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for $i, j \in \{1, \ldots, m-1\}$. If*

$$\pi = \Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i} \Big) \in \Pi_{m,X} \tag{4.133}$$

and the $m \times (m-1)$ matrix A such that $A_j^i = \delta_{i,w_j}$ satisfy $(A, h_{m,X}(\pi)) \in A_{u,v,w}$, then

$$\pi = f_{u,v,w}(A, h_{m,X}(\pi)). \tag{4.134}$$

By Theorem 4.12, the components of a parameter $\pi \in h_{m,X}^{-1}(q)$ that are part of the matrix A given by (4.121) are constrained as follows.

Definition 153. If k is a strictly positive integer, let P_k denote the set of permutations of $\{1, \ldots, k\}$, i.e. the set of bijections from $\{1, \ldots, k\}$ to $\{1, \ldots, k\}$.

Lemma 4.15. *Let $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \ldots, m-1\}$. If*

$$q = h_{m,X}\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i} \Big) \tag{4.135}$$

satisfy $\alpha_{\{u,v\}}^m(q) \neq 0$, then there exist permutations $\sigma_1, \ldots, \sigma_{m-1} \in P_m$ such that, for all $t \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, m-1\}$, we have

$$\delta_{t,w_j} = -r_{\sigma_j(t),j}, \tag{4.136}$$

where $\{r_{1,j}, \ldots, r_{m,j}\}$ are the m roots of $v_{w_j,\{u,v\}}^m(q)$ for $j \in \{1, \ldots, m-1\}$.

Proof. By Theorem 4.12, we have

$$\{\delta_{1,w_j}, \ldots, \delta_{m,w_j}\} = \{(-r_{1,j}), \ldots, (-r_{m,j})\} \tag{4.137}$$

for $j \in \{1, \ldots, m-1\}$. Hence, there exist $m-1$ permutations $\sigma_1, \ldots, \sigma_{m-1} \in P_m$ such that

$$\delta_{t,w_j} = -r_{\sigma_j(t),j}, \tag{4.138}$$

for $t \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, m-1\}$. □

Lemma 4.14 and Lemma 4.15 readily suggest the following algorithm. Its assumptions are described using the sets $Q_{m,X}$ and $\Lambda'_{\{u,v,w\}}$.

Definition 154. If X is a finite and non-empty set of discrete random variables and m is an integer ≥ 2, the set $Q_{m,X}$ is defined by

$$\{u, v, w\} \in Q_{m,X} \tag{4.139}$$

if, and only if,

- $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$,
- $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \ldots, m-1\}$,
- $u_i \neq u_j$, $v_i \neq v_j$, and $w_i \neq w_j$ for $i \neq j$,
- there is no $X_k \in X$ such that $\mathcal{X}_k \subseteq \{u_1, \ldots, u_{m-1}\}$, $\mathcal{X}_k \subseteq \{v_1, \ldots, v_{m-1}\}$, or $\mathcal{X}_k \subseteq \{w_1, \ldots, w_{m-1}\}$.

Definition 155. If $\{u, v, w\} \in Q_{m,X}$, the set $\Lambda'_{\{u,v,w\}}$ is defined by

$$\Lambda'_{\{u,v,w\}} = \left\{ q \in R_X \middle| \alpha^m_{\{u,v\}}(q) \neq 0, \alpha^m_{\{u,w\}}(q) \neq 0, \alpha^m_{\{v,w\}}(q) \neq 0 \right\}. \tag{4.140}$$

Remark 80. In the definition of $Q_{m,X}$, the first two constraints on u, v, and w simply ensure that the functions $\alpha^m_{\{u,v\}}$ and $\nu^m_{w_i,\{u,v\}}$ with $i \in \{1, \ldots, m-1\}$ are defined. The last two constraints are necessary to have $h_{m,X}(\Pi_{m,X}) \cap \Lambda'_{\{u,v,w\}} \neq \emptyset$. Indeed, given $q = h_{m,X}(\pi)$, let A be the $m \times (m-1)$ matrix such that $A^i_j = \delta_{i,u_j}$. If $u_i = u_j$ with $i \neq j$, the ith and jth columns of A are identical. If $\mathcal{X}_k \subseteq \{u_1, \ldots, u_{m-1}\}$, the columns of A corresponding to the values of X_k sum to zero. In both cases, by Theorem 4.9, we thus have

$$\alpha^m_{\{u,v\}}(q) = \alpha^m_{\{u,w\}}(q) = 0. \tag{4.141}$$

Remark 81. By Theorem 4.9, one of the assertions $\alpha^m_{\{u,v\}}(q) \neq 0$, $\alpha^m_{\{u,w\}}(q) \neq 0$ and $\alpha^m_{\{v,w\}}(q) \neq 0$ is redundant if $q \in h_{m,X}(\Pi_{m,X})$.

The following algorithm takes for input

- u, v, w such that $\{u, v, w\} \in Q_{m,X}$,
- $q \in h_{m,X}(\Pi_{m,X}) \cap \Lambda'_{\{u,v,w\}}$,

and returns a set $S \subseteq h^{-1}_{m,X}(q)$.

Algorithm 8

```
1. For $j \in \{1, \ldots, m-1\}$, compute the $m$ roots $r_{1,j}, \ldots, r_{m,j}$ of $\nu^m_{w_j,\{u,v\}}(q)$.
2. Set $S := \emptyset$.
3. Set $\sigma_{m-1} \in P_m$ such that $\sigma_{m-1}(t) := t$ for $t \in \{1, \ldots, m\}$.
4. For each $(\sigma_1, \ldots, \sigma_{m-2}) \in (P_m)^{m-2}$,
    (a) Set $A \in \mathbb{R}^{m \times (m-1)}$ such that $A^i_j := -r_{\sigma_j(i),j}$.
    (b) If $(A, q) \in A_{u,v,w}$,
        i. Compute $\pi := f_{u,v,w}(A, q)$.
        ii. If $\pi \in \Pi_{m,X}$ and $q = h_{m,X}(\pi)$, set $S := S \cup \{\pi\}$.
5. Return $S$.
```

□

Remark 82. The roots of the polynomial $v^m_{w_j,\{u,v\}}(q)$ are counted with their multiplicity. Since its leading coefficient is $\alpha^m_{\{u,v\}}(q) \neq 0$, $v^m_{w_j,\{u,v\}}(q)$ does have m roots.

The fiber $h^{-1}_{m,X}(q)$ can be obtained easily from the output of Algorithm 8.

Theorem 4.16. *If S is the result of Algorithm 8 with inputs u, v, w, and q, then*

$$\left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in h^{-1}_{m,X}(q) \tag{4.142}$$

if, and only if, there exists $\sigma \in P_m$ such that

$$\left((\omega_{\sigma(t)}, (\delta_{\sigma(t),x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in S. \tag{4.143}$$

Proof. Let T be the set such that

$$\left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in T \tag{4.144}$$

if, and only if, there exists $\sigma \in P_m$, such that

$$\left((\omega_{\sigma(t)}, (\delta_{\sigma(t),x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in S. \tag{4.145}$$

1. Let us show that $T \subseteq h^{-1}_{m,X}(q)$. By Step 4(b)ii of Algorithm 8, we have $S \subseteq h^{-1}_{m,X}(q)$. If $\sigma \in P_m$ and

$$\pi = \left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in \Pi_{m,X}, \tag{4.146}$$

   then

$$\pi' = \left((\omega_{\sigma(t)}, (\delta_{\sigma(t),x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in \Pi_{m,X} \tag{4.147}$$

   and $h_{m,X}(\pi) = h_{m,X}(\pi')$. Hence, $S \subseteq h^{-1}_{m,X}(q)$ implies $T \subseteq h^{-1}_{m,X}(q)$.

2. Let us show that $h^{-1}_{m,X}(q) \subseteq T$. Consider

$$\pi = \left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in h^{-1}_{m,X}(q). \tag{4.148}$$

   For $j \in \{1, \ldots, m-1\}$, let $\{r_{1,j}, \ldots, r_{m,j}\}$ be the roots of $v^m_{w_j,\{u,v\}}(q)$ computed at Step 1 of Algorithm 8. By Lemma 4.15, there exist permutations $\sigma_1, \ldots, \sigma_{m-1} \in P_m$ such that

$$\delta_{t,w_j} = -r_{\sigma_j(t),j} \tag{4.149}$$

   for $t \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, m-1\}$. Let $\sigma = (\sigma_{m-1})^{-1} \in P_m$ and

$$\pi' = \left((\omega_{\sigma(t)}, (\delta_{\sigma(t),x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right). \tag{4.150}$$

   As discussed above, $\pi' \in h^{-1}_{m,X}(q)$. Let us show that $\pi' \in S$. For $j \in \{1, \ldots, m-1\}$, let

$$\sigma'_j = \sigma_j \circ \sigma \in P_m, \tag{4.151}$$

and let A be the $m \times (m-1)$ matrix such that $A^i_j = -r_{\sigma'_j(i),j}$. By (4.149), we have

$$A^i_j = \delta_{\sigma(i),w_j}. \tag{4.152}$$

By Theorem 4.9, the hypotheses $\alpha^m_{\{w,u\}}(q) \neq 0$ and $\pi' \in h^{-1}_{m,X}(q)$ imply $(A, q) \in A_{u,v,w}$. By Lemma 4.14, we have

$$\pi' = f_{u,v,w}(A, q). \tag{4.153}$$

Hence, $\pi' \in S$. To conclude, note that $\pi' \in S$ implies $\pi \in T$. Therefore, $h^{-1}_{m,X}(q) \subseteq T$. □

Remark 83. Algorithm 8 computes S instead of $h^{-1}_{m,X}(q)$ because S is smaller and contains all the information necessary to generate $h^{-1}_{m,X}(q)$.

Remark 84. The sets S and $h^{-1}_{m,X}(q)$ are finite, with $|S| \leq |(P_m)^{m-2}| = (m!)^{m-2}$ and $h^{-1}_{m,X}(q) \leq (m!)^{m-1}$.

The computational complexity of Algorithm 8 increases very quickly with the number m of hidden classes since the set $(P_m)^{m-2}$, which has $(m!)^{m-2}$ elements, is enumerated. Moreover, the computation of the polynomials and the computation of their roots may be costly. Finally, checking whether $q = h_{m,(}\pi)$ may also be costly. On the other hand, the complexity increases more slowly with the number of observable variables and their cardinalities. In particular, the complexity of the computation of $f_{u,v,w}(A, q)$ in Step 4(b)i grows linearly with $\sum_{X_i \in X} |\mathcal{X}_i|$.

4.4.2 A More Efficient Algorithm

Using Theorem 4.13, a computationaly more efficient algorithm can be defined. Its assumptions are slightly different.

Definition 156. If X is a finite and non-empty set of discrete random variables and m is an integer ≥ 2, the set $Q'_{m,X}$ is defined by

$$(\{u, v\}, w) \in Q'_{m,X} \tag{4.154}$$

if, and only if, $\{u, v, w\} \in Q_{m,X}$ and $W_{m-1} \notin \{W_1, \ldots, W_{m-2}\}$.

Remark 85. The requirement $W_{m-1} \notin \{W_1, \ldots, W_{m-2}\}$ simply ensures that the function $\zeta^m_{(w_{m-1},w_i),\{u,v\}}$ is defined for $i \in \{1, \ldots, m-2\}$.

Suppose that $(\{u, v\}, w) \in Q'_{m,X}$ and

$$q = \Big((\lambda_{x_i})_{x_i \in \cup^n_{i=1} \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big) = h_{m,X}\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup^n_{i=1} \mathcal{X}_i})^m_{t=1}, (\lambda_{x_i})_{x_i \in \cup^n_{i=1} \mathcal{X}_i}\Big). \tag{4.155}$$

For $t \in \{1, \dots, m\}$ and $j \in \{1, \dots, m-1\}$, the components δ_{t,w_j} satisfy

$$\sum_{j=1}^{m} \delta_{j,w_i} \delta_{j,w_{m-1}} = q_{\{w_i,w_{m-1}\}} + \frac{\zeta^m_{\{w_i,w_{m-1}\},\{u,v\}}(q)}{\alpha^m_{\{u,v\}}(q)} \tag{4.156}$$

by Theorem 4.13. This observation suggest the following algorithm. It takes for input

- u, v, w such that $(\{u, v\}, w) \in Q'_{m,X}$, and
- $q = \left((\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}, ((q_s)_{s \in S})_{S \subseteq X}\right) \in h_{m,X}(\Pi_{m,X}) \cap \Lambda'_{\{u,v,w\}}$,

and returns a set $S \subseteq h^{-1}_{m,X}(q)$.

Algorithm 9

```
1. For $j \in \{1, \dots, m-1\}$, compute the $m$ roots $r_{1,j}, \dots, r_{m,j}$ of $v^m_{w_j,\{u,v\}}(q)$.
2. For $i \in \{1, \dots, m-2\}$, compute the set $T_i$ of permutations $\sigma_i \in P_m$ such that
       $$\sum_{j=1}^{m} r_{\sigma_i(j),w_i} r_{j,w_{m-1}} = q_{\{w_i,w_{m-1}\}} + \frac{\zeta^m_{\{w_i,w_{m-1}\},\{u,v\}}(q)}{\alpha^m_{\{u,v\}}(q)}.  (4.157)$$
3. Set $S := \emptyset$.
4. Set $\sigma_{m-1} \in P_m$ such that $\sigma_{m-1}(t) := t$ for $t \in \{1, \dots, m\}$.
5. For each $(\sigma_1, \dots, \sigma_{m-2}) \in T_1 \times \dots \times T_{m-2}$,
   (a) Set $A \in \mathbb{R}^{m \times (m-1)}$ such that $A^i_j := -r_{\sigma_j(i),j}$.
   (b) If $(A, q) \in A_{u,v,w}$,
       i. Compute $\pi := f_{u,v,w}(A, q)$.
       ii. If $\pi \in \Pi_{m,X}$ and $q = h_{m,X}(\pi)$, set $S := S \cup \{\pi\}$.
6. Return $S$.
```
□

The hope behind this algorithm is that each set T_i will be much smaller than P_m in practice, so that Step 5 will only be performed a few times.

By (4.156), Algorithm 9 and Algorithm 8 applied to $q \in h_{m,X}(\Pi_{m,X}) \cap \Lambda'_{\{u,v,w\}}$ and u, v, w such that $(\{u, v\}, w) \in Q'_{m,X}$ return the same set S. Hence, the following theorem holds.

Theorem 4.17. *If S is the result of Algorithm 9 with inputs u, v, w, and q, then*

$$\left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in h^{-1}_{m,X}(q) \tag{4.158}$$

if, and only if, there exists $\sigma \in P_m$ such that

$$\left((\omega_{\sigma(t)}, (\delta_{\sigma(t),x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in S. \tag{4.159}$$

4.4.3 Discussion of the Assumptions

To compute the preimage $h^{-1}_{m,X}(q)$ of $q \in h_{m,X}(\Pi_{m,X})$ with Algorithm 8 (resp. Algorithm 9), it is necessary to identify u, v, w such that $\{u, v, w\} \in Q_{m,X}$ (resp. $(\{u, v\}, w) \in Q'_{m,X}$), and $q \in \Lambda'_{\{u,v,w\}}$. Let us divide the assumptions to apply our algorithms:

- it is necessary to have $Q_{m,X} \neq \emptyset$ (or $Q'_{m,X} \neq \emptyset$) and
- given $\{u, v, w\} \in Q_{m,X}$ (or $(\{u, v\}, w) \in Q'_{m,X}$), it is necessary to have $q \in \Lambda'_{\{u,v,w\}}$.

It is not straightforward to interpret the assumptions $Q_{m,X} \neq \emptyset$ and $Q'_{m,X} \neq \emptyset$. However, one can see that $Q_{m,X} = \emptyset$ if $|X| < 3$, and $Q'_{m,X} = \emptyset$ if $|X| < 4$. The following result may be helpful to understand the assumption $Q_{m,X} \neq \emptyset$.

Proposition 4.18. *There exists a partition* $\{P_1, P_2, P_3\}$ *of* X *such that*

$$\sum_{X_i \in P_j} (|\mathcal{X}_i| - 1) \geq m - 1 \tag{4.160}$$

for $j \in \{1, 2, 3\}$ *if, and only if,* $Q_{m,X} \neq \emptyset$.

Proof.

1. Consider $\{u, v, w\} \in Q_{m,X}$. If

$$P_1 = \{X_i \in X \big| \mathcal{X}_i \cap \{u_1, \dots, u_{m-1}\} \neq \emptyset\}, \tag{4.161}$$
$$P_2 = \{X_i \in X \big| \mathcal{X}_i \cap \{v_1, \dots, v_{m-1}\} \neq \emptyset\}, \tag{4.162}$$
$$P_3 = \{X_i \in X \big| \mathcal{X}_i \cap \{w_1, \dots, w_{m-1}\} \neq \emptyset\} \cup (X \setminus (P_1 \cup P_2)), \tag{4.163}$$

   then $\{P_1, P_2, P_3\}$ is a partition of X such that $\sum_{X_i \in P_j}(|\mathcal{X}_i| - 1) \geq m - 1$ for $j \in \{1, 2, 3\}$.

2. Consider a partition $\{P_1, P_2, P_3\}$ of X such that $\sum_{X_i \in P_j}(|\mathcal{X}_i| - 1) \geq m - 1$ for $j \in \{1, 2, 3\}$. For $j \in \{1, 2, 3\}$, let

$$A_j = \bigcup_{X_i \in P_j} \mathcal{X}_i \setminus \{x_i^0\}, \tag{4.164}$$

   where x_i^0 is some arbitrary value of X_i. If u (resp. v and w) is a $m - 1$-dimensional vector whose components are distinct elements of A_1 (resp. A_2 and A_3), then $\{u, v, w\} \in Q_{m,X}$. □

If Algorithm 8 can be applied to $\{u, v, w\} \in Q_{m,X}$ and $\psi(p) \in h_{m,X}(\Pi_{m,X}) \cap \Lambda'_{\{u,v,w\}}$, the preimage $f^{-1}_{m,X}(p)$ is finite. Intuitively, if the dimension of parameter space $\Theta_{m,X}$ is larger than the dimension of the space of distributions S_X, then the preimage should not be finite and it should be impossible to apply the algorithm. The following result confirms this intuition.

Proposition 4.19. *If $d(\Theta_{m,X}) > d(S_X)$, then $Q_{m,X} = \emptyset$.*

Proof. By Proposition 4.18, it is sufficient to prove that $d(S_X) - d(\Theta_{m,X}) \geq 0$ if there exists a partition $\{P_1, P_2, P_3\}$ of X such that $\sum_{X_i \in P_j}(|\mathcal{X}_i| - 1) \geq m - 1$ for $j \in \{1, 2, 3\}$.

1. First, let us show inductively that $\prod_{i=1}^{n} a_i \geq 1 + \sum_{i=1}^{n}(a_i - 1)$ if $a_i \geq 1$ for $i \in \{1, \dots, n\}$. For $n = 1$, we have
$$a_1 \geq 1 + (a_1 - 1) = a_1. \tag{4.165}$$
For $n > 1$, we have
$$\prod_{i=1}^{n} a_i = (\prod_{i=1}^{n-1} a_i)a_n \geq (1 + \sum_{i=1}^{n-1}(a_i - 1))a_n \tag{4.166}$$
by inductive hypothesis. Also, we have
$$(1 + \sum_{i=1}^{n-1}(a_i - 1))a_n = 1 + \sum_{i=1}^{n}(a_i - 1) + (a_n - 1)\sum_{i=1}^{n-1}(a_i - 1) \geq 1 + \sum_{i=1}^{n}(a_i - 1). \tag{4.167}$$
Hence, we have
$$\prod_{i=1}^{n} a_i \geq 1 + \sum_{i=1}^{n}(a_i - 1). \tag{4.168}$$
2. We have
$$d(S_X) - d(\Theta_{m,X}) = (\prod_{X_i \in X} |\mathcal{X}_i|) - m(1 + \sum_{X_i \in X}(|\mathcal{X}_i| - 1)). \tag{4.169}$$
As shown above, we have
$$(\prod_{X_i \in X} |\mathcal{X}_i|) = \prod_{j=1}^{3} \prod_{X_i \in P_j} |\mathcal{X}_i| \geq \prod_{j=1}^{3}(1 + \sum_{X_i \in P_j}(|\mathcal{X}_i| - 1)). \tag{4.170}$$
For $j \in \{1, 2, 3\}$, let $x_j = 1 + \sum_{X_i \in P_j}(|\mathcal{X}_i| - 1)$. We thus have
$$d(S_X) - d(\Theta_{m,X}) \geq x_1x_2x_3 - m(x_1 + x_2 + x_3 - 2). \tag{4.171}$$
By hypothesis, $x_j \geq m$ for $j \in \{1, 2, 3\}$. If we let $a_j = x_j - m$, we obtain
$$x_1x_2x_3 - m(x_1 + x_2 + x_3 - 2) = a_1a_2a_3 + m \\ \Big(a_1a_2 + a_1a_3 + a_2a_3 + (m - 1)(a_1 + a_2 + a_3 + (m - 2))\Big). \tag{4.172}$$
The terms of the above expression are non-negative, and we conclude that $d(S_X) - d(\Theta_{m,X}) \geq 0$. □

If $\{u, v, w\} \in Q_{m,X}$ (or $(\{u, v\}, w) \in Q'_{m,X}$) and $q \in h_{m,X}(\Pi_{m,X})$, the assumption $q \in \Lambda'_{\{u,v,w\}}$ is similar to a faithfulness assumption in the context of Bayesian network models. If a parameter π is randomly picked in $\Pi_{m,X}$, then $h_{m,X}(\pi) \in \Lambda'_{\{u,v,w\}}$ with probability one. In practice, this does not necessarily mean that the hypothesis $q \in \Lambda'_{\{u,v,w\}}$ is not important. In particular, it does not hold when $q \in h_{m-1,X}(\Pi_{m-1,X})$ by Corollary 4.10.

4.4.4 Computation of the Fibers with Two Hidden Classes

In this section, we suppose that $m = 2$ and present additional results. The assumptions made to apply Algorithm 8 are easy to interpret. Let U and V be distinct random variables, $p \in S_X$ and $q = \psi(p)$. Since $\alpha^2_{\{(u),(v)\}}(q) = q_{(u,v)} = p_{(u,v)} - p_{(u)}p_{(v)}$,

- $\alpha^2_{\{(u),(v)\}}(q) \neq 0$ for some $u \in \mathcal{U}$ and $v \in \mathcal{V}$ implies that U and V are not independent in p,
- $\alpha^2_{\{(u),(v)\}}(q) = 0$ for all $u \in \mathcal{U}$ and $v \in \mathcal{V}$ implies that U and V are independent in p.

Moreover, if

$$q = h_{2,X}\Big(\big(\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big)_{t=1}^2, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big), \tag{4.173}$$

then, by Theorem 4.8 and Theorem 4.9,

$$\alpha^2_{\{(u),(v)\}}(q) = -\delta_{1,u}\delta_{2,v} = -\delta_{1,v}\delta_{2,u}. \tag{4.174}$$

Hence, three distinct random variables that are not pairwise independent are needed to apply Algorithm 8. Let us compute the fibers of $h_{2,X}$ when three such variables do not exist. The cases where all the variables are pairwise independent and the cases where only two variables are independent are considered separately.

All the Variables Are Pairwise Independent

This case has the following simple interpretation.

Proposition 4.20. *All the variables are pairwise independent in $p \in \mathcal{NB}_{2,X}$ if, and only if, $p \in \mathcal{NB}_{1,X}$.*

Proof.

1. If $p \in \mathcal{NB}_{1,X}$, then $p \in \mathcal{NB}_{2,X}$ by Proposition 4.1, and all the variables are pairwise independent by Corollary 4.10.

2. Suppose $p \in \mathcal{NB}_{2,X}$ and all the variables are pairwise independent. Let

$$\psi(p) = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big), \tag{4.175}$$

and let

$$\pi = \Big(\big(1, (0)_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big), (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \in \Pi_{1,X}. \tag{4.176}$$

To show that $p \in \mathcal{NB}_{1,X}$, let us show that $\psi(p) = h_{1,X}(\pi)$. By Example 61, it is sufficient to show that $q_s = 0$ for $s \in \mathcal{S}$, $S \subseteq X$, and $S \neq \emptyset$. Consider

$$\Big(\big(\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\big)_{t=1}^2, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) \in h_{2,X}^{-1}(\psi(p)). \tag{4.177}$$

Since $\omega_1 > 0$, $\omega_2 > 0$, and $\omega_1\delta_{1,x_i} + \omega_2\delta_{2,x_i} = 0$ for $x_i \in \cup_{X_i \in X}\mathcal{X}_i$, we have $\delta_{1,x_i} = 0$ if, and only if, $\delta_{2,x_i} = 0$. Moreover, by (4.174), $\delta_{1,u}\delta_{2,v} = \delta_{1,v}\delta_{2,u} =$

0 for $u, v \in \cup_{X_i \in X} \mathcal{X}_i$ such that $U \neq V$. Hence, there is at most one variable $U \in X$ such that $\delta_{1,u} \neq 0$ and $\delta_{2,u} \neq 0$ for some $u \in \mathcal{U}$. For $V \in X \setminus \{U\}$, we have $\delta_{1,v}\delta_{2,v} = 0$. By definition of $h_{2,X}$, we thus have $q_s = 0$ for $s \in \mathcal{S}$, $S \subseteq X$ and $S \neq \emptyset$. □

The preimage in $\Theta_{2,X}$ of a distribution in $\mathcal{NB}_{1,X}$ is described as follows.

Proposition 4.21. *If*

$$q = \left((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\right) \in h_{1,X}(\Pi_{1,X}), \tag{4.178}$$

then

$$\pi = \left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^2, (\lambda'_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\right) \in h_{2,X}^{-1}(q) \tag{4.179}$$

if, and only if,

$$\begin{aligned} &\pi \in \Pi_{2,X} && && (4.180)\\ &\lambda'_{x_i} = \lambda_{x_i} && \text{for } x_i \in \cup_{X_i \in X} \mathcal{X}_i && (4.181)\\ &\delta_{1,x_i}\delta_{2,x_j} = 0 && \text{for } x_i, x_j \in \cup_{X_i \in X} \mathcal{X}_i \text{ such that } X_i \neq X_j. && (4.182)\end{aligned}$$

Proof.

1. Suppose that $\pi \in h_{2,X}^{-1}(q)$. Then, $\pi \in \Pi_{2,X}$ and $\lambda'_{x_i} = \lambda_{x_i}$. By Proposition 4.20, $q \in h_{1,X}(\Pi_{1,X})$ implies $\alpha^2_{\{(u),(v)\}}(q) = 0$ for $u, v \in \cup_{X_i \in X} \mathcal{X}_i$ such that $U \neq V$. By (4.174), we thus have $\alpha^2_{\{(u),(v)\}}(q) = -\delta_{1,u}\delta_{2,v} = 0$.
2. Suppose that $\pi \in \Pi_{2,X}$, $\lambda'_{x_i} = \lambda_{x_i}$ for $x_i \in \cup_{X_i \in X} \mathcal{X}_i$, and $\delta_{1,x_i}\delta_{2,x_j} = 0$ for $x_i, x_j \in \cup_{X_i \in X} \mathcal{X}_i$ such that $X_i \neq X_j$. If

$$h_{2,X}(\pi) = \left((\lambda'_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q'_s)_{s \in \mathcal{S}})_{S \subseteq X}\right), \tag{4.183}$$

let us show that $h_{2,X}(\pi) = q$. By hypothesis, $\lambda'_{x_i} = \lambda_{x_i}$. As shown in Example 61, $q \in h_{1,X}(\Pi_{1,X})$ implies $q_s = 0$ for $s \in \mathcal{S}$, $S \subseteq X$ and $S \neq \emptyset$. On the other hand, as shown in the proof of Proposition 4.20, $\delta_{1,x_i}\delta_{2,x_j} = 0$ for $x_i, x_j \in \cup_{X_i \in X} \mathcal{X}_i$ such that $X_i \neq X_j$ implies that $q'_s = 0$ for $s \in \mathcal{S}$, $S \subseteq X$, and $S \neq \emptyset$. □

Only Two Variables Are Not Independent

The preimage of a distribution in $\mathcal{NB}_{2,X}$ where only two variables are not independent is described as follows.

Proposition 4.22. *Let*

$$q = \left((\lambda'_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q'_s)_{s \in \mathcal{S}})_{S \subseteq X}\right) \in h_{2,X}(\Pi_{2,X}), \tag{4.184}$$

$$\pi = \left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^2, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\right). \tag{4.185}$$

If there exist distinct variables $U, V \in X$ such that

- *$U \not\perp V$,*
- *$X_i \perp X_j$ for $\{X_i, X_j\} \neq \{U, V\}$*

in the distribution $\psi^{-1}(q)$, then $\pi \in h^{-1}_{2,X}(q)$ if, and only if,

$$0 > \delta_{1,v_0}\delta_{2,v_0} \tag{4.186}$$

$$\lambda_{x_i} = \lambda'_{x_i} \qquad \textit{for } x_i \in \cup_{X_i \in X}\mathcal{X}_i \tag{4.187}$$

$$\omega_1 = \frac{-\delta_{2,v_0}}{-\delta_{2,v_0} + \delta_{1,v_0}} \tag{4.188}$$

$$\omega_2 = \frac{\delta_{1,v_0}}{-\delta_{2,v_0} + \delta_{1,v_0}} \tag{4.189}$$

$$\delta_{t,x_i} = 0 \qquad \textit{for } t \in \{1, 2\} \textit{ and } x_i \in \cup_{X_i \in X \setminus \{U,V\}}\mathcal{X}_i \tag{4.190}$$

$$\delta_{t,v} = \delta_{t,v_0}\frac{\alpha^2_{\{(u_0),(v)\}}(q)}{\alpha^2_{\{(u_0),(v_0)\}}(q)} > -\lambda'_v \qquad \textit{for } t \in \{1, 2\} \textit{ and } v \in \mathcal{V} \tag{4.191}$$

$$\delta_{1,u} = -\frac{\alpha^2_{\{(u),(v_0)\}}(q)}{\delta_{2,v_0}} > -\lambda'_u \qquad \textit{for } u \in \mathcal{U} \tag{4.192}$$

$$\delta_{2,u} = -\frac{\alpha^2_{\{(u),(v_0)\}}(q)}{\delta_{1,v_0}} > -\lambda'_u \qquad \textit{for } u \in \mathcal{U}, \tag{4.193}$$

where $u_0 \in \mathcal{U}$ and $v_0 \in \mathcal{V}$ are such that $\alpha^2_{\{(u_0),(v_0)\}}(q) \neq 0$.

Proof.

1. Suppose that $\pi \in h^{-1}_{2,X}(q)$. Let us show that (4.186) to (4.193) hold. First, $\pi \in \Pi_{2,X}$ implies $\delta_{t,x_i} > -\lambda_{x_i} = -\lambda'_{x_i}$ for $x_i \in \cup_{X_i \in X}\mathcal{X}_i$ and $t \in \{1, 2\}$.

   (a) By (4.174), $\alpha^2_{\{(u_0),(v_0)\}}(q) \neq 0$ implies that $\delta_{1,v_0} \neq 0$ and $\delta_{2,v_0} \neq 0$. Since $\omega_1\delta_{1,v_0} + \omega_2\delta_{2,v_0} = 0$, $\omega_1 > 0$ and $\omega_2 > 0$, we have $\delta_{1,v_0}\delta_{2,v_0} = -\frac{\omega_2}{\omega_1}(\delta_{2,v_0})^2$. Hence, (4.186) holds.

   (b) By definition of $h_{2,X}$, (4.187) holds.

   (c) Since $\omega_1\delta_{1,v_0} + \omega_2\delta_{2,v_0} = 0$, $\delta_{1,v_0} = \delta_{2,v_0}$ would imply that $\omega_1 + \omega_2 = 0$. This contradicts $\omega_1 + \omega_2 = 1$, and thus $\delta_{1,v_0} \neq \delta_{2,v_0}$. By Theorem 4.8, (4.188) and (4.189) thus hold.

   (d) By (4.174), (4.190) to (4.193) hold.

2. Suppose that (4.186) to (4.193) hold. It is straighforward to see that $\pi \in \Pi_{2,X}$. Let us show that $h_{2,X}(\pi) = q$. Let

$$h_{2,X}(\pi) = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big). \tag{4.194}$$

   By (4.187), $\lambda_{x_i} = \lambda'_{x_i}$. By (4.190), $q_s = 0$ for $S \neq \emptyset, \{U, V\}$ and $s \in \mathcal{S}$. As shown in the first part of the proof, (4.190) holds for a parameter in $h^{-1}_{2,X}(q)$, and thus $q'_s = 0$ for $S \neq \emptyset, \{U, V\}$ and $s \in \mathcal{S}$. For $S = \{U, V\}$, we have

$$q'_{(u,v)} = \frac{\alpha^2_{\{(u_0),(v)\}}(q)\alpha^2_{\{(u),(v_0)\}}(q)}{\alpha^2_{\{(u_0),(v_0)\}}(q)}. \tag{4.195}$$

By (4.174), we thus have $q'_{(u,v)} = \alpha^2_{\{(u),(v)\}}(q) = q_{(u,v)}$. □

4.4.5 Extensions

Using Algorithm 8 or Algorithm 9, it may be possible to compute fibers of the parametrization maps of other classes of discrete Bayesian network models with hidden variables. This section presents two examples, leaving their generalization and in-depth analysis for future work.

A First Example

Consider a set $\{H, X_1, \dots, X_6\}$ of discrete random variables, and consider the BN model $\mathcal{M}_H = f_H(\Theta)$ with hidden variable H and observable variables $\{X_1, \dots, X_6\}$ obtained from the discrete Bayesian network model with structure given in Figure 4.5. If $p = f_H(\theta) \in \mathcal{M}_H$, then

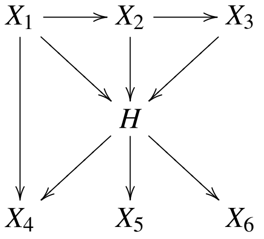


Figure 4.5: A Bayesian network structure over $\{H, X_1, \dots, X_6\}$

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \sum_{h \in \mathcal{H}} \theta^{X_1}_{x_1} \theta^{X_2,x_1}_{x_2} \theta^{X_3,x_2}_{x_3} \theta^{H,(x_1,x_2,x_3)}_{h} \theta^{X_4,(x_1,h)}_{x_4} \theta^{X_5,h}_{x_5} \theta^{X_6,h}_{x_6}. \tag{4.196}$$

It is straightforward to see that

$$p(x_1, x_2, x_3) = \theta^{X_1}_{x_1} \theta^{X_2,x_1}_{x_2} \theta^{X_3,x_2}_{x_3}, \tag{4.197}$$

$$p(x_4, x_5, x_6 | x_1, x_2, x_3) = \sum_{h \in \mathcal{H}} \theta^{H,(x_1,x_2,x_3)}_{h} \theta^{X_4,(x_1,h)}_{x_4} \theta^{X_5,h}_{x_5} \theta^{X_6,h}_{x_6}. \tag{4.198}$$

The marginal distribution $p(x_1, x_2, x_3)$ is an element of the discrete Bayesian network model for $\{X_1, X_2, X_3\}$ with structure $X_1 \to X_2 \to X_3$. Hence, the values of the parameters $\theta^{X_1}_{x_1}$, $\theta^{X_2,x_1}_{x_2}$ and $\theta^{X_3,x_2}_{x_3}$ can be obtained by (1.33). The conditional distribution $p(x_4, x_5, x_6 | x_1, x_2, x_3)$ is an element of $\mathcal{NB}_{|\mathcal{H}|,\{X_4,X_5,X_6\}}$, and it may be possible to apply the theory developped in this chapter.

A Discrete HLC Model

Consider the HLC model $\mathcal{M}_H = f_H(\Theta)$ presented in Example 21. If $p = f_H(\theta) \in \mathcal{M}_H$, then we have by marginalization

$$p(x_1, x_2, x_3) = \sum_{h_2 \in \mathcal{H}_2} (\sum_{h_1 \in \mathcal{H}_1} \theta_{h_2}^{H_2,h_1}) \theta_{x_1}^{X_1,h_2} \theta_{x_2}^{X_2,h_2} \theta_{x_3}^{X_3,h_2}, \tag{4.199}$$

$$p(x_i, x_4, x_j) = \sum_{h_1 \in \mathcal{H}_1} \theta_{h_1}^{H_1} \theta_{x_4}^{X_4,h_1} (\sum_{h_2 \in \mathcal{H}_2} \theta_{h_2}^{H_2,h_1} \theta_{x_i}^{X_i,h_2})(\sum_{h_3 \in \mathcal{H}_3} \theta_{h_3}^{H_3,h_1} \theta_{x_j}^{X_j,h_3}), \tag{4.200}$$

$$p(x_5, x_6, x_7) = \sum_{h_3 \in \mathcal{H}_3} (\sum_{h_1 \in \mathcal{H}_1} \theta_{h_3}^{H_3,h_1}) \theta_{x_5}^{X_5,h_3} \theta_{x_6}^{X_6,h_3} \theta_{x_7}^{X_7,h_3} \tag{4.201}$$

for $i \in \{1, 2, 3\}$ and $j \in \{5, 6, 7\}$. Hence, we have

$$p(x_1, x_2, x_3) \in \mathcal{NB}_{|\mathcal{H}_2|,\{X_1,X_2,X_3\}}, \tag{4.202}$$

$$p(x_i, x_4, x_j) \in \mathcal{NB}_{|\mathcal{H}_1|,\{X_i,X_4,X_j\}}, \tag{4.203}$$

$$p(x_5, x_6, x_7) \in \mathcal{NB}_{|\mathcal{H}_3|,\{X_5,X_6,X_7\}}. \tag{4.204}$$

If we have

$$\{u, v, w\} \in Q_{|\mathcal{H}_2|,\{X_1,X_2,X_3\}}, \qquad p(x_1, x_2, x_3) \in \Lambda'_{\{u,v,w\}}, \tag{4.205}$$

$$\{a, b, c\} \in Q_{|\mathcal{H}_1|,\{X_i,X_4,X_j\}}, \qquad p(x_i, x_4, x_j) \in \Lambda'_{\{a,b,c\}}, \tag{4.206}$$

$$\{d, e, f\} \in Q_{|\mathcal{H}_3|,\{X_5,X_6,X_7\}}, \qquad p(x_5, x_6, x_7) \in \Lambda'_{\{d,e,f\}}, \tag{4.207}$$

then it is possible to compute with Algorithm 8 the components $\theta_{x_1}^{X_1,h_2}$, $\theta_{x_2}^{X_2,h_2}$, $\theta_{x_3}^{X_3,h_2}$, $\theta_{x_4}^{X_4,h_1}$, $\theta_{h_1}^{H_1}$, $\theta_{x_5}^{X_5,h_3}$, $\theta_{x_6}^{X_6,h_3}$, and $\theta_{x_7}^{X_7,h_3}$ of the parameter θ and

$$\theta_{x_i}^{X_i,h_1} = \sum_{h_2 \in \mathcal{H}_2} \theta_{h_2}^{H_2,h_1} \theta_{x_i}^{X_i,h_2}, \tag{4.208}$$

$$\theta_{x_j}^{X_j,h_1} = \sum_{h_3 \in \mathcal{H}_3} \theta_{h_3}^{H_3,h_1} \theta_{x_j}^{X_j,h_3}. \tag{4.209}$$

Let us show that no additional assumption is needed to compute the remaining components $\theta_{h_2}^{H_2,h_1}$ and $\theta_{h_3}^{H_3,h_1}$. For $k \in \{1, \ldots, |\mathcal{H}_2| - 1\}$, let $\delta_{w_k}^{W_k,h_2} = \theta_{w_k}^{W_k,h_2} - p(w_k)$. Then, (4.208) implies that

$$\sum_{h_2 \in \mathcal{H}_2} \theta_{h_2}^{H_2,h_1} \delta_{w_k}^{W_k,h_2} = \theta_{w_k}^{W_k,h_1} - p(w_k). \tag{4.210}$$

To express this relation in matrix form, label the values of H_1 as $h_{1,1}, \ldots, h_{1,|\mathcal{H}_1|}$ and the values of H_2 as $h_{2,1}, \ldots, h_{2,|\mathcal{H}_2|}$, and let A, B and C be the matrices such that

$$A \in \mathbb{R}^{|\mathcal{H}_1| \times (|\mathcal{H}_2|-1)}, \qquad A_j^i = \theta_{h_{2,j}}^{H_2,h_{1,i}} \tag{4.211}$$

$$B \in \mathbb{R}^{(|\mathcal{H}_2|-1) \times (|\mathcal{H}_2|-1)}, \qquad B_k^j = \delta_{w_k}^{W_k,h_{2,j}} \tag{4.212}$$

$$C \in \mathbb{R}^{|\mathcal{H}_1| \times (|\mathcal{H}_2|-1)}, \qquad C_k^i = \theta_{w_k}^{W_k,h_{1,i}} - p(w_k). \tag{4.213}$$

Then, (4.210) is equivalent to $AB = C$. By Theorem 4.8 and Theorem 4.9, the assumption $p(x_1, x_2, x_3) \in \Lambda'_{\{u,v,w\}}$ implies $\det B \neq 0$. Hence, we have $A = CB^{-1}$. Moreover, we have

$$\theta^{H_2,h_1}_{h_{2,|\mathcal{H}_2|}} = 1 - \sum_{j=1}^{|\mathcal{H}_2|-1} \theta^{H_2,h_1}_{h_{2,j}}. \tag{4.214}$$

Similarly, it is straighforward to obtain $\theta^{H_3,h_1}_{h_3}$.

4.5 Projections for Parameter Learning

In this section, Algorithm 8 and Algorithm 9 are modified so that each returns a single parameter and leads to a continuous projection when the input distribution is sufficiently close to $h_{m,X}(\Pi_{m,X})$. As will be shown, it is sufficient to adapt the parts of the algorithms where we test for equality and to keep the real part of the roots.

4.5.1 Projections Based on Algorithm 8

The following algorithm is adapted from Algorithm 8. It takes for input

- u, v, w such that $\{u, v, w\} \in Q_{m,X}$,
- $q \in \Lambda'_{\{u,v,w\}}$,

and returns a parameter $\pi \in \Pi_{m,X}$ or $\emptyset$.

Algorithm 10

```
1. For $j \in \{1, \dots, m-1\}$, compute the real parts $r_{1,j}, \dots, r_{m,j}$ of the $m$ roots of $v^m_{w_j,\{u,v\}}(q)$.
2. Set $S := \emptyset$.
3. Set $\sigma_{m-1} \in P_m$ such that $\sigma_{m-1}(t) := t$ for $t \in \{1, \dots, m\}$.
4. For each $(\sigma_1, \dots, \sigma_{m-2}) \in (P_m)^{m-2}$,
    (a) Set $A \in \mathbb{R}^{m\times(m-1)}$ such that $A^i_j := -r_{\sigma_j(i),j}$.
    (b) If $(A, q) \in A_{u,v,w}$,
        i. Compute $\pi := f_{u,v,w}(A, q)$.
        ii. If $\pi \in \Pi_{m,X}$, set $S := S \cup \{\pi\}$.
5. If $S = \emptyset$, return $\emptyset$. Otherwise, return an element of the set
        $\arg\min_{\pi\in S} D(\psi^{-1}(q) \parallel \psi^{-1}(h_{m,X}(\pi))).$    (4.215)
```

□

Remark 86. In Algorithm 8, the preimage $h^{-1}_{m,X}(q)$ is contained in S, and equality between q and $h_{m,X}(\pi)$, $\pi \in S$ is tested explicitely. In Algorithm 10, elements of S should be interpreted as candidate projections of q. Instead of testing for equality, the parameter in S that minimizes the KL distance to q is selected.

Remark 87. Algorithm 10 is non-deterministic because an order for the roots of the polynomials $v^m_{w_j,\{u,v\}}(q)$ and an element in $\arg\min_{\pi\in S} D(\psi^{-1}(q) \parallel \psi^{-1}(h_{m,X}(\pi)))$ are chosen implicitely.

Using Algorithm 10, a family of projections can be derived. To formally define functions based on the output of the algorithm, its non-determinism must be eliminated. However, the properties of the functions defined do not depend on the particular choice and it can be done implicitely.

Definition 157. If u, v, w satisfy $\{u, v, w\} \in Q_{m,X}$, the set $\Lambda_{u,v,w}$ is the set of elements $q \in \Lambda'_{\{u,v,w\}}$ such that Algorithm 10 applied to u, v, w and q returns a parameter $\pi \in \Pi$ (and not $\emptyset$).

Remark 88. In other words, $\Lambda_{u,v,w}$ is the set of elements $q \in \Lambda'_{\{u,v,w\}}$ such that there exist permutations $\sigma_1, \ldots, \sigma_{m-1} \in P_m$ such that

- σ_{m-1} is the identity,
- $(A, q) \in A_{u,v,w}$, and
- $f_{u,v,w}(A, q) \in \Pi_{m,X}$

where A is the $m \times (m-1)$ matrix such that $A^i_j = -r_{\sigma_j(i),w_j}$ and $r_{1,w_j}, \ldots, r_{m,w_j}$ are the real parts of the m roots of $v^m_{w_j,\{u,v\}}(q)$ for $j \in \{1, \ldots, m-1\}$.

Remark 89. It is easy to see that $\Lambda_{u,v,w} \cap h_{m,X}(\Pi_{m,X}) = \Lambda'_{\{u,v,w\}} \cap h_{m,X}(\Pi_{m,X})$.

Definition 158. If u, v, w satisfy $\{u, v, w\} \in Q_{m,X}$, the function $\pi_{u,v,w} : \Lambda_{u,v,w} \to \Pi_{m,X}$ associates to $q \in \Lambda_{u,v,w}$ the parameter obtained by applying Algorithm 10 to u, v, w and q.

Let us check that the constraints on $\pi_{u,v,w}$ introduced in Section 4.1 hold.

Proposition 4.23. *The set $\Lambda_{u,v,w}$ is open.*

The proof of Proposition 4.23 uses the following intermediate result.

Definition 159. The set $\Upsilon_{m,X}$ is defined by

$$\left((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n X_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n X_i}\right) \in \Upsilon_{m,X} \tag{4.216}$$

if, and only if,

$$\sum_{t=1}^{m} \omega_t = 1, \qquad \sum_{t=1}^{m} \omega_t \delta_{t,x_i} = 0, \tag{4.217}$$

$$\sum_{x_i \in \mathcal{X}_i} \lambda_{x_i} = 1, \qquad \sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = 0 \tag{4.218}$$

for $t \in \{1, \dots, m\}$, $X_i \in X$ and $x_i \in \mathcal{X}_i$.

Lemma 4.24. *Let $u, v, w \in (\cup_{X_i \in X} \mathcal{X}_i)^{m-1}$ satisfy $U_i \neq V_j$, $U_i \neq W_j$, and $V_i \neq W_j$ for all $i, j \in \{1, \dots, m-1\}$. If $(M, q) \in A_{u,v,w}$, then $f_{u,v,w}(M, q) \in \Upsilon_{m,X}$.*

Proof. Let us check that

$$\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^{n} \mathcal{X}_i})_{t=1}^{m}, (\lambda_{x_i})_{x_i \in \cup_{i=1}^{n} \mathcal{X}_i} \Big) = f_{u,v,w}(M, q) \in \Upsilon_{m,X}. \tag{4.219}$$

1. Trivially, we have $\sum_{t=1}^{m} \omega_t = 1$.
2. Trivially, we have $\sum_{x_i \in \mathcal{X}_i} \lambda_{x_i} = 1$.
3. Let us show that $\sum_{t=1}^{m} \omega_t \delta_{t,x_i} = 0$ holds.

   (a) If $x_i \in \{w_1, \dots, w_{m-1}\}$, we have

   $$\sum_{t=1}^{m} \omega_t \delta_{t,x_i} = \sum_{t=1}^{m} \frac{(-1)^t \det M^{\hat{t}}}{\sum_{j=1}^{m} (-1)^j \det M^{\hat{j}}} M_i^t = \frac{\det M'}{\sum_{j=1}^{m} (-1)^j \det M^{\hat{j}}} \tag{4.220}$$

   where M' is the $m \times m$ matrix such that $M'_{(2,\dots,m)} = M$ and $M'_1 = M_i$. We have $\det M' = 0$ since the first and ith column of M' are identical.

   (b) If $X_i \in X \setminus \cup_{i=1}^{m-1} \{V_i\}$ and $x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\}$, we have

   $$\sum_{t=1}^{m} \omega_t \delta_{t,x_i} = \sum_{t=1}^{m} \omega_t \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j \delta_{t,w_j} \alpha^m_{\{(w_{\hat{j}}, x_i), v\}}(q) \tag{4.221}$$

   $$= \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j \Big(\sum_{t=1}^{m} \omega_t \delta_{t,w_j} \Big) \alpha^m_{\{(w_{\hat{j}}, x_i), v\}}(q) \tag{4.222}$$

   $$= 0. \tag{4.223}$$

   (c) The case $X_i \in \cup_{i=1}^{m-1} \{V_i\}$ and $x_i \in \mathcal{X}_i \setminus \{w_1, \dots, w_{m-1}\}$ is similar to the previous case, and one can see that $\sum_{t=1}^{m} \omega_t \delta_{t,x_i} = 0$.

4. Let us show that $\sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = 0$ holds. For $i, j \in \{1, \dots, m-1\}$, we have

$$\alpha^m_{\{(w_{\hat{j}}, w_i), u\}}(q) = \begin{cases} 0 & \text{if } i \neq j, \\ (-1)^{(m-1)-i} \alpha^m_{\{w,u\}}(q) & \text{if } i = j, \end{cases} \tag{4.224}$$

$$\alpha^m_{\{(w_{\hat{j}}, w_i), v\}}(q) = \begin{cases} 0 & \text{if } i \neq j, \\ (-1)^{(m-1)-i} \alpha^m_{\{w,v\}}(q) & \text{if } i = j. \end{cases} \tag{4.225}$$

Hence, we have

$$\delta_{w_i} = \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j \delta_{w_j} \alpha^m_{\{(w_{\hat{j}},w_i),v\}}(q), \tag{4.226}$$

$$= \frac{(-1)^{m+1}}{\alpha^m_{\{w,u\}}(q)} \sum_{j=1}^{m-1} (-1)^j \delta_{w_j} \alpha^m_{\{(w_{\hat{j}},w_i),u\}}(q). \tag{4.227}$$

(a) For $X_i \in X \setminus \cup_{i=1}^{m-1}\{V_i\}$, we thus have

$$\sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = \sum_{x_i \in \mathcal{X}_i} \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j \delta_{t,w_j} \alpha^m_{\{(w_{\hat{j}},x_i),v\}}(q) \tag{4.228}$$

$$= \frac{(-1)^{m+1}}{\alpha^m_{\{w,v\}}(q)} \sum_{j=1}^{m-1} (-1)^j \delta_{t,w_j} \sum_{x_i \in \mathcal{X}_i} \alpha^m_{\{(w_{\hat{j}},x_i),v\}}(q). \tag{4.229}$$

If $q = \Big((\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}, ((q_s)_{s \in \mathcal{S}})_{S \subseteq X}\Big)$, then

$$\alpha^m_{\{(w_{\hat{j}},x_i),v\}}(q) = \det \begin{pmatrix} q_{(w_1,v_1)} & \cdots & q_{(w_1,v_{m-1}} \\ \vdots & & \vdots \\ q_{(w_{j-1},v_1)} & \cdots & q_{(w_{j-1},v_{m-1})} \\ q_{(w_{j+1},v_1)} & \cdots & q_{(w_{j+1},v_{m-1})} \\ \vdots & & \vdots \\ q_{(w_{m-1},v_1)} & \cdots & q_{(w_{m-1},v_{m-1})} \\ q_{(x_i,v_1)} & \cdots & q_{(x_i,v_{m-1})} \end{pmatrix}, \tag{4.230}$$

and thus

$$\sum_{x_i \in \mathcal{X}_i} \alpha^m_{\{(w_{\hat{j}},x_i),v\}}(q) = \det \begin{pmatrix} q_{(w_1,v_1)} & \cdots & q_{(w_1,v_{m-1}} \\ \vdots & & \vdots \\ q_{(w_{j-1},v_1)} & \cdots & q_{(w_{j-1},v_{m-1})} \\ q_{(w_{j+1},v_1)} & \cdots & q_{(w_{j+1},v_{m-1})} \\ \vdots & & \vdots \\ q_{(w_{m-1},v_1)} & \cdots & q_{(w_{m-1},v_{m-1})} \\ \sum_{x_i \in \mathcal{X}_i} q_{(x_i,v_1)} & \cdots & \sum_{x_i \in \mathcal{X}_i} q_{(x_i,v_{m-1})} \end{pmatrix} \tag{4.231}$$

The last line of the above matrix contains only zeros. Its determinant is thus zero, and we conclude that $\sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = 0$.

(b) The case $X_i \in \cup_{i=1}^{m-1}\{V_i\}$ is similar to the previous case, and one can see that $\sum_{x_i \in \mathcal{X}_i} \delta_{t,x_i} = 0$. □

Proof (Proposition 4.23). Recall that $\Lambda_{u,v,w}$ is open if, and only if,

$$(\forall q \in \Lambda_{u,v,w})(\exists \delta > 0 \text{ s.t. } (r \in R_X \text{ and } |q-r| < \delta) \Rightarrow r \in \Lambda_{u,v,w}). \tag{4.232}$$

Also, recall that a distribution $x \in \Lambda_{u,v,w}$ if, and only if, $x \in \Lambda'_{\{u,v,w\}}$ and Algorithm 10 applied to u, v, w and x return $S \neq \emptyset$. Consider $q \in \Lambda_{u,v,w}$.

1. Since $q \in \Lambda'_{\{u,v,w\}}$ and $\Lambda'_{\{u,v,w\}}$ is open, we have

$$\exists \delta > 0 \text{ s.t. } (r \in R_X \text{ and } |q-r| < \delta) \Rightarrow r \in \Lambda'_{\{u,v,w\}}. \tag{4.233}$$

2. Let us show that

$$\exists \delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q-r| < \delta) \Rightarrow r \in \Lambda_{u,v,w}. \tag{4.234}$$

For $x \in \Lambda'_{\{u,v,w\}}$ and $j \in \{1,\dots,m-1\}$, let $u^x_{1,w_j},\dots,u^x_{m,w_j}$ be the real parts of the m roots of $v^m_{w_j,\{u,v\}}(x)$ (in some fixed order). For $x \in \Lambda'_{\{u,v,w\}}$ and $\sigma = (\sigma_1,\dots,\sigma_{m-1}) \in (P_m)^{m-1}$, let $M(x,\sigma)$ be the $m \times (m-1)$ matrix such that $M(x,\sigma)^i_j = u^x_{\sigma_j(i),w_j}$. If $S_q \neq \emptyset$ is the result of Algorithm 10 applied to u, v, w and q, there exists $\sigma^q = (\sigma^q_1,\dots,\sigma^q_{m-1}) \in (P_m)^{m-1}$ such that σ^q_{m-1} is the identity, $(M(q,\sigma^q), q) \in A_{u,v,w}$ and

$$\Big((\omega_t, (\delta_{t,x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i})_{t=1}^m, (\lambda_{x_i})_{x_i \in \cup_{i=1}^n \mathcal{X}_i}\Big) = f_{u,v,w}(M(q,\sigma^q), q) \in S_q, \tag{4.235}$$

with $\delta_{i,w_j} = -u^q_{\sigma^q_j(i),w_j}$ for $j \in \{1,\dots,m-1\}$.

(a) By continuity of the coefficients of $v^m_{w,\{u,v\}}$, continuity of the roots of a polynomial and continuity of the function taking the real part of a complex number, we have

$$\begin{aligned}(\forall \epsilon > 0)\Big(&\exists \delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q-r| < \delta) \Rightarrow \\ &(\exists (\sigma_1,\dots,\sigma_{m-1}) \in (P_m)^{m-1} \text{ s.t. } |u^r_{\sigma_j(i),w_j} - u^q_{i,w_j}| < \epsilon)\Big).\end{aligned} \tag{4.236}$$

Letting $\sigma^r_j = \sigma^{-1}_{m-1} \circ \sigma^q_j \circ \sigma_j$ for $j \in \{1,\dots,m-1\}$, we obtain

$$\begin{aligned}(\forall \epsilon > 0)\Big(&\exists \delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q-r| < \delta) \Rightarrow \\ &(\exists (\sigma^r_1,\dots,\sigma^r_{m-1}) \in (P_m)^{m-1} \text{ s.t. } |u^r_{\sigma^r_j(i),w_j} - u^q_{(\sigma^{-1}_{m-1} \circ \sigma^q_j)(i),w_j}| < \epsilon \\ &\qquad \text{and } \sigma^r_{m-1} \text{ is the identity})\Big).\end{aligned} \tag{4.237}$$

(b) Since $(M(q,\sigma^q), q) \in A_{u,v,w}$, we have $(M(q, (\sigma \circ \sigma^q_1, \dots, \sigma \circ \sigma^q_{m-1})), q) \in A_{u,v,w}$ for any permutation $\sigma \in P_m$, in particular σ^{-1}_{m-1}. By (4.237) and because $A_{u,v,w}$ is open, we thus have

$$\begin{aligned}&\exists \delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q-r| < \delta) \Rightarrow \\ &\quad \Big(\exists (\sigma^r_1,\dots,\sigma^r_{m-1}) \in (P_m)^{m-1} \text{ s.t. } (M(r,\sigma^r), r) \in A_{u,v,w} \\ &\qquad \text{and } \sigma^r_{m-1} \text{ is the identity}\Big).\end{aligned} \tag{4.238}$$

(c) By continuity of $f_{u,v,w}$, (4.237) and (4.238), we have

$$\begin{aligned}&(\forall\epsilon > 0)\Big(\exists\delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q - r| < \delta) \Rightarrow\\ &\quad\Big(\exists(\sigma_1^r, \ldots, \sigma_{m-1}^r) \in (P_m)^{m-1} \text{ s.t. } (M(r, \sigma^r), r) \in A_{u,v,w},\\ &\quad\Big|\pi - f_{u,v,w}(M(r, \sigma^r), r)\Big| < \epsilon, \text{ and } \sigma_{m-1}^r \text{ is the identity}\Big)\Big), \qquad (4.239)\end{aligned}$$

where

$$\pi = f_{u,v,w}(M(q, (\sigma \circ \sigma_1^q, \ldots, \sigma \circ \sigma_{m-1}^q)), q) \in \Pi_{m,X}. \qquad (4.240)$$

(Note that $\pi \in h_{m,X}^{-1}(q)$ if $q \in h_{m,X}(\Pi_{m,X}) \cap \Lambda_{u,v,w}$.) By Lemma 4.24 and because $\Pi_{m,X}$ is open in the topology induced by $\Upsilon_{m,X}$, we thus have

$$\begin{aligned}&\exists\delta > 0 \text{ s.t. } (r \in \Lambda'_{\{u,v,w\}} \text{ and } |q - r| < \delta) \Rightarrow\\ &\quad\Big(\exists(\sigma_1^r, \ldots, \sigma_{m-1}^r) \in (P_m)^{m-1} \text{ s.t. } (M(r, \sigma^r), r) \in A_{u,v,w},\\ &\quad f_{u,v,w}(M(r, \sigma^r), r) \in \Pi_{m,X}, \text{ and } \sigma_{m-1}^r \text{ is the identity}\Big), \qquad (4.241)\end{aligned}$$

By (4.233) and (4.241), we have

$$\exists\delta > 0 \text{ s.t. } (r \in R_X \text{ and } |q - r| < \delta) \Rightarrow r \in \Lambda_{u,v,w}. \qquad (4.242)$$

□

Corollary 4.25. *The set $h_{m,X}(\Pi_{m,X}) \cap \Lambda_{u,v,w}$ is included in the interior of $\Lambda_{u,v,w}$.*

Proposition 4.26. *If $q \in h_{m,X}(\Pi_{m,X}) \cap \Lambda_{u,v,w}$, then $\pi_{u,v,w}(q) \in h_{m,X}^{-1}(q)$.*

Proof. The proof of Proposition 4.26 is very similar to the proof of Theorem 4.16. It is easy to see that the set S obtained after Step 4 of Algorithm 10 satisfies $S \cap h_{m,X}^{-1}(q) \neq \emptyset$. Hence, $\min_{\pi\in S} D(\psi^{-1}(q) \,\|\, \psi^{-1}(h_{m,X}(\pi))) = 0$ and

$$\arg\min_{\pi\in S} D(\psi^{-1}(q) \,\|\, \psi^{-1}(h_{m,X}(\pi))) = S \cap h_{m,X}^{-1}(q). \qquad (4.243)$$

□

Proposition 4.27. *The function $h_{m,X} \circ \pi_{u,v,w}$ is continuous on $\Lambda_{u,v,w} \cap h_{m,X}(\Pi_{m,X})$.*

Proof. Recall that $h_{m,X} \circ \pi_{u,v,w}$ is continuous on $\Lambda_{u,v,w} \cap h_{m,X}(\Pi_{m,X})$ if, and only if,

$$\begin{aligned}&(\forall q \in \Lambda_{u,v,w} \cap h_{m,X}(\Pi_{m,X}))(\forall\epsilon > 0)\Big(\exists\delta > 0 \text{ s.t. } (r \in \Lambda_{u,v,w} \text{ and } |r - q| < \delta)\\ &\qquad \Rightarrow (|h_{m,X} \circ \pi_{u,v,w}(r) - h_{m,X} \circ \pi_{u,v,w}(q)| < \epsilon)\Big). \qquad (4.244)\end{aligned}$$

Consider $q \in \Lambda_{u,v,w} \cap h_{m,X}(\Pi_{m,X})$.

1. By Proposition 4.26, it is sufficient to show that

$$(\forall \epsilon > 0)\big(\exists \delta > 0 \text{ s.t. } (r \in \Lambda_{u,v,w} \text{ and } |r - q| < \delta) \Rightarrow |h_{m,X} \circ \pi_{u,v,w}(r) - q| < \epsilon\big). \tag{4.245}$$

2. By Pinsker's inequality (see (C.3) in Appendix C), we have

$$(\forall p, p' \in S_X)\big(D(p \,\|\, p') < \delta \Rightarrow (\forall x \in \mathcal{X})(|p_x - p'_x| < \sqrt{2 \ln 2\delta})\big). \tag{4.246}$$

By continuity of the $L2$-norm and continuity of ψ, we thus have

$$(\forall q, q' \in R_X)(\forall \epsilon > 0)\big(\exists \delta > 0 \text{ s.t. } D(\psi^{-1}(q) \,\|\, \psi^{-1}(q')) < \delta \Rightarrow |q - q'| < \epsilon\big). \tag{4.247}$$

To prove the proposition, it is thus sufficient to show that

$$(\forall \epsilon > 0)\big(\exists \delta > 0 \text{ s.t. } (r \in \Lambda_{u,v,w} \text{ and } |r - q| < \delta) \Rightarrow D(\psi^{-1}(q) \,\|\, \psi^{-1}(h_{m,X}(\pi_{u,v,w}(r)))) < \epsilon\big). \tag{4.248}$$

3. By definition of Algorithm 10, it is sufficient to show that

$$(\forall \epsilon > 0)\big(\exists \delta > 0 \text{ s.t. } (r \in \Lambda_{u,v,w} \text{ and } |r - q| < \delta) \Rightarrow (\exists \sigma = (\sigma_1, \dots, \sigma_{m-1}) \in (P_m)^{m-1} \text{ s.t. } (M(r, \sigma), r) \in A_{u,v,w}, f_{u,v,w}(M(r, \sigma), r) \in \Pi_{m,X}, \sigma_{m-1} \text{ is the identity, and } D(\psi^{-1}(q) \,\|\, \psi^{-1}(h_{m,X}(f_{u,v,w}(A(r, \sigma), r)))) < \epsilon)\big) \tag{4.249}$$

where the matrix $M(r, \sigma)$ is defined in the proof of Proposition 4.23.

4. By continuity of $h_{m,X}$, ψ^{-1}, and D, it is sufficient to show that

$$(\forall \epsilon > 0)\big(\exists \delta > 0 \text{ s.t. } (r \in \Lambda_{u,v,w} \text{ and } |r - q| < \delta) \Rightarrow (\exists \pi \in h^{-1}_{m,X}(q) \text{ and } \exists \sigma = (\sigma_1, \dots, \sigma_{m-1}) \in (P_m)^{m-1} \text{ s.t. } (M(r, \sigma), r) \in A_{u,v,w}, f_{u,v,w}(M(r, \sigma), r) \in \Pi_{m,X}, \sigma_{m-1} \text{ is the identity, and } |\pi - f_{u,v,w}(M(r, \sigma), r)| < \epsilon)\big). \tag{4.250}$$

This assertion is a consequence of (4.239) and (4.240). □

Remark 90. The domain of definition $\Lambda_{u,v,w}$ of our projection function may be enlarged by slightly modifying Algorithm 10 as follows. Consider $\pi = f_{u,v,w}(A, q)$ obtained at Step 4(b)i. At the next step, instead of rejecting π if it does not belong to $\Pi_{m,X}$, one could project it onto $\Pi_{m,X}$ while minimizing the euclidian distance. It is straighforward to see that Proposition 4.23, Proposition 4.26 and Proposition 4.27 still hold with this modification. Hence, the asymptotic properties of the algorithm are preserved. Although such an euclidian projection may not make sense in terms of the distributions represented, this variant should be explored and tested in future research.

4.5.2 Projections Based on Algorithm 9

The following algorithm is adapted from Algorithm 9. It takes for input

- $\alpha > 0$,
- u, v, w such that $(\{u, v\}, w) \in Q'_{m,X}$, and
- $q \in \Lambda'_{\{u,v,w\}}$,

and returns a parameter $\pi \in \Pi_{m,X}$ or $\emptyset$.

Algorithm 11

1. For $j \in \{1, \ldots, m-1\}$, compute the real parts $r_{1,j}, \ldots, r_{m,j}$ of the m roots of $\nu^m_{w_j,\{u,v\}}(q)$.
2. For $i \in \{1, \ldots, m-2\}$, compute the set T_i of permutations $\sigma_i \in P_m$ such that
$$\Big|\sum_{j=1}^{m} r_{\sigma_i(j),w_i} r_{j,w_{m-1}} - q_{\{w_i,w_{m-1}\}} - \frac{\zeta^m_{\{w_i,w_{m-1}\},\{u,v\}}(q)}{\alpha^m_{\{u,v\}}(q)}\Big| < \alpha. \tag{4.251}$$
3. Set $S := \emptyset$.
4. Set $\sigma_{m-1} \in P_m$ such that $\sigma_{m-1}(t) := t$ for $t \in \{1, \ldots, m\}$.
5. For each $(\sigma_1, \ldots, \sigma_{m-2}) \in T_1 \times \cdots \times T_{m-2}$,
    - (a) Set $A \in \mathbb{R}^{m \times (m-1)}$ such that $A^i_j := -r_{\sigma_j(i),j}$.
    - (b) If $(A, q) \in A_{u,v,w}$,
        - i. Compute $\pi := f_{u,v,w}(A, q)$.
        - ii. If $\pi \in \Pi_{m,X}$, set $S := S \cup \{\pi\}$.
6. If $S = \emptyset$, return $\emptyset$. Otherwise, return an element of the set
$$\arg\min_{\pi \in S} D(\psi^{-1}(q) \parallel \psi^{-1}(h_{m,X}(\pi))). \tag{4.252}$$

□

Remark 91. The parameter α of Algorithm 11 influences the size of each set T_i. It is easy to see that $T_i \subseteq T'_i$ for $\alpha \leq \alpha'$ and $T_i = P_m$ for α sufficiently large.

Using Algorithm 11, a family of projections is defined. As before, the non-determinism of the algorithm is implicitely eliminated.

Definition 160. If u, v, w satisfy $(\{u, v\}, w) \in Q'_{m,X}$ and $\alpha > 0$, the set $\Lambda_{u,v,w,\alpha}$ is the set of elements $q \in \Lambda'_{\{u,v,w\}}$ such that Algorithm 11 applied to α, u, v, w and q returns a parameter $\pi \in \Pi$ (and not $\emptyset$).

Remark 92. We have $\Lambda_{u,v,w,\alpha} \cap h_{m,X}(\Pi_{m,X}) = \Lambda'_{\{u,v,w\}} \cap h_{m,X}(\Pi_{m,X})$.

Definition 161. If u, v, w satisfy $(\{u, v\}, w) \in Q'_{m,X}$ and $\alpha > 0$, the function $\pi_{u,v,w,\alpha} : \Lambda_{u,v,w,\alpha} \to \Pi_{m,X}$ associates to $q \in \Lambda_{u,v,w,\alpha}$ the parameter obtained by applying Algorithm 11 to α, u, v, w and q.

Let us check that the constraints on $\pi_{u,v,w,\alpha}$ introduced in Section 4.1 hold.

Proposition 4.28. *The set $\Lambda_{u,v,w,\alpha}$ is open.*

Proof. Consider $q \in \Lambda_{u,v,w,\alpha}$. It is sufficient to show that there exists $\delta > 0$ such that $r \in R_X$ and $|r - q| < \delta$ imply $r \in \Lambda'_{\{u,v,w\}}$ and there exists $\sigma = (\sigma_1, \ldots, \sigma_{m-1}) \in (P_m)^{m-1}$ such that

1. σ_{m-1} is the identity,
2. $\left| \sum_{j=1}^{m} u_{\sigma_i(j),w_i} u_{j,w_{m-1}} - q_{\{w_i,w_{m-1}\}} - \frac{\zeta^m_{\{w_i,w_{m-1}\},\{u,v\}}(r)}{\alpha^m_{\{u,v\}}(r)} \right| < \alpha,$
3. $(A(r, \sigma), r) \in A_{u,v,w},$
4. $f_{u,v,w}(A(r, \sigma), r) \in \Pi_{m,X}$

where $u_{1,w_j}, \ldots, u_{m,w_j}$ are the real parts of roots of $v^m_{w_j,\{u,v\}}(r)$ for $j \in \{1, \ldots, m-1\}$ and $A(r, \sigma)$ is the $m \times (m-1)$ matrix such that $A^i_j = -u_{\sigma_j(i),w_j}$. The proof is similar to the proof of Proposition 4.23, the only difference is related to the second constraint. If we observe that the expression

$$\sum_{j=1}^{m} u_{\sigma_i(j),w_i} u_{j,w_{m-1}} - r_{\{w_i,w_{m-1}\}} - \frac{\zeta^m_{\{w_i,w_{m-1}\},\{u,v\}}(r)}{\alpha^m_{\{u,v\}}(r)} \tag{4.253}$$

is continuous with respect to $u_{\sigma_i(j),w_i}$, $u_{j,w_{m-1}}$ and r, it is straightforward to transpose the proof of Proposition 4.23 to this case. □

Corollary 4.29. *The set $h_{m,X}(\Pi_{m,X}) \cap \Lambda_{u,v,w,\alpha}$ is included in the interior of $\Lambda_{u,v,w,\alpha}$.*

Proposition 4.30. *If $q \in h_{m,X}(\Pi_{m,X}) \cap \Lambda_{u,v,w,\alpha}$, then $\pi_{u,v,w,\alpha}(q) \in h^{-1}_{m,X}(q)$.*

Proof. It is straightforward to transpose the proof of Proposition 4.26 to this case. □

Proposition 4.31. *The function $h_{m,X} \circ \pi_{u,v,w,\alpha}$ is continuous on $\Lambda_{u,v,w,\alpha} \cap h_{m,X}(\Pi_{m,X})$.*

Proof. It is straightforward to transpose the proof of Proposition 4.27 to this case. □

The parameter α influences the applicability of $\pi_{u,v,w,\alpha}$ and the computational complexity of its evaluation with Algorithm 11. By Remark 91,

- $\Lambda_{u,v,w,\alpha} \subseteq \Lambda_{u,v,w,\alpha'}$ for $\alpha' \geq \alpha$, and $\Lambda_{u,v,w,\alpha} = \Lambda_{u,v,w}$ for sufficiently large α;
- the computational complexity increases with α until $T_1 = \cdots = T_{m-2} = P_m$.

For small values, it is likely that α controls a trade-off between the sample complexity and the computational complexity of the parameter learning procedure using $\pi_{u,v,w,\alpha}$.

4.6 Conclusion

The contributions of this chapter can be summarized as follows. First, an *alternative parametrization* of discrete Naive Bayes models with hidden class variable is provided. Using this new parametrization, two *families of functions* $\alpha^m_{\{u,v\}}$ *and* $\nu^m_{w,\{u,v\}}$ describing probability distributions are introduced and *series of results at the heart of our algorithms* is derived. Second, these developments are applied to design *two algorithms that compute the set of parameters mapped to a given Naive Bayes distribution* under certain technical assumptions formulated using the family $\alpha^m_{\{u,v\}}$. This generalization of the case of two hidden classes and binary observable variables presented in [Pea88] and [GHKM01] is not trivial. Then, promising research directions extending our work to other classes of discrete Bayesian network models with hidden variables are proposed. Finally, our two algorithms are converted into *two projection algorithms suitable for parameter learning*. The resulting parameter learning procedures are *asymptotically correct* in the following sense: if the distribution generating the observations belongs to the discrete Naive Bayes model under consideration and satisfies other technical assumptions, then, with probability one in the limit of a large sample size, the learned parameter is mapped to a distribution converging towards the generating distribution.

The content of this chapter is purely theoretical, and much research remains to be done. Our parameter learning algorithms should be extensively tested and compared to other methods such as the E.M. algorithm. We anticipate that the following points will need to be addressed. First, guidance for the choice of parameters u, v, w (and α) should be provided. Naturally, different parameter settings can be compared by the likelihood of the solutions obtained. However, the set of admissible parameters may be difficult to enumerate, prohibitively large, and contain many elements that result in the same solution. For guidance a priori, a better interpretation of the hypothesis $q \in \Lambda'_{\{u,v,w\}}$ may be helpful, in particular to determine what it entails and understand the algorithms behavior should it not hold. Also, a method to test the validity of the hypothesis using the observations only is clearly of interest. To that end, we suspect that the projection algorithms return an empty answer for a sufficiently large number of observations generated by a distribution violating the hypothesis. Under the appropriate hypotheses, our parameter learning algorithms are asymptotically correct. While this property is remarkable, the behavior for a finite sample size and the convergence speed should be studied. Note that the convergence speed probably depends on u, v, w (and α), and is thus relevant to their choice. Also, note that a value $\alpha^m_{\{u,v\}}(q)$ or $\nu^m_{w,\{u,v\}}(q)$ only depend on marginal distributions of $p = \psi^{-1}(q)$ of at most three random variables. Second, the computational complexity increases rapidly with the number m of hidden classes. The cause of this problem is that the preimage of a distribution is not obtained completely analytically, but some enumeration is involved. An interesting and closely related question is whether the non-injectivity of the parametrization map is only due to aliasing at distributions p such that $\psi(p) \in \Lambda'_{\{u,v,w\}}$.

Besides parameter learning, our results may have other potential applications that were not pursued here. First, they may be useful to derive implicit descriptions of discrete Naive Bayes models, in particular Corollary 4.10 (see Remark 79) and Theorem 4.12. For the latter result, recall that there exists constraints on the coefficients of a polynomial ensuring that its roots are real. Moreover, if we have an explicit description of the roots as functions of the coefficients, then some sign constraints of the parameter space can be transposed. Second, our results may be useful to study the geometric properties of $\mathcal{NB}_{m,X}$, in particular its dimension. Recall that $f^{-1}_{m,X}(p)$ is finite if there exists $\{u, v, w\} \in Q_{m,X}$ such that $\psi(p) \in \Lambda'_{\{u,v,w\}}$. Also, $Q_{m,X} \neq \emptyset$ implies $d(\Theta_{m,X}) \leq d(S_X)$. If $Q_{m,X} \neq \emptyset$, the questions of whether the dimension of $\mathcal{NB}_{m,X}$ is maximal, i.e. $d(\mathcal{NB}_{m,X}) = d(\Theta_{m,X})$, and whether the intersection of $\mathcal{NB}_{m,X}$ with some set $\Lambda'_{\{u,v,w\}}$ is a $d(\Theta_{m,X})$-dimensional manifold embedded in $\mathbb{R}^{|X|}$ seem interesting to us. Third, our results may be useful to estimate the number m of hidden classes of a Naive Bayes distribution. Suppose that $p \in \mathcal{NB}_{m,X}$, $Q_{m',X} \neq \emptyset$ for some $m' > m$, and we attempt to project $\hat{p}$ onto $\mathcal{NB}_{n,X}$. For a sufficiently large sample size, we suspect that Algorithm 10 will return an empty answer for $m < n \leq m'$ and possibly for some values $n \in \{1, \ldots, m-1\}$. If Algorithm 10 does return a parameter for n in some subset of $\{1, \ldots, m\}$, we suspect that the parameter maximizing the data likelihood corresponds to the case $n = m$. Finally, our results are also relevant to the computation of the marginal likelihood of discrete Naive Bayes models. In particular, if the preimage of the distribution in $\mathcal{NB}_{m,X}$ maximizing the likelihood is finite, Rusakov and Geiger propose to simply approximate the marginal likelihood by the BIC score (see [RG03] and Section 2.5.1). Our results allows us to identify situations where the preimage is finite.

4.7 Proofs of the Core Results

To generalize the results of Section 4.3.3 and dispose of unnecessary notations and hypotheses, let us define a function $g_{m,k}$.

Definition 162. The set $\Sigma_{m,k}$ is defined by $(\omega, A) \in \Sigma_{m,k}$ if, and only if,

$$\omega \in \mathbb{R}^{1\times m}, \qquad \sum_{t=1}^{m} \omega_t = 1, \tag{4.254}$$

$$A \in \mathbb{R}^{m\times k}, \qquad \omega A = 0. \tag{4.255}$$

Definition 163. The function $g_{m,k} : \Sigma_{m,k} \to \mathbb{R}^{2^k}$ is defined by

$$g_{m,k}(\omega, A) = (q_i)_{i\in 2^{\{1,\ldots,k\}}} \tag{4.256}$$

where

$$q_i = \sum_{t=1}^{m} \omega_t \prod_{j\in i} A^t_j. \tag{4.257}$$

The set $\Sigma_{m,k}$ and the function $g_{m,k}$ are similar to respectively $\Pi_{m,X}$ and $h_{m,X}$.

4.7.1 Theorem 4.8

Theorem 4.8 is a special case of the following result.

Theorem 4.32. *If $(\omega, A) \in \Sigma_{m,m-1}$ and $t \in \{1, \ldots, m\}$, we have*

$$(-1)^t \det A^{\hat{t}} = \omega_t \sum_{j=1}^{m} (-1)^j \det A^{\hat{j}}. \tag{4.258}$$

Proof. To prove the theorem, let us show that

$$\omega_{\hat{t}} A^{\hat{t}} \mathcal{A}(A^{\hat{t}})^T \mathbf{1} = \det A^{\hat{t}} (1 - \omega_t) = (-1)^t \omega_t \sum_{\substack{j=1 \\ j \neq t}}^{m} (-1)^j \det A^{\hat{j}} \tag{4.259}$$

where $\mathcal{A}(A^{\hat{t}})$ is the matrix of the algebraic minors of $A^{\hat{t}}$ and $\mathbf{1}$ is the $(m-1)$-dimensional column vector of ones.

1. By associativity, we have

$$\omega_{\hat{t}} A^{\hat{t}} \mathcal{A}(A^{\hat{t}})^T \mathbf{1} = \omega_{\hat{t}} (A^{\hat{t}} \mathcal{A}(A^{\hat{t}})^T) \mathbf{1} = \det A^{\hat{t}} \omega_{\hat{t}} 1 = \det A^{\hat{t}} (1 - \omega_t). \tag{4.260}$$

2. By associativity and because $\omega A = 0$, we have

$$\omega_{\hat{t}} A^{\hat{t}} \mathcal{A}(A^{\hat{t}})^T \mathbf{1} = \Big((\omega_{\hat{t}} A^{\hat{t}}) \mathcal{A}(A^{\hat{t}})^T \Big) \mathbf{1} \tag{4.261}$$

$$= \sum_{j=1}^{m-1} \sum_{l=1}^{m-1} \Big(\mathcal{A}(A^{\hat{t}})^T \Big)_j^l \sum_{\substack{p=1 \\ p \neq t}}^{m} \omega_p A_l^p, \tag{4.262}$$

$$= -\omega_t \sum_{j=1}^{m-1} \sum_{l=1}^{m-1} A_l^t \Big(\mathcal{A}(A^{\hat{t}}) \Big)_l^j. \tag{4.263}$$

For $j < t$, we have

$$\sum_{l=1}^{m-1} A_l^t \Big(\mathcal{A}(A^{\hat{t}}) \Big)_l^j = \sum_{l=1}^{m-1} A_l^t (-1)^{j+l} \det A_{\hat{l}}^{\{\hat{j},t\}} = (-1)^{t+j-1} \det A^{\hat{j}}. \tag{4.264}$$

For $j \geq t$, we have

$$\sum_{l=1}^{m-1} A_l^t \Big(\mathcal{A}(A^{\hat{t}}) \Big)_l^j = \sum_{l=1}^{m-1} A_l^t (-1)^{j+l} \det A_{\hat{l}}^{\{\widehat{j+1},t\}} = (-1)^{t+j} \det A^{\widehat{j+1}}. \tag{4.265}$$

Hence, we have

$$\omega_{\hat{t}} A^{\hat{t}} \mathcal{A}(A^{\hat{t}})^T \mathbf{1} = (-1)^t \omega_t \sum_{\substack{j=1 \\ j \neq t}}^{m} (-1)^j \det A^{\hat{j}}. \tag{4.266}$$

□

Remark 93. Theorem 4.32 can be formulated using the notion of vector product: if $(\omega, A) \in \Sigma_{m,m-1}$, then

$$A_1 \wedge \cdots \wedge A_{m-1} = \omega \sum_{t=1}^{m} (A_1 \wedge \cdots \wedge A_{m-1})_t. \tag{4.267}$$

4.7.2 Theorem 4.9 and Corollary 4.10

The following family of functions generalizes $\alpha^m_{\{u,v\}}$.

Definition 164. If $u, v \in \{1, \ldots, k\}^{m-1}$ satisfy $u_i \neq v_j$ for all $i, j \in \{1, \ldots, m-1\}$, the function $\alpha'^m_{\{u,v\}} : \mathbb{R}^{2^k} \to \mathbb{R}$ is defined by

$$\alpha'^m_{\{u,v\}}((q_r)_{r \in 2^{\{1,\ldots,k\}}}) = \det B \tag{4.268}$$

where B is the $(m-1) \times (m-1)$ matrix such that $B^i_j = q_{\{u_i, v_j\}}$.

Theorem 4.9 is a special case of the following result.

Theorem 4.33. *Let $u, v \in \{1, \ldots, k\}^{m-1}$ satisfy $u_i \neq v_j$ for all $i, j \in \{1, \ldots, m-1\}$. If $x = g_{m,k}(\omega, A)$, then*

$$\alpha'^m_{\{u,v\}}(x) = \Big(\prod_{j=1}^{m} \omega_j\Big)\Big(\sum_{j=1}^{m} (-1)^j \det A^{\hat{j}}_u\Big)\Big(\sum_{j=1}^{m} (-1)^j \det A^{\hat{j}}_v\Big). \tag{4.269}$$

To prove Theorem 4.33, we use the following notation.

Definition 165. If $\sigma \in P_k$, let $P(\sigma)$ denote the matrix obtained by permuting the rows of the $k \times k$ identity matrix I according to σ, i.e.

$$P(\sigma)^i_j = I^{\sigma(i)}_j. \tag{4.270}$$

Proof (Theorem 4.33). Suppose that $x = (q_r)_{r \in 2^{\{1,\ldots,k\}}}$. By definition of $\alpha'^m_{\{u,v\}}$ and $g_{m,k}$, we have

$$\alpha'^m_{\{u,v\}}(x) = \sum_{\sigma \in P_{m-1}} \det P(\sigma) \prod_{j=1}^{m-1} q_{\{u_j, v_{\sigma(j)}\}} = \sum_{\sigma \in P_{m-1}} \det P(\sigma) \prod_{j=1}^{m-1} \sum_{t_j=1}^{m} \omega_{t_j} A^{t_j}_{u_j} A^{t_j}_{v_{\sigma(j)}}. \tag{4.271}$$

Reorganizing the terms, we have

$$\begin{aligned} \alpha'^m_{\{u,v\}}(x) &= \sum_{1 \le t_1, \ldots, t_{m-1} \le m} \Big(\sum_{\sigma \in P_{m-1}} \det P(\sigma) \prod_{j=1}^{m-1} A^{t_j}_{v_{\sigma(j)}}\Big)\Big(\prod_{j=1}^{m-1} \omega_{t_j} A^{t_j}_{u_j}\Big) \\ &= \sum_{1 \le t_1, \ldots, t_{m-1} \le m} \det A^{(t_1, \ldots, t_{m-1})}_v \prod_{j=1}^{m-1} \omega_{t_j} A^{t_j}_{u_j}. \end{aligned} \tag{4.272}$$

Because $\det A_v^{(t_1,\dots,t_{m-1})} = 0$ if $t_i = t_j$ for some $i \neq j$, the range of the sum in (4.272) can be restricted to distinct t_i, $i = 1, \dots, m-1$. Hence, we suppose that the t_i are distinct. Let t_m be the only element in $\{1, \dots, m\} \setminus \{t_1, \dots, t_{m-1}\}$, and let $\sigma \in P_m$ be such that $\sigma(i) = t_i$. We have

$$A_v^{(\sigma(1),\dots,\sigma(m-1))} = P\,(\sigma)^{\hat{m}}_{\sigma(\hat{m})} A_v^{\sigma(\hat{m})}. \tag{4.273}$$

Hence, we have

$$\det A_v^{(\sigma(1),\dots,\sigma(m-1))} = (-1)^{m+\sigma(m)} \det P\,(\sigma) \det A_v^{\sigma(\hat{m})}. \tag{4.274}$$

and, by Theorem 4.32,

$$\det A_v^{(\sigma(1),\dots,\sigma(m-1))} = (-1)^m \omega_{\sigma(m)} \det P\,(\sigma) \sum_{j=1}^{m} (-1)^j \det A_v^{\hat{j}}. \tag{4.275}$$

Inserting this result in (4.272), we see that $\alpha'^m_{\{u,v\}}(x)$ is equal to

$$\Big(\prod_{j=1}^{m} \omega_j\Big)\Big(\sum_{j=1}^{m} (-1)^j \det A_v^{\hat{j}}\Big)\Big(\sum_{\sigma \in P_m} (-1)^m \det P\,(\sigma) \prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)}\Big). \tag{4.276}$$

Consider the last factor of this product. The sum over P_m can be decomposed so that the factor is equal to

$$\sum_{i=1}^{m} \sum_{\sigma' \in P_{m-1}} (-1)^m \det P\,(\sigma) \prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)}, \tag{4.277}$$

where $\sigma \in P_m$ is defined by

$$\sigma(j) = \begin{cases} \sigma'(j) & \text{for } j < m \text{ and } \sigma'(j) < i \\ \sigma'(j) + 1 & \text{for } j < m \text{ and } \sigma'(j) \geq i \\ i & \text{for } j = m. \end{cases} \tag{4.278}$$

For $j \in \{1, \dots, m-1\}$, we thus have

$$A_{u_j}^{\sigma(j)} = (A^{\hat{i}})_{u_j}^{\sigma'(j)} = (A_u^{\hat{i}})_j^{\sigma'(j)}. \tag{4.279}$$

We have

$$P\,(\sigma') = P\,(\sigma)^{\hat{m}}_{\hat{i}}, \tag{4.280}$$

and thus we have

$$\det P\,(\sigma) = (-1)^{m+i} \det P\,(\sigma'). \tag{4.281}$$

Therefore, we have

$$\begin{aligned} \sum_{\sigma \in P_m} (-1)^m \det P\,(\sigma) \prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)} &= \sum_{i=1}^{m} (-1)^i \sum_{\sigma' \in P_{m-1}} \det P\,(\sigma') \prod_{j=1}^{m-1} (A_u^{\hat{i}})_j^{\sigma'(j)} \\ &= \sum_{i=1}^{m} (-1)^i \det A_u^{\hat{i}}, \end{aligned} \tag{4.282}$$

and we can conclude the proof. □

Corollary 4.10 is a special case of the following result.

Corollary 4.34. *Let $u, v \in \{1, \dots, k\}^m$ satisfy $u_i \neq v_j$ for all $i, j \in \{1, \dots, m\}$. If $x = g_{m,k}(\omega, A)$, then*

$$\alpha'^{(m+1)}_{\{u,v\}}(x) = 0. \tag{4.283}$$

Proof. There exists $(\omega', A') \in \Sigma_{(m+1),k}$ such that

$$(\omega')_{\{m,\hat{m}+1\}} = \omega_{\hat{m}} \tag{4.284}$$

$$\omega'_m + \omega'_{m+1} = \omega_m \tag{4.285}$$

$$(A')^{\widehat{m+1}} = A \tag{4.286}$$

$$(A')^{m+1} = A^m. \tag{4.287}$$

Moreover, it is easy to see that $x = g_{m,k}(\omega, A) = g_{(m+1),k}(\omega', A')$. By Theorem 4.33, $\alpha'^{(m+1)}_{\{u,v\}} x$ is equal to

$$\Big(\prod_{j=1}^{m+1} \omega'_j\Big)\Big(\sum_{j=1}^{m+1} (-1)^j \det{(A')}^{\hat{j}}_u\Big)\Big(\sum_{j=1}^{m+1} (-1)^j \det{(A')}^{\hat{j}}_v\Big). \tag{4.288}$$

To conclude the proof, let us show that $\det{(A')}^{\hat{j}}_u = 0$ for all $j \in \{1, \dots, m+1\}$.

1. If $j < m$, the last two rows of $(A')^{\hat{j}}_u$ are identical.

2. If $j \in \{m, m+1\}$, we have $\omega (A')^{\hat{j}}_u = \omega A_u = 0$ with $\omega \neq 0$ since $\sum_{t=1}^m \omega_t = 1$.□

4.7.3 Theorem 4.11

Theorem 4.11 is a special case of the following result.

Theorem 4.35. *Let $u \in \{1, \dots, k\}^m$ and $v \in \{1, \dots, k\}^{m-1}$ satisfy $u_i \neq v_j$ for all $i \in \{1, \dots, m\}$ and all $j \in \{1, \dots, m-1\}$. If $x = g_{m,k}(\omega, A)$, then, for all $t \in \{1, \dots, m\}$,*

$$\sum_{j=1}^{m} (-1)^j A^t_{u_j} \alpha'^m_{\{u_{\hat{j}},v\}}(x) = 0. \tag{4.289}$$

Proof. By Theorem 4.33, we have

$$\sum_{j=1}^{m} (-1)^j A^t_{u_j} \alpha'^m_{\{u_{\hat{j}},v\}}(x) = \Big(\prod_{p=1}^{m} \omega_p\Big)\Big(\sum_{p=1}^{m} (-1)^p \det A^{\hat{p}}_v\Big)\Big(\sum_{j=1}^{m} (-1)^j A^t_{u_j} \sum_{p=1}^{m} (-1)^p \det A^{\hat{p}}_{u^{\hat{j}}}\Big). \tag{4.290}$$

By Theorem 4.32, we thus have

$$\sum_{j=1}^{m}(-1)^j A^t_{u_j}\alpha'^m_{\{u_{\hat{j}},v\}}(x) = \Big(\prod_{\substack{p=1\\p\neq t}}^{m}\omega_p\Big)\Big(\sum_{p=1}^{m}(-1)^p \det A^{\hat{p}}_v\Big)\Big(\sum_{j=1}^{m}(-1)^j A^t_{u_j}(-1)^t \det A^{\hat{t}}_{u^{\hat{j}}}\Big) \tag{4.291}$$

$$= \Big(\prod_{\substack{p=1\\p\neq t}}^{m}\omega_p\Big)\Big(\sum_{p=1}^{m}(-1)^p \det A^{\hat{p}}_v\Big)\det A_u. \tag{4.292}$$

We have $\omega A_u = 0$ and $\omega \neq 0$ because $\sum_{t=1}^m \omega_t = 1$. Hence, $\det A_u = 0$. □

4.7.4 Theorem 4.12

The following families of functions generalize $\beta^m_{w,\{u,v\},p}$, $\gamma^m_{w,\{u,v\},p}$ and $\nu^m_{w,\{u,v\}}$.

Definition 166. If $w \in \{1,\dots,k\}$ and $u, v \in \{1,\dots,k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$, and $u_i \neq v_j$ for all $i, j \in \{1,\dots,m-1\}$ and if p is an integer such that $1 \leq p \leq m$, the function $\beta'^m_{w,\{u,v\},p} : \mathbb{R}^{2^k} \to \mathbb{R}$ is defined by

$$\beta'^m_{w,\{u,v\},p}\big((q_r)_{r\in 2^{\{1,\dots,k\}}}\big) = \begin{cases} \sum_{(P_1,P_2)\in P_{m,p}} \det B_{P_1,P_2} & \text{if } 1 \leq p \leq m-1 \\ 0 & \text{if } p = m \end{cases} \tag{4.293}$$

where $P_{m,p}$ is the set of pairs (P_1, P_2) such that $\{P_1, P_2\}$ is a partition of $\{1,\dots,m-1\}$, $|P_1| = m-1-p$, and $|P_2| = p$ and B_{P_1,P_2} is the $(m-1)\times(m-1)$ matrix such that

$$(B_{P_1,P_2})^i_j = \begin{cases} q_{\{u_i,v_j\}} & \text{if } i \in P_1, \\ q_{\{w,u_i,v_j\}} & \text{if } i \in P_2. \end{cases} \tag{4.294}$$

Definition 167. If $w \in \{1,\dots,k\}$ and $u, v \in \{1,\dots,k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$, and $u_i \neq v_j$ for all $i, j \in \{1,\dots,m-1\}$ and if p is an integer such that $1 \leq p \leq m$, the function $\gamma'^m_{w,\{u,v\},p} : \mathbb{R}^{2^k} \to \mathbb{R}$ is defined by

$$\gamma'^m_{w,\{u,v\},p}\big((q_r)_{r\in 2^{\{1,\dots,k\}}}\big) = \begin{cases} 0 & \text{if } p = 1, \\ \sum_{(P_1,P_2,P_3)\in P'_{m,p}} \det B_{P_1,P_2,P_3} & \text{if } 2 \leq p \leq m, \end{cases} \tag{4.295}$$

where $P'_{m,p}$ is the set of triples (P_1, P_2, P_3) such that $\{P_1, P_2, P_3\}$ is a partition of $\{1,\dots,m-1\}$, $|P_1| = m-p$, $|P_2| = p-2$ and $|P_3| = 1$, and B_{P_1,P_2,P_3} is the $(m-1)\times(m-1)$ matrix such that

$$(B_{P_1,P_2,P_3})^i_j = \begin{cases} q_{\{u_i,v_j\}} & \text{if } i \in P_1, \\ q_{\{w,u_i,v_j\}} & \text{if } i \in P_2, \\ q_{\{w,u_i\}}q_{\{w,v_j\}} & \text{if } i \in P_3. \end{cases} \tag{4.296}$$

Definition 168. If $w \in \{1, \dots, k\}$ and $u, v \in \{1, \dots, k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$, and $u_i \neq v_j$ for all $i, j \in \{1, \dots, m-1\}$, let $\nu'^m_{w,\{u,v\}}$ be the function defined on $\mathbb{R}^{2^k}$ that returns the polynomial in s with real coefficients given by

$$\nu'^m_{w,\{u,v\}}(q) = s^m \alpha'^m_{\{u,v\}}(q) + \sum_{p=1}^{m} s^{m-p}(\beta'^m_{w,\{u,v\},p}(q) - \gamma'^m_{w,\{u,v\},p}(q)). \tag{4.297}$$

Theorem 4.12 is a special case of the following result.

Theorem 4.36. *Let $w \in \{1, \dots, k\}$, and $u, v \in \{1, \dots, k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$ and $u_i \neq v_j$ for all $i, j \in \{1, \dots, m-1\}$. If $x = g_{m,k}(\omega, A)$, then*

$$\nu'^m_{w,\{u,v\}}(x) = \alpha'^m_{\{u,v\}}(x) \prod_{j=1}^{m} (s + A^j_w). \tag{4.298}$$

To prove Theorem 4.36, we use Lemma 4.37 and Lemma 4.38.

Lemma 4.37. *Let p be an integer such that $1 \leq p \leq m-1$, and let $w \in \{1, \dots, k\}$ and $u, v \in \{1, \dots, k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$ and $u_i \neq v_j$ for all $i, j \in \{1, \dots, m-1\}$. If $x = g_{m,k}(\omega, A)$, then*

$$\beta'^m_{w,\{u,v\},p}(x) = \alpha'^m_{\{u,v\}}(x) \sum_{1 \leq t_1 < \cdots < t_p \leq m} \Big(\prod_{j=1}^{p} A^{t_j}_w\Big)\Big(\sum_{t \in \mathbf{m} \setminus \{t_1, \dots, t_p\}} \omega_t\Big). \tag{4.299}$$

Proof. Suppose that $x = (q_r)_{r \in 2^{\{1,\dots,k\}}}$. By definition of $g_{m,k}$ and $\beta'^m_{w,\{u,v\},p}$, we have

$$\beta'^m_{w,\{u,v\},p}(x) = \sum_{(P_1,P_2) \in P_{m,p}} \sum_{\sigma \in P_{m-1}} \det P(\sigma)\Big(\prod_{j \in P_1} q_{\{u_j, v_{\sigma(j)}\}}\Big)\Big(\prod_{j \in P_2} q_{\{w, u_j, v_{\sigma(j)}\}}\Big) \tag{4.300}$$

$$= \sum_{(P_1,P_2) \in P_{m,p}} \sum_{\sigma \in P_{\mathbf{m-1}}} \det P(\sigma)\Big(\prod_{j \in P_1} \sum_{t_j=1}^{m} \omega_{t_j} A^{t_j}_{u_j} A^{t_j}_{v_{\sigma(j)}}\Big) \tag{4.301}$$

$$\Big(\prod_{j \in P_2} \sum_{t_j=1}^{m} \omega_{t_j} A^{t_j}_w A^{t_j}_{u_j} A^{t_j}_{v_{\sigma(j)}}\Big).$$

Reorganizing the terms, we obtain

$$\beta'^m_{w,\{u,v\},p}(x) = \sum_{1 \leq t_1, \dots, t_{m-1} \leq m} \det A^{(t_1,\dots,t_{m-1})}_v \Big(\prod_{j=1}^{m-1} \omega_{t_j} A^{t_j}_{u_j}\Big)\Big(\sum_{(P_1,P_2) \in P_{m,p}} \prod_{j \in P_2} A^{t_j}_w\Big). \tag{4.302}$$

Because $\det A^{(t_1,\dots,t_{m-1})}_v = 0$ if $t_i = t_j$ for some $i \neq j$, the range of the leftmost sum can be restricted to distinct t_i, $i = 1, \dots, m-1$. Hence, we suppose that the t_i are distinct. Let t_m be the only element in $\{1, \dots, m\} \setminus \{t_1, \dots, t_{m-1}\}$, and let $\sigma \in P_m$ be such that $\sigma(i) = t_i$. We have

$$A^{(\sigma(1),\dots,\sigma(m-1))}_v = P(\sigma)^{\hat{m}}_{\sigma(\hat{m})} A^{\sigma(\hat{m})}_v. \tag{4.303}$$

Hence, we have

$$\det A_v^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^{m+\sigma(m)} \det P\,(\sigma) \det A_v^{\widehat{\sigma(m)}}. \tag{4.304}$$

and, by Theorem 4.32,

$$\det A_v^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^m \omega_{\sigma(m)} \det P\,(\sigma) \sum_{j=1}^{m} (-1)^j \det A_v^{\hat{j}}. \tag{4.305}$$

Hence, $\beta'^{m}_{w,\{u,v\},p}(x)$ is equal to

$$\Big(\prod_{j=1}^{m} \omega_j\Big)\Big(\sum_{j=1}^{m} (-1)^j \det A_v^{\hat{j}}\Big)\Big(\sum_{\sigma\in P_m} (-1)^m \det P\,(\sigma)\Big(\prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)}\Big)\Big(\sum_{\substack{S\subseteq\mathbf{m-1}\\ |S|=p}} \prod_{j\in S} A_w^{\sigma(j)}\Big)\Big). \tag{4.306}$$

Consider the last factor of this product. The sum over P_m can be decomposed so that the factor is equal to

$$\sum_{i=1}^{m} \sum_{\sigma'\in P_{m-1}} (-1)^m \det P\,(\sigma)\Big(\prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)}\Big)\Big(\sum_{\substack{S\subseteq\mathbf{m-1}\\ |S|=p}} \prod_{j\in S} A_w^{\sigma(j)}\Big). \tag{4.307}$$

where $\sigma \in P_m$ is defined by

$$\sigma(j) = \begin{cases} \sigma'(j) & \text{for } j < m \text{ and } \sigma'(j) < i \\ \sigma'(j) + 1 & \text{for } j < m \text{ and } \sigma'(j) \geq i \\ i & \text{for } j = m. \end{cases} \tag{4.308}$$

For $j \in \{1, \ldots, m-1\}$, we thus have

$$A_{u_j}^{\sigma(j)} = (A^{\hat{i}})_{u_j}^{\sigma'(j)} \tag{4.309}$$

$$A_{w}^{\sigma(j)} = (A^{\hat{i}})_{w}^{\sigma'(j)}. \tag{4.310}$$

Also, we have

$$P\,(\sigma') = P\,(\sigma)^{\hat{m}}_{\hat{i}}, \tag{4.311}$$

and thus we have

$$\det P\,(\sigma) = (-1)^{m+i} \det P\,(\sigma'). \tag{4.312}$$

Moreover, we have

$$\sum_{\substack{S\subseteq\mathbf{m-1}\\ |S|=p}} \prod_{j\in S} (A^{\hat{i}})_w^{\sigma'(j)} = \sum_{\substack{S\subseteq\mathbf{m-1}\\ |S|=p}} \prod_{j\in S} (A^{\hat{i}})_w^{j}. \tag{4.313}$$

By (4.307) to (4.313) we thus have

$$\sum_{\sigma\in P_m}(-1)^m \det P(\sigma)\Big(\prod_{j=1}^{m-1} A_{u_j}^{\sigma(j)}\Big)\Big(\sum_{\substack{S\subseteq \mathbf{m-1}\\ |S|=p}}\prod_{j\in S} A_w^{\sigma(j)}\Big)$$
$$= \sum_{i=1}^{m}(-1)^i\Big(\sum_{\substack{S\subseteq \mathbf{m-1}\\ |S|=p}}\prod_{j\in S}(A^{\hat{i}})_w^j\Big)\det A_u^{\hat{i}}. \tag{4.314}$$

By Theorem 4.32 and Theorem 4.33, we thus have

$$\beta'^m_{w,\{u,v\},p}(x) = \alpha'^m_{\{u,v\}}(x)\sum_{i=1}^{m}\omega_i\Big(\sum_{\substack{S\subseteq \mathbf{m-1}\\ |S|=p}}\prod_{j\in S}(A^{\hat{i}})_w^j\Big). \tag{4.315}$$

Reorganizing the terms, we conclude the proof. □

Lemma 4.38. *Let p be an integer such that $2 \le p \le m$, and let $w \in \{1,\dots,k\}$ and $u, v \in \{1,\dots,k\}^{m-1}$ satisfy $w \neq u_i$, $w \neq v_j$, and $u_i \neq v_j$ for all $i, j \in \{1,\dots,m-1\}$. If $x = g_{m,k}(\omega, A)$, then*

$$\gamma'^m_{w,\{u,v\},p}(x) = -\alpha'^m_{\{u,v\}}(x)\sum_{1\le t_1<\dots<t_p\le m}\Big(\prod_{j=1}^{p} A_w^{t_j}\Big)\Big(\sum_{j=1}^{p}\omega_{t_j}\Big). \tag{4.316}$$

Proof. Suppose that $x = (q_r)_{r\in 2^{\{1,\dots,k\}}}$. If $\sigma \in P_{m-1}$, let $F(\sigma)$ be the $(m-1)\times(m-1)$ matrix defined by

$$F(\sigma)^i_j = \begin{cases} q_{\{w,u_{\sigma(i)}\}}q_{\{w,v_j\}} & \text{for } i = 1,\\ q_{\{w,u_{\sigma(i)},v_j\}} & \text{for } i = 2,\dots,p-1,\\ q_{\{u_{\sigma(i)},v_j\}} & \text{for } i = p,\dots,m-1.\end{cases} \tag{4.317}$$

If $(P_1, P_2, P_3) \in P'_{m,p}$, there are $(m-p)!(p-2)!$ permutations $\sigma \in P_{m-1}$ satisfying

$$\sigma(i) \in \begin{cases} P_3 & \text{if } i = 1\\ P_2 & \text{if } i = 2,\dots,p-1\\ P_1 & \text{if } i = p,\dots,m-1.\end{cases} \tag{4.318}$$

Moreover, for each such permutation σ, we have

$$B_{P_1,P_2,P_3} = P(\sigma)F(\sigma). \tag{4.319}$$

Hence, we have

$$\gamma'^m_{w,\{u,v\},p}(x) = \frac{1}{(m-p)!(p-2)!}\sum_{\sigma_u\in P_{m-1}}\det(P(\sigma_u)F(\sigma_u)), \tag{4.320}$$

and, by definition of F, we have

$$\gamma'^{m}_{w,\{u,v\},p}(x) = \frac{1}{(m-p)!(p-2)!} \sum_{\sigma_u \in P_{m-1}} \sum_{\sigma_v \in P_{m-1}} \det P(\sigma_u) \det P(\sigma_v)$$
$$q_{\{w,u_{\sigma_u(1)}\}} q_{\{w,v_{\sigma_v(1)}\}} \Big(\prod_{j=2}^{p-1} q_{\{w,u_{\sigma_u(j)},v_{\sigma_v(j)}\}}\Big)\Big(\prod_{j=p}^{m-1} q_{\{u_{\sigma_u(j)},v_{\sigma_v(j)}\}}\Big). \quad (4.321)$$

By definition of $g_{m,k}$, $\gamma'^{m}_{w,\{u,v\},p}(x)$ is thus equal to

$$\frac{1}{(m-p)!(p-2)!} \sum_{\sigma_u \in P_{m-1}} \sum_{\sigma_v \in P_{m-1}} \det P(\sigma_u) \det P(\sigma_v) (\sum_{t_1=1}^{m} \omega_{t_1} A_w^{t_1} A_{u_{\sigma_u(1)}}^{t_1})$$
$$(\sum_{t_m=1}^{m} \omega_{t_m} A_w^{t_m} A_{v_{\sigma_v(1)}}^{t_m})(\prod_{j=2}^{p-1} \sum_{t_j=1}^{m} \omega_{t_j} A_w^{t_j} A_{u_{\sigma_u(j)}}^{t_j} A_{v_{\sigma_v(j)}}^{t_j})(\prod_{j=p}^{m-1} \sum_{t_j=1}^{m} \omega_{t_j} A_{u_{\sigma_u(j)}}^{t_j} A_{v_{\sigma_v(j)}}^{t_j}). \quad (4.322)$$

Reorganizing the terms, we obtain

$$\frac{1}{(m-p)!(p-2)!} \sum_{1 \le t_1,\ldots,t_m \le m} A_w^{t_m} (\prod_{j=1}^{p-1} A_w^{t_j})(\prod_{j=1}^{m} \omega_{t_j}) \det A_u^{(t_1,\ldots,t_{m-1})} \det A_v^{(t_m,t_2,\ldots,t_{m-1})}. \quad (4.323)$$

Because of the determinants, a term of the above sum will vanish if $t_i = t_j$ for distinct $i, j \in \{1, \ldots m-1\}$ or distinct $i, j \in \{2, \ldots, m\}$. Hence, the sum can be restricted to avoid these cases. Let us consider the terms where $t_1 = t_m$ and the terms where $t_1 \neq t_m$ separately.

1. Consider the terms of the sum in (4.323) where $t_1 = t_m$. Let $\sigma \in P_m$ be such that $\sigma(i) = t_i$ for $i \in \{1, \ldots, m-1\}$, and let $\sigma(m)$ be the only element of $\{1, \ldots, m\} \setminus \{t_1, \ldots, t_{m-1}\}$. The terms of the sum where $t_1 = t_m$ can be written as

$$\frac{1}{(m-p)!(p-2)!} \sum_{\sigma \in P_m} \omega_{\sigma(1)} A_w^{\sigma(1)} (\prod_{j=1}^{p-1} A_w^{\sigma(j)})(\prod_{j=1}^{m-1} \omega_{\sigma(j)})$$
$$\det A_u^{(\sigma(1),\ldots,\sigma(m-1))} \det A_v^{(\sigma(1),\ldots,\sigma(m-1))}. \quad (4.324)$$

We have

$$A_u^{(\sigma(1),\ldots,\sigma(m-1))} = P(\sigma)^{\hat{m}}_{\hat{\sigma(m)}} A_u^{\hat{\sigma(m)}}, \quad (4.325)$$
$$A_v^{(\sigma(1),\ldots,\sigma(m-1))} = P(\sigma)^{\hat{m}}_{\hat{\sigma(m)}} A_v^{\hat{\sigma(m)}}, \quad (4.326)$$

and thus we have

$$A_u^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^{m+\sigma(m)} \det P(\sigma) \det A_u^{\hat{\sigma(m)}}, \quad (4.327)$$
$$A_v^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^{m+\sigma(m)} \det P(\sigma) \det A_v^{\hat{\sigma(m)}}. \quad (4.328)$$

By Theorem 4.32 and Theorem 4.33, (4.324) is thus equal to

$$\frac{\alpha'^{m}_{\{u,v\}}(x)}{(m-p)!(p-2)!}\sum_{\sigma\in P_m}\omega_{\sigma(1)}\omega_{\sigma(m)}A_w^{\sigma(1)}(\prod_{j=1}^{k-1}A_w^{\sigma(j)}). \tag{4.329}$$

2. Consider the terms of the sum in (4.323) where $t_1 \neq t_m$. Let $\sigma \in P_m$ be such that $\sigma(i) = t_i$ for $i \in \{1, \ldots, m\}$. These terms can be written as

$$\frac{1}{(m-p)!(p-2)!}(\prod_{j=1}^{m}\omega_j)\sum_{\sigma\in P_m}A_w^{\sigma(m)}(\prod_{j=1}^{p-1}A_w^{\sigma(j)})$$
$$\det A_u^{(\sigma(1),\ldots,\sigma(m-1))}\det A_v^{(\sigma(m),\sigma(2),\ldots,\sigma(m-1))}. \tag{4.330}$$

Let us consider the two determinants separately.

(a) We have

$$A_u^{(\sigma(1),\ldots,\sigma(m-1))} = P(\sigma)^{\hat{m}}_{\sigma(\hat{m})}A_u^{\sigma(\hat{m})}. \tag{4.331}$$

Hence, we have

$$\det A_u^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^{\sigma(m)+m}\det P(\sigma)\det A_u^{\sigma(\hat{m})}, \tag{4.332}$$

and, by Theorem 4.32, we have

$$\det A_u^{(\sigma(1),\ldots,\sigma(m-1))} = (-1)^{m}\det P(\sigma)\omega_{\sigma(m)}\sum_{j=1}^{m}(-1)^{j}\det A_u^{\hat{j}}. \tag{4.333}$$

(b) We have

$$A_v^{(\sigma(2),\ldots,\sigma(m))} = P(\sigma)^{\hat{1}}_{\sigma(\hat{1})}A_v^{\sigma(\hat{1})}. \tag{4.334}$$

Moreover, $A_v^{(\sigma(m),\sigma(2),\ldots,\sigma(m-1))}$ can be obtained from $A_v^{(\sigma(2),\ldots,\sigma(m))}$ by $m-$ 2 row permutations. Hence, we have

$$\det A_v^{(\sigma(m),\sigma(2),\ldots,\sigma(m-1))} = (-1)^{\sigma(1)+m-1}\det P(\sigma)\det A_v^{\sigma(\hat{1})}, \tag{4.335}$$

and, by Theorem 4.32, we have

$$\det A_v^{(\sigma(m),\sigma(2),\ldots,\sigma(m-1))} = -(-1)^{m}\det P(\sigma)\omega_{\sigma(1)}\sum_{j=1}^{m}(-1)^{j}\det A_v^{\hat{j}}. \tag{4.336}$$

By (4.333), (4.336) and Theorem 4.33, we see that (4.330) is equal to

$$-\frac{\alpha'^{m}_{\{u,v\}}(x)}{(m-p)!(p-2)!}\sum_{\sigma\in P_m}\omega_{\sigma(1)}\omega_{\sigma(m)}A_w^{\sigma(m)}(\prod_{j=1}^{k-1}A_w^{\sigma(j)}). \tag{4.337}$$

Adding (4.329) and (4.337), we see that

$$\gamma'^m_{w,\{u,v\},p}(x) = \frac{\alpha'^m_{\{u,v\}}(x)}{(m-p)!(p-2)!}\sum_{\sigma\in P_m}\omega_{\sigma(1)}\omega_{\sigma(m)}(A_w^{\sigma(1)}-A_w^{\sigma(m)})(\prod_{j=1}^{p-1}A_w^{\sigma(j)}). \tag{4.338}$$

Decomposing $\sum_{\sigma\in P_m}$ into $\sum_{t_1\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_1,\ldots,t_{m-1}\}}$ with $t_i=\sigma(i)$ for $i=1,\ldots,m$ and rearranging the terms, we obtain

$$\begin{aligned}\gamma'^m_{w,\{u,v\},p}(x) = \frac{\alpha'^m_{\{u,v\}}(x)}{(m-p)!(p-2)!}\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}\omega_{t_m}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}A_w^j)\\ \sum_{t_1\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}\omega_{t_1}(A_w^{t_1}-A_w^{t_m})\sum_{t_2\in\mathbf{m}\setminus\{t_1,t_p,\ldots,t_m\}}\cdots\sum_{t_{p-1}\in\mathbf{m}\setminus\{t_1,\ldots,t_{p-2},t_p,\ldots,t_m\}}1.\end{aligned} \tag{4.339}$$

Since there are $(p-2)!$ terms in the sums over $t_2,\ldots,t_{p-1}$, we have

$$\sum_{t_2\in\mathbf{m}\setminus\{t_1,t_p,\ldots,t_m\}}\cdots\sum_{t_{p-1}\in\mathbf{m}\setminus\{t_1,\ldots,t_{p-2},t_p,\ldots,t_m\}}1=(p-2)!. \tag{4.340}$$

Also, because $\sum_{t=1}^m\omega_t=1$ and $\omega A=0$, we have

$$\sum_{t_1\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}\omega_{t_1}(A_w^{t_1}-A_w^{t_m})=-(\sum_{j=p}^m\omega_{t_j}A_w^{t_j})-A_w^{t_m}(1-\sum_{j=p}^m\omega_{t_j}). \tag{4.341}$$

Hence, we have

$$\begin{aligned}\gamma'^m_{w,\{u,v\},p}(x) = -\frac{\alpha'^m_{\{u,v\}}(x)}{(m-p)!}\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}A_w^j)\\ (\omega_{t_m}A_w^{t_m}+\omega_{t_m}\sum_{j=p}^m\omega_{t_j}A_w^{t_j}-\omega_{t_m}A_w^{t_m}\sum_{j=p}^m\omega_{t_j}).\end{aligned} \tag{4.342}$$

For $j\in\{p,\ldots,m\}$, one can see by permuting the labels of t_j and t_m that

$$\begin{aligned}&\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}A_w^j)\omega_{t_m}\omega_{t_j}A_w^{t_j}\\ &\qquad=\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}A_w^j)\omega_{t_j}\omega_{t_m}A_w^{t_m}.\end{aligned} \tag{4.343}$$

Hence, we have

$$\gamma'^m_{w,\{u,v\},p}(x) = -\frac{\alpha'^m_{\{u,v\}}(x)}{(m-p)!}\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_m\}}A_w^j)\omega_{t_m}A_w^{t_m}, \tag{4.344}$$

and thus

$$\gamma'^m_{w,\{u,v\},p}(x) = -\frac{\alpha'^m_{\{u,v\}}(x)}{(m-p)!}\sum_{t_p\in\mathbf{m}}\cdots\sum_{t_{m-1}\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-2}\}}(\prod_{j\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}A_w^j)\sum_{t_m\in\mathbf{m}\setminus\{t_p,\ldots,t_{m-1}\}}\omega_{t_m}. \tag{4.345}$$

Reorganizing the terms of the above sums, we conclude the proof. □

Proof (Theorem 4.36). To prove the theorem, let us show that we have

$$\beta'^{m}_{w,\{u,v\},p}(x) - \gamma'^{m}_{w,\{u,v\},p}(x) = \alpha'^{m}_{\{u,v\}}(x) \sum_{1\leq t_1<\cdots<t_p\leq m} (\prod_{j=1}^{p} A^{t_j}_{w}) \tag{4.346}$$

for $1 \leq p \leq m$.

1. Suppose that $p = 1$. Then, $\gamma'^{m}_{w,\{u,v\},1}x = 0$, and we have

$$\begin{aligned}\beta'^{m}_{w,\{u,v\},1}x &= \alpha'^{m}_{\{u,v\}}(x) \sum_{t_1=1}^{m} A^{t_1}_{w} \sum_{t\in\mathbf{m}\setminus\{t_1\}} \omega_t && \text{by Lemma 4.37} && (4.347)\\ &= \alpha'^{m}_{\{u,v\}}(x) \sum_{t_1=1}^{m} A^{t_1}_{w}(1-\omega_{t_1}) && \text{because } \sum_{t=1}^{m} \omega_t = 1 && (4.348)\\ &= \alpha'^{m}_{\{u,v\}}(x) \sum_{t_1=1}^{m} A^{t_1}_{w} && \text{because } \omega A = 0. && (4.349)\end{aligned}$$

2. Suppose that $2 \leq p \leq m - 1$. By Lemma 4.37 and Lemma 4.38, we have

$$\beta'^{m}_{w,\{u,v\},p}(x) - \gamma'^{m}_{w,\{u,v\},p}(x) = \alpha'^{m}_{\{u,v\}}(x) \sum_{1\leq t_1<\cdots<t_p\leq m} (\prod_{j=1}^{p} A^{t_j}_{w})(\sum_{t=1}^{m} \omega_t), \tag{4.350}$$

and, because $\sum_{t=1}^{m} \omega_t = 1$, we have

$$\beta'^{m}_{w,\{u,v\},p}(x) - \gamma'^{m}_{w,\{u,v\},p}(x) = \alpha'^{m}_{\{u,v\}}(x) \sum_{1\leq t_1<\cdots<t_p\leq m} (\prod_{j=1}^{p} A^{t_j}_{w}). \tag{4.351}$$

3. Suppose that $p = m$. Then, $\beta'^{m}_{w,\{u,v\},p}(x) = 0$, and we have

$$-\gamma'^{m}_{w,\{u,v\},p}(x) = \alpha'^{m}_{\{u,v\}}(x) \sum_{1\leq t_1<\cdots<t_m\leq m} (\prod_{j=1}^{m} A^{t_j}_{w}) \tag{4.352}$$

by Lemma 4.38 and because $\sum_{t=1}^{m} \omega_t = 1$. □

4.7.5 Theorem 4.13

The following family of functions generalizes $\zeta'^{m}_{s,\{u,v\}}$.

Definition 169. If $s \subseteq \{1, \ldots, k\}$ and $u, v \in \{1, \ldots, k\}^{m-1}$ satisfy $u_i \notin s$, $v_j \notin s$, and $u_i \neq v_j$ for all $i, j \in \{1, \ldots, m-1\}$, the function $\zeta'^{m}_{s,\{u,v\}} : \mathbb{R}^{2^k} \to \mathbb{R}$ is defined by

$$\zeta'^{m}_{s,\{u,v\}}((q_r)_{r\in 2^{\{1,\ldots,k\}}}) = \sum_{p=1}^{m-1} \det B(p) \tag{4.353}$$

where $B(p)$ is the $(m-1)\times(m-1)$ matrix such that

$$B(p)^i_j = \begin{cases} q_{s\cup\{u_i,v_j\}} & \text{for } i = p, \\ q_{\{u_i,v_j\}} & \text{for } i \neq p. \end{cases} \tag{4.354}$$

Theorem 4.13 is a special case of the following result.

Theorem 4.39. *Let $s \subseteq \{1,\ldots,k\}$ and $u, v \in \{1,\ldots,k\}^{m-1}$ satisfy $u_i \notin s$, $v_j \notin s$, and $u_i \neq v_j$ for all $i, j \in \{1,\ldots,m-1\}$. If $x = (q_r)_{r\in 2^{\{1,\ldots,k\}}} = g_{m,k}(\omega, A)$, then*

$$\alpha'^m_{\{u,v\}}(x)(\sum_{i=1}^m \prod_{j\in s} A^i_j) = \alpha'^m_{\{u,v\}}(x)q_s + \zeta'^m_{s,\{u,v\}}(x). \tag{4.355}$$

Proof. Let us consider separately $\alpha'^m_{\{u,v\}}(x)(\sum_{i=1}^m \prod_{j\in s} A^i_j)$, $\alpha'^m_{\{u,v\}}(x)q_s$, and $\zeta'^m_{s,\{u,v\}}(x)$.

1. By (4.272), we have

$$\alpha'^m_{\{u,v\}}(x)(\sum_{i=1}^m \prod_{j\in s} A^i_j) = (\sum_{i=1}^m \prod_{j\in s} A^i_j) \sum_{1\leq t_1,\ldots,t_{m-1}\leq m} \det A_v^{(t_1,\ldots,t_{m-1})} \prod_{j=1}^{m-1} \omega_{t_j} A^{t_j}_{u_j}, \tag{4.356}$$

and thus

$$\alpha'^m_{\{u,v\}}(x)(\sum_{i=1}^m \prod_{j\in s} A^i_j) = \sum_{\sigma\in P_m} (\sum_{i=1}^m \prod_{j\in s} A^{\sigma(i)}_j) \det A_v^{(\sigma(1),\ldots,\sigma(m-1))} \prod_{j=1}^{m-1} \omega_{\sigma(j)} A^{\sigma(j)}_{u_j}. \tag{4.357}$$

2. By Theorem 4.33 and by definition of $g_{m,k}$, we have

$$\alpha'^m_{\{u,v\}}(x)q_s = \Big(\prod_{j=1}^m \omega_j\Big)\Big(\sum_{j=1}^m (-1)^j \det A_v^{\hat{j}}\Big)\Big(\sum_{j=1}^m (-1)^j \det A_u^{\hat{j}}\Big)\Big(\sum_{t_m=1}^m \omega_{t_m} \prod_{i\in s} A^{t_m}_i\Big). \tag{4.358}$$

By Theorem 4.32, we thus have

$$\alpha'^m_{\{u,v\}}(x)q_s = \Big(\prod_{j=1}^m \omega_j\Big)\Big(\sum_{j=1}^m (-1)^j \det A_v^{\hat{j}}\Big)\Big(\sum_{t_m=1}^m (-1)^{t_m} \det A_u^{\hat{t_m}} \prod_{i\in s} A^{t_m}_i\Big). \tag{4.359}$$

We have

$$\begin{aligned} &\sum_{t_m=1}^m (-1)^{t_m} \det A_u^{\hat{t_m}} \prod_{i\in s} A^{t_m}_i \\ &\qquad = \sum_{t_m=1}^m \sum_{\sigma'\in P_{m-1}} (-1)^{t_m}\Big(\prod_{i\in s} A^{t_m}_i\Big) \det P(\sigma')\Big(\prod_{j=1}^{m-1} (A_u^{\hat{t_m}})^{\sigma'(j)}_j\Big). \end{aligned} \tag{4.360}$$

Hence, we have

$$\sum_{t_m=1}^{m}(-1)^{t_m}\det A_u^{\hat{t_m}}\prod_{i\in s}A_i^{t_m} = \sum_{\sigma\in P_m}(-1)^{t_m}\Big(\prod_{i\in s}A_i^{t_m}\Big)\det P(\sigma')\Big(\prod_{j=1}^{m-1}(A_u^{\hat{t_m}})_j^{\sigma'(j)}\Big) \tag{4.361}$$

where $t_m = \sigma(m)$ and $\sigma' \in P_{m-1}$ is such that

$$\sigma'(i) = \begin{cases}\sigma(i) & \text{if } \sigma(i) < \sigma(m),\\ \sigma(i)-1 & \text{if } \sigma(i) > \sigma(m).\end{cases} \tag{4.362}$$

For $j \in \{1,\ldots,m-1\}$, we have

$$(A_u^{\hat{t_m}})_j^{\sigma'(j)} = (A_u)_j^{\sigma(j)} = A_{u_j}^{\sigma(j)}. \tag{4.363}$$

Also, we have

$$P(\sigma') = P(\sigma)^{\hat{m}}_{\hat{\sigma(m)}}, \tag{4.364}$$

and thus we have

$$\det P(\sigma') = (-1)^{m+\sigma(m)}\det P(\sigma). \tag{4.365}$$

Hence, we have

$$\sum_{t_m=1}^{m}(-1)^{t_m}\det A_u^{\hat{t_m}}\prod_{i\in s}A_i^{\sigma(m)} = (-1)^m\sum_{\sigma\in P_m}\det P(\sigma)\Big(\prod_{i\in s}A_i^{\sigma(m)}\Big)\Big(\prod_{j=1}^{m-1}A_{u_j}^{\sigma(j)}\Big). \tag{4.366}$$

By (4.359) and (4.366), we have

$$\begin{aligned}\alpha'^{m}_{\{u,v\}}(x)q_s = (-1)^m\sum_{\sigma\in P_m}\omega_{\sigma(m)}\Big(\sum_{j=1}^{m}(-1)^j\det A_v^{\hat{j}}\Big)&\\ \det P(\sigma)\Big(\prod_{i\in s}A_i^{\sigma(m)}\Big)\Big(\prod_{j=1}^{m-1}\omega_{\sigma(j)}A_{u_j}^{\sigma(j)}\Big).&\end{aligned} \tag{4.367}$$

By Theorem 4.32, we have

$$\begin{aligned}\alpha'^{m}_{\{u,v\}}(x)q_s = \sum_{\sigma\in P_m}(-1)^{m+\sigma(m)}\det A_v^{\hat{\sigma(m)}}\det P(\sigma)&\\ \Big(\prod_{i\in s}A_i^{\sigma(m)}\Big)\Big(\prod_{j=1}^{m-1}\omega_{\sigma(j)}A_{u_j}^{\sigma(j)}\Big).&\end{aligned} \tag{4.368}$$

Moreover, we have

$$A_v^{(\sigma(1),\ldots,\sigma(m-1))} = P(\sigma)^{\hat{m}}_{\hat{\sigma(m)}}A_v^{\hat{\sigma(m)}}, \tag{4.369}$$

and thus

$$\alpha'^m_{\{u,v\}}(x)q_s = \sum_{\sigma\in P_m} \det A_v^{(\sigma(1),\ldots,\sigma(m-1))}\Big(\prod_{i\in s} A_i^{\sigma(m)}\Big)\Big(\prod_{j=1}^{m-1} \omega_{\sigma(j)}A_{u_j}^{\sigma(j)}\Big). \tag{4.370}$$

3. By definition of $\zeta'^m_{s,\{u,v\}}$ and $g_{m,k}$, we have

$$\zeta'^m_{s,\{u,v\}}(x) = \sum_{p=1}^{m-1}\sum_{\sigma\in P_{m-1}} \det P(\sigma) q_{s\cup\{u_p,v_{\sigma(p)}\}}\Big(\prod_{\substack{j=1\\ j\neq p}}^{m-1} q_{\{u_j,v_{\sigma(j)}\}}\Big) \tag{4.371}$$

$$= \sum_{p=1}^{m-1}\sum_{\sigma\in P_{m-1}} \det P(\sigma)\Big(\prod_{\substack{j=1\\ j\neq p}}^{m-1}\sum_{t_j=1}^{m} \omega_{t_j}A_{u_j}^{t_j}A_{v_{\sigma(j)}}^{t_j}\Big) \tag{4.372}$$

$$\Big(\sum_{t_p=1}^{m} \omega_{t_p}A_{u_p}^{t_p}A_{v_{\sigma(p)}}^{t_p}\prod_{i\in s} A_i^{t_p}\Big).$$

Reorganizing the terms, we obtain

$$\zeta'^m_{s,\{u,v\}}(x) = \sum_{1\le t_1,\ldots,t_{m-1}\le m}\Big(\sum_{p=1}^{m-1}\prod_{i\in s} A_i^{t_p}\Big)\det A_v^{(t_1,\ldots,t_{m-1})}\Big(\prod_{j=1}^{m-1}\omega_{t_j}A_{u_j}^{t_j}\Big). \tag{4.373}$$

Because of the determinant, we suppose that the t_i are distinct and we have

$$\zeta'^m_{s,\{u,v\}}(x) = \sum_{\sigma\in P_m}\Big(\sum_{p=1}^{m-1}\prod_{i\in s} A_i^{\sigma(p)}\Big)\det A_v^{(\sigma(1),\ldots,\sigma(m-1))}\Big(\prod_{j=1}^{m-1}\omega_{\sigma(j)}A_{u_j}^{\sigma(j)}\Big). \tag{4.374}$$

Using (4.357), (4.370) and (4.374), it is straightforward to conclude the proof. □

Conclusion

The contributions of this dissertation can be summarized as follows. Chapter 3 proposed efficient algorithms to compute the inclusion boundary of an arbitrary essential graph. We found that elements of the boundary can be identified simply and efficiently. Moreover, the difference in score between an essential graph and an element of its boundary can be evaluated efficiently when the scoring criterion is decomposable. Finally, the number of elements in the inclusion boundary of an essential graph may sometimes be exponential in its number of vertices. When it is used as a neighborhood in a greedy structure learning algorithm, the potentially large size of the boundary may be an issue worth exploring.

The main contribution of Chapter 4 consists of the results gathered in Section 4.7. There, a function very close to the parametrization map of a discrete Naive Bayes model with hidden class variable is defined. Then, several polynomial equations satisfied by the elements of the graph of this function are introduced. These polynomial equations are remarkable because each involve only a small number of input variables and with a low degree. Furthermore, they allow us to compute explicitely some fibers of the function. Another contribution of Chapter 4 is the application of these results to compute fibers of discrete Naive Bayes models with hidden class variable and learn their parameter from data. First, an alternative parametrization of the discrete Naive Bayes models is introduced. Then, two algorithms that compute fibers of the parametrization map under appropriate technical assumptions are derived. Finally, these algorithms are converted into parameter learning algorithms for a special class of discrete Naive Bayes models. Under appropriate assumptions, the resulting algorithms have nice asymptotic properties, but also many rough edges that future research should address. Among other problems, their applicability is limited, they have high computational complexity, and many parameters to choose. We believe that one key to improve our algorithms is a description of fibers that does not involve any enumeration. A second potential research direction is to extend our algorithms and results to other classes of Bayesian network models with hidden variables, such as HLC models. Finally, we suspect that our results may be useful to study the geometry of discrete Naive Bayes models and estimate the number of hidden classes.

Appendix

A Common Densities and Families

Definition 170. Let X be a continuous random variable, let μ be a real, and let σ^2 be a strictly positive real. The *(univariate) Gaussian density* $\mathcal{N}(x|\mu,\sigma^2)$ is

$$(2\pi\sigma^2)^{-1/2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}. \tag{A.1}$$

The mean and variance of $\mathcal{N}(x|\mu,\sigma^2)$ are μ and σ^2.

Definition 171. Let X be a k-dimensional vector of continuous random variables, let μ be a vector in $\mathbb{R}^k$, and let Σ be a symmetric positive definite matrix in $\mathbb{R}^{k\times k}$. The *multivariate Gaussian density* $\mathcal{N}(x|\mu,\Sigma)$ is

$$(2\pi)^{-k/2}(\det\Sigma)^{-1/2}e^{\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^k. \tag{A.2}$$

The mean vector and covariance matrix of $\mathcal{N}(x|\mu,\Sigma)$ are μ and Σ.

A Dirichlet density defines a probability distribution over the set of strictly positive distributions of a discrete random variable.

Definition 172. Let X be a discrete random variable, let $p = (p_x)_{x\in\mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}$, let $\alpha \in \mathbb{R}_{>0}$, and let $m \in S^+_{\mathcal{X}}$. The *Dirichlet density* $\mathcal{D}(p|\alpha,m)$ is

$$\frac{\Gamma(\alpha)}{\prod_{x\in\mathcal{X}}\Gamma(\alpha m_x)}\prod_{x\in\mathcal{X}} p_x^{\alpha m_x - 1}\delta(\sum_{x\in\mathcal{X}} p_x - 1), \quad p \in \mathbb{R}^{|\mathcal{X}|} \tag{A.3}$$

where δ is the Dirac delta function.

The mean vector of $\mathcal{D}(p|\alpha,m)$ is m and the variance of p_x is $m_x(1-m_x)/(\alpha+1)$ for $x \in \mathcal{X}$. Let us now define a very general class of families.

Definition 173. Let $\Theta \subseteq \mathbb{R}^k$ be a parameter space and let μ be a σ-finite measure on a sample space $\mathcal{X}$. Let h and t be functions from $\mathcal{X}$ to respectively $\mathbb{R}_{\geq 0}$ and $\mathbb{R}^k$. The family of densities w.r.t. μ given by $p = f(\theta)$ with

$$p(x) = h(x)e^{\theta^T t(x) - \psi(\theta)} \tag{A.4}$$

is called an *exponential family*.

The *natural parameter space* is

$$N = \{\theta \in \mathbb{R}^k \big| \int_X h(x) e^{\theta^T t(x)} d\mu(x) < \infty\}. \tag{A.5}$$

The family is said to be *full* if $\Theta = N$, *regular* if N is open, and *minimal* if it can not be reparametrized by a vector with less than k components. In the minimal case, the *dimension* of the family is k. Many common densities can be parametrized as (A.4). For example, one can show that Gaussian, Dirichlet, and discrete densities are exponential. Moreover, learning (see chapter 2) with exponential families is often tractable because they admit conjugate families (see e.g. [Rob94] for details).

Some submodels of an exponential family are noteworthy (see [GHKM01]).

Definition 174. Consider a full exponential model of dimension k and natural parameter space N. The submodel induced by $\Theta \subseteq N$ is a

- *curved exponential model* of dimension d if Θ is a smooth d-dimensional manifold in $\mathbb{R}^k$;
- *stratified exponential* model of dimension d if Θ is a d-dimensional stratified set in $\mathbb{R}^k$.

B Semi-Algebraic Sets

This section defines semi-algebraic sets and presents some elementary properties. For more details, see e.g. [BCR87].

Definition 175. A subset V of $\mathbb{R}^n$ is *semi-algebraic* if it admits a representation of the form

$$V = \cup_{i=1}^{s} \cap_{j=1}^{r_i} V_{ij}, \tag{B.1}$$

where, for each $i = 1, \ldots, s$ and $j = 1, \ldots r_i$, V_{ij} is $\{x \in \mathbb{R}^n | P_{ij}(x) > 0\}$ or $\{x \in \mathbb{R}^n | P_{ij}(x) = 0\}$ for a real polynomial P_{ij}.

Definition 176. A *stratification* of a subset E of $\mathbb{R}^n$ is a finite partition $\{A_i\}_{i \in I}$ of E such that

1. each A_i is a d_i-dimensional smooth manifold in $\mathbb{R}^n$ called a *stratum*
2. if $A_j \cap \overline{A_i} \neq \phi$, then $A_j \subseteq \overline{A_i}$ and $d_j < d_i$.

The *dimension* of a stratified set is the largest dimension of a stratum. A stratification is *semi-algebraic* if each stratum is a semi-algebraic set.

The following theorem holds.

Theorem 5.40. *Every semi-algebraic set admits a semi-algebraic stratification.*

The dimension of a semi-algebraic set is its dimension as a stratified set. It can sometimes be computed with the following theorem (from [GHKM01]).

Theorem 5.41. *Let $g : A \subseteq \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping where A is a semi-algebraic open set. Let $J(x) = \frac{\delta g}{\delta x}$ be the Jacobian matrix at x. The maximal rank of $J(x)$ over A is equal to the dimension of $g(A)$.*

C Kullback-Leibler Distance

Definition 177 (*Kullback-Leibler distance*). The *Kullback-Leibler distance* $D(p \parallel q)$ (or *relative entropy*, see [CT91]) between two probability distributions p and q defined on the same finite sample space $\mathcal{X}$ is

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{C.1}$$

The following conventions are adopted by continuity: $0 \log \frac{0}{q(x)} = 0$ for $q(x) \geq 0$ and $p \log \frac{p(x)}{0} = \infty$ for $p(x) > 0$.

The KL distance has the following properties:

- it is non-negative
$$D(p \parallel q) \geq 0, \tag{C.2}$$
with equality only if $p = q$,

- it satisfies Pinsker's inequality
$$D(p \parallel q) \geq \frac{1}{2 \ln 2} \Big(\sum_{x \in \mathcal{X}} |p(x) - q(x)| \Big)^2, \tag{C.3}$$

- it is not symmetric in general
$$D(p \parallel q) \neq D(q \parallel p). \tag{C.4}$$

Although the KL distance is not symmetric, it is often used as a distance between p and q.

The notions of KL distance and data likelihood can be related as follows.

Lemma 5.42. *If X is a discrete random variable, p is a distribution on X, and $\mathcal{M}$ is a set of distributions on X, then*

$$\arg\max_{q \in \mathcal{M}} \prod_{x \in \mathcal{X}} q(x)^{p(x)} = \arg\min_{q \in \mathcal{M}} D(p \parallel q). \tag{C.5}$$

Proof. We have

$$\begin{aligned} D(p \parallel q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad &(C.6) \\ &= \sum_{x \in \mathcal{X}} p(x) \log p(x) - \log \prod_{x \in \mathcal{X}} q(x)^{p(x)}. \quad &(C.7) \end{aligned}$$

Hence, we have

$$\arg\max_{q\in\mathcal{M}} \log \prod_{x\in\mathcal{X}} q(x)^{p(x)} = \arg\min_{q\in\mathcal{M}} D(p \parallel q). \tag{C.8}$$

Since log is a strictly increasing function, we have

$$\arg\max_{q\in\mathcal{M}} \log \prod_{x\in\mathcal{X}} q(x)^{p(x)} = \arg\max_{q\in\mathcal{M}} \prod_{x\in\mathcal{X}} q(x)^{p(x)}. \tag{C.9}$$

□

Corollary 5.43. *If X is a discrete random variable, $\mathcal{M}$ is a set of distributions on X, and $p \in \mathcal{M}$, then*

$$\arg\max_{q\in\mathcal{M}} \prod_{x\in\mathcal{X}} q(x)^{p(x)} = \{p\}. \tag{C.10}$$

Proof. Because $p \in \mathcal{M}$, $D(p \parallel q) \geq 0$ and $D(p \parallel p) = 0$, we have

$$\min_{q\in\mathcal{M}} D(p \parallel q) = D(p \parallel p) = 0. \tag{C.11}$$

Because $D(p \parallel q) = 0$ only if $p = q$, we have

$$\arg\min_{q\in\mathcal{M}} D(p \parallel q) = \{p\}. \tag{C.12}$$

We conclude the proof by Lemma 5.42. □

Corollary 5.44. *If X is a discrete random variable, $\mathcal{M}$ is a set of distributions on X and $o[1], \ldots, o[n]$ is a sequence of values of X, then*

$$\arg\max_{q\in\mathcal{M}} \prod_{i=1}^{n} q(o[i]) = \arg\min_{q\in\mathcal{M}} D(\hat{p} \parallel q) \tag{C.13}$$

where the distribution of relative frequencies $\hat{p}$ is defined by (4.2).

Proof. We have

$$\prod_{i=1}^{n} q(o[i]) = \prod_{x\in\mathcal{X}} q(x)^{n_x} = \Big(\prod_{x\in\mathcal{X}} q(x)^{n_x/n}\Big)^n = \Big(\prod_{x\in\mathcal{X}} q(x)^{\hat{p}(x)}\Big)^n. \tag{C.14}$$

Because $n \geq 1$, we have

$$\arg\max_{q\in\mathcal{M}}\Big(\prod_{x\in\mathcal{X}} q(x)^{\hat{p}(x)}\Big)^n = \arg\max_{q\in\mathcal{M}} \prod_{x\in\mathcal{X}} q(x)^{\hat{p}(x)}. \tag{C.15}$$

By Lemma 5.42, we have

$$\arg\max_{q\in\mathcal{M}} \prod_{x\in\mathcal{X}} q(x)^{\hat{p}(x)} = \arg\min_{q\in\mathcal{M}} D(\hat{p} \parallel q), \tag{C.16}$$

and we conclude the proof. □

Bibliography

[AGW06] Vincent Auvray, Pierre Geurts, and Louis Wehenkel. A semi-algebraic description of discrete Naive Bayes models with two hidden classes. In *Ninth International Symposium on Artificial Intelligence and Mathematics*, January 2006.

[AMP97] Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

[AW02] Vincent Auvray and Louis Wehenkel. On the construction of the inclusion boundary neighbourhood for Markov equivalent classes of Bayesian network structures. In A. Darwiche and N. Friedman, editors, *Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 26–35. Morgan Kaufmann, August 2002.

[BCR87] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Géométrie algébrique réelle*. Springer-Verlag, Berlin, 1987.

[Bil79] Patrick Billingsley. *Probability and Measure*. Wiley, New York, 1979.

[BKRK97] John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2-3):213–244, 1997.

[CDLS99] Robert Cowell, A. Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[CH97] David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29(2-3):181–212, November 1997.

[Chi95] David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 87–98, San Francisco, CA, 1995. Morgan Kaufmann.

[Chi02a] David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, February 2002.

[Chi02b] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, November 2002.

[CK03] Robert Castelo and Tomáš Kočka. On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research*, 4:527–574, September 2003.

[CM02] David Chickering and Christopher Meek. Finding optimal bayesian networks. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, pages 94–102, San Francisco, CA, 2002. Morgan Kaufmann.

[CMH03] David Chickering, Christopher Meek, and David Heckerman. Large-sample learning of bayesian networks is NP-hard. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 124–13, San Francisco, CA, 2003. Morgan Kaufmann.

[Cow98] Robert G. Cowell. Mixture reduction via predictive scores. *Statistics and Computing*, 8:97–103, 1998.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[DT92] Dorit Dor and Michael Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. Technical Report R-185, UCLA Cognitive Systems Laboratory, 1992.

[EF01] Gal Elidan and Nir Friedman. Learning the dimensionality of hidden variables. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)*, pages 144–15, San Francisco, CA, 2001. Morgan Kaufmann.

[ELFK00] Gal Elidan, Noam Lotner, Nir Friedman, and Daphne Koller. Discovering hidden variables: A structure based-approach. In *Proceedings of the Neural Information Processing Systems conference (NIPS)*, 2000.

[FK03] Nir Friedman and Daphne Koller. Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–125, 2003.

[Fri98] Nir Friedman. The bayesian structural EM algorithm. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 129–13, San Francisco, CA, 1998. Morgan Kaufmann.

[Gar04] Luis David Garcia. Algebraic statistics in model selection. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 177–184, Arlington, Virginia, 2004. AUAI Press.

[GH94] Dan Geiger and David Heckerman. Learning gaussian networks. Technical report, Microsoft Research, 1994. MSR-TR-94-10.

[GHKM01] Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001.

[GHM96] Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 283–29, San Francisco, CA, 1996. Morgan Kaufmann.

[GM98] Dan Geiger and Christopher Meek. Graphical models and exponential families. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 156–165, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

[Hau88] Dominique M. A. Haughton. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.

[Hec98] David Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical models*, pages 301–354. Kluwer Academic Publishers, 1998.

[HGC95] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[KC01] Tomáš Kočka and Robert Castelo. Improved learning of Bayesian networks. In *Proceedings of Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2001.

[KZ02] Tomáš Kočka and Nevin Zhang. Dimension correction for hierarchical latent class models. In *Proceedings of the 18th Annual Conference* *on Uncertainty in Artificial Intelligence (UAI-02)*, pages 267–27, San Francisco, CA, 2002. Morgan Kaufmann.

[KZ03] Tomáš Kočka and Nevin Zhang. Effective dimension of partially observed polytrees. In *Proceedings of the Seventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-03)*, 2003.

[Lau95] Steffen L. Lauritzen. The em algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19:191–201, 1995.

[Lau96] Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, United Kingdom, 1996.

[Mac98] David J. C. MacKay. Choice of basis for the Laplace approximation. *Machine Learning*, 33(1):77–86, 1998.

[Mac03] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, United Kingdom, 2003.

[Mee95] Christopher Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 411–41, San Francisco, CA, 1995. Morgan Kaufmann.

[Nea92] Radford M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.

[Nea03] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, Upper Saddle River, NJ, 2003.

[NH98] Radford M. Neal and Geoffrey E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical models*, pages 355–368. Kluwer Academic Publishers, 1998.

[NKP03] Jens Nielsen, Tomas Kocka, and Jose Peña. On local optima in learning Bayesian networks. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 435–44, San Francisco, CA, 2003. Morgan Kaufmann.

[NMR06] Radu Stefan Niculescu, Tom M. Mitchell, and R. Bharat Rao. Bayesian network learning with parameter constraints. *Journal of Machine Learning Research*, 7:1357–1383, July 2006.

[NWL+04] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. *Réseaux bayésiens*. Eyrolles, Paris, second edition, 2004.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, 1988.

[RG03] Dmitry Rusakov and Dan Geiger. Automated analytic asymptotic evaluation of the marginal likelihood for latent model. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 501–50, San Francisco, CA, 2003. Morgan Kaufmann.

[RG05] Dmitry Rusakov and Dan Geiger. Asymptotic model selection for Naive Bayesian networks. *Journal of Machine Learning Research*, 6:1–35, January 2005.

[Rob77] R. W. Robinson. Counting unlabeled acyclic digraphs. In *Lecture Notes in Mathematics, 622: Combinatorial Mathematics V*, New York, 1977. Springer Verlag.

[Rob94] Christian P. Robert. *The Bayesian Choice: a decision-theoretic motivation*. Springer, New York, 1994.

[SGS01] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. MIT Press, second edition, 2001.

[SK89] Ross D. Shachter and C. Robert Kenley. Gaussian influence diagrams. *Management Science*, 35(5):527–550, 1989.

[SSGS06] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Sprites. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, February 2006.

[Stu05] Milan Studený. Characterization of inclusion neighbourhood in terms of the essential graph. *International Journal of Approximate Reasoning*, 38:283–309, 2005.

[VS07] Jiří Vomlel and Milan Studený. Graphical and algebraic representatives of conditional independence models. In Peter Lucas, José Gamez, and Antonio Salmeron, editors, *Advances in Probabilistic Graphical Models*, volume 213. Springer, 2007. Preliminary version.

[Zha04] Nevin L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, June 2004.

Index

List of Theorems

List of Definitions

List of Figures