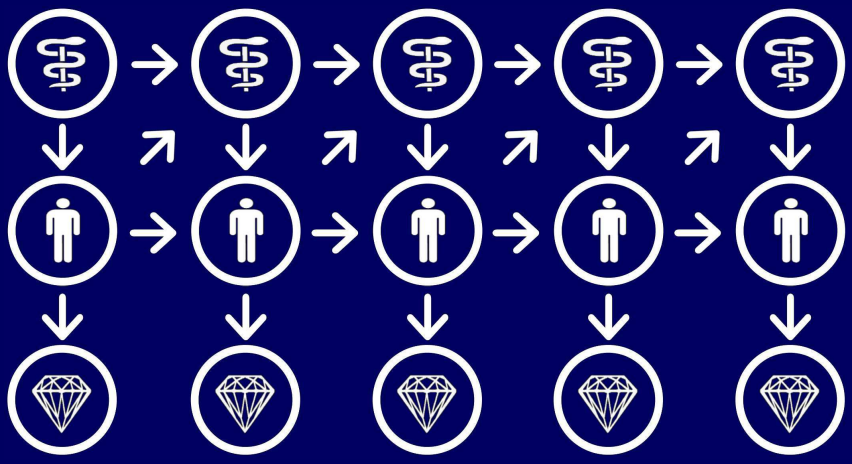
Bayesian Networks for Clinical Decision Support

A Rational Approach to Dynamic Decision-Making under Uncertainty



Marcel van Gerven

BAYESIAN NETWORKS FOR CLINICAL DECISION SUPPORT

A Rational Approach to Dynamic Decision-Making under Uncertainty

Een wetenschappelijke proeve op het gebied van de
Natuurwetenschappen, Wiskunde & Informatica

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S. C. J. J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
woensdag 5 september 2007 om 15.30 precies

door

Marcel Antonius Johannes van Gerven
geboren op 4 september 1976
te ’s-Hertogenbosch

Promotor:
Prof. dr. ir. T. P. van der Weide

Copromotor:
Dr. P. J. F. Lucas

Manuscriptcommittee:
Prof. dr. H. J. Kappen
Prof. dr. ir. A. Hasman (Universiteit van Amsterdam)
Prof. dr. P. Larrañaga (Universidad del País Vasco)
Dr. F. J. Díez Vegas (Universidad Nacional de Educación a Distancia)
Dr. B. G. Taal (Nederlands Kanker Instituut)

SIKS Dissertation Series No. 2007–12
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems, and the Institute for Computing and Information Sciences of the Radboud University Nijmegen.

This research has been funded by the Netherlands Organisation for Scientific Research (NWO) under grant number 612.066.201.

ISBN: 978-90-9021999-8

Dankwoord

Ten eerste wil ik mijn copromotor Peter Lucas bedanken dat hij mij de mogelijkheid heeft geboden om onderzoek te doen binnen het ProBayes project. Het vastleggen van een vaag begrip als medische expertise in termen van Bayesiaanse netwerken was een interessante uitdaging. Zowel Peter als mijn promotor Theo van der Weide wil ik bedanken voor de begeleiding gedurende het project. Rasa, je morele ondersteuning de afgelopen vier jaar heeft me erg geholpen. Perry, ook jij hebt me vaak het vertrouwen gegeven om door te zetten. Natuurlijk wil ik ook mijn andere collega's bedanken voor de samenwerking en gezelligheid.

Een substantieel deel van mijn tijd besteedde ik aan de ontwikkeling van een model voor de prognose van patiënten met carcinoïde tumoren. Ik werkte hierbij nauw samen met Babs Taal van het Nederlands Kanker Instituut en de lange reis naar Amsterdam was altijd de moeite waard. Soms raakten we beiden gefrustreerd door de complexiteit van de ziekte, zodat Babs wanhopig uitriep: "alles is met alles verbonden!" Op andere momenten waren we beiden onder de indruk van diezelfde complexiteit. Babs kon zich dan verwonderen over de grote hoeveelheid kennis die er in het model schuil ging. Babs, ik voelde me altijd welkom en ik wil je bedanken voor het delen van je tijd en expertise.

Mijn bezoek aan de afdeling kunstmatige intelligentie van de UNED te Madrid was een leuke en leerzame tijd. Ik wil Javier Díez, zijn familie en de mensen op de afdeling graag bedanken voor een onvergetelijke ervaring! Mijn dank gaat tevens uit naar de mensen die kritisch naar mijn proefschrift gekeken hebben: Peter Lucas, Perry Groot, Rasa Jurgelenaite, Arjen Hommersom, Stijn Hoppenbrouwers, Ward van Dieten en Esther Booltink. Ook dank ik de leden van de manuscriptcommissie: Bert Kappen, Pedro Larrañaga, Javier Díez, Arie Hasman en Babs Taal. Ameen Abu-Hanna wil ik bedanken voor zijn deelname aan mijn promotiecommissie. Tom Heskes wil ik expliciet bedanken omdat hij mij de mogelijkheid heeft geboden om me verder te ontwikkelen binnen een erg spannend vakgebied.

In het bijzonder gaat mijn dank uit naar de volgende mensen: Ron van Gerven en Ward van Dieten, bedankt voor jullie moedige optreden als mijn paranimfen. Pap en mam, bedankt voor jullie onvoorwaardelijke steun. Tot slot, lieve Es, bedankt dat je er voor me was de afgelopen jaren. Ik kijk uit naar de avonturen die voor ons liggen!

Contents

Chapter 1

Introduction

With the current rate of scientific progress and rising costs of healthcare, *Evidence-Based Medicine* (EBM) is becoming increasingly important (Woolf, 2000). EBM is the notion that medical intervention should be based on scientific evidence, thus maintaining a high level of healthcare, justifying both the interventions being made and their associated costs (Lucas and Abu-Hanna, 1999). In practice, EBM implies the need for an integration of individual clinical expertise and available external scientific evidence, where the preferences, desires, and expectations of the patient should be central to the decision-making process (Offringa et al., 2003). These requirements make adequate decision-making during clinical patient management more and more difficult for the physician. Advances in (medical) informatics in general, and artificial intelligence in particular, suggest that computers may help improve healthcare quality (Hasman and Takeda, 2003).

At present, automated support of physicians during clinical patient management using mathematically sound techniques for representing and reasoning with clinical knowledge is possible. However, the use of these techniques is difficult in practice since there are few guidelines that describe how to get from the specification of a clinical problem to a system that solves the problem using the techniques in question. Furthermore, for real-world clinical problems it is hard to obtain the required medical knowledge and/or clinical data. The subject matter of this thesis is therefore to provide techniques that allow the solution of real-world problems in clinical decision-making using mathematically sound techniques.

1.1 Medical informatics

The role of medical informatics for the improvement of healthcare quality has been recognized as early as the 1950's, when Ledley and Lusted presented their classical paper on the formal concepts underlying medical reasoning (Ledley and Lusted, 1959). Medical informatics is a broad field, ranging from healthcare ICT and electronic patient records to the development of electronic guidelines and clinical decision support systems (Shortliffe et al., 2001). In this thesis, our interest is in the

support of physicians during clinical patient management. Electronic guidelines are evidently important in this respect since they provide a formal basis for best-practice medicine, and can be developed using guideline-representation languages such as Asbru, PROforma, and GLIF (Peleg et al., 2003). An alternative to supporting clinicians in their decision-making tasks is offered by decision-support systems. In contrast to guidelines, these systems offer support adapted to the individual patient. Clinical decision support systems (CDSSs) are defined as: *active knowledge systems which use patient data to generate case-specific advice* (after (Wyatt and Spiegelhalter, 1990)). A CDSS that makes extensive use of expert knowledge is also called an *expert system* (Jackson, 1990), and we will use this term throughout.

1.2 Expert system development

As pointed out in (Patel et al., 2004), improving medical practice by understanding the thought processes that are involved in clinical reasoning has been on the research agenda for at least a century (Osler, 1906). An understanding of these thought processes is needed, since it is recognized that medical decision-making should rely more on formal techniques instead of clinical intuition (Macartney, 1988; Lucas, 1995). The use of expert systems for clinical decision support has become commonplace, with many potential benefits in terms of improving patient safety, quality of care, and efficiency in health care delivery (Lucas and van der Gaag, 1991; Coiera, 2003). Cognitive scientists have devoted much research to the analysis of problem solving strategies that are used in humans (Newell and Simon, 1972; Elstein et al., 1978). Artificial intelligence researchers have used these problem solving strategies in the development of expert systems (Schreiber et al., 2000).

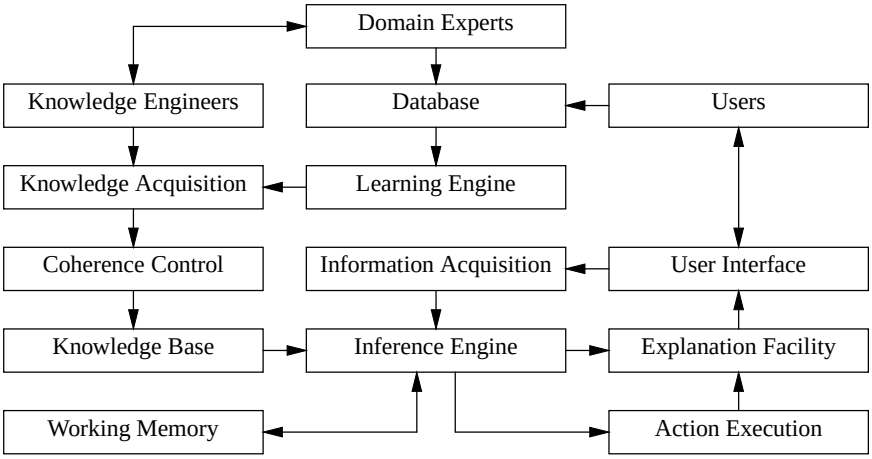


Figure 1.1: The information flow between expert system components.

Figure 1.1 depicts the components that make up an expert system and represents

the information flow between expert system components (adapted from (Castillo et al., 1997)). One of the distinguishing features of an expert system is that domain knowledge, as represented in the *knowledge base*, is separated from the problem solving strategies, as embodied by the *inference engine* (Clancey, 1983). The knowledge base is constructed from knowledge that is obtained from *domain experts* by *knowledge engineers* and/or from statistical data contained in a *database*. The task of a *learning engine* is to process the data and convert it into input to the knowledge base. The *knowledge acquisition* component combines the knowledge obtained from domain experts and statistical data. Consistency of the obtained knowledge is enforced by the *coherence control* component. Queries made by the *users* are kept in *working memory* and are processed by the *inference engine*, while the interaction between the user and the expert system takes place via a *user interface*. Queries are processed by an *information acquisition* component, and transferred to working memory. The inference engine processes the query, possibly taking actions by means of an *action execution* component, and transferring the conclusions and explanations thereof, as generated by an *explanation facility*, to the user.

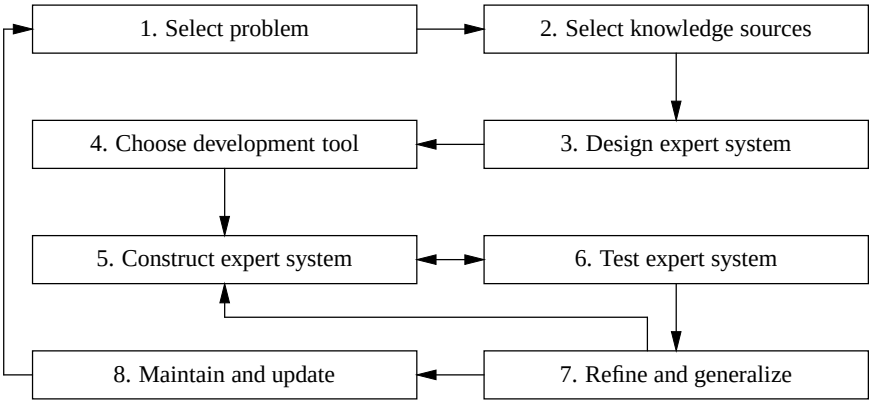


Figure 1.2: The expert system life-cycle.

Expert system development consist of a number of steps, which is shown in Fig. 1.2, and known as the expert system life-cycle (Weiss and Kulikowski, 1984; Castillo et al., 1997). As shown, the first step is to select an appropriate problem. This requires the identification of the task we wish to solve as well as the domain we wish to solve for. This first step is crucial since it may determine the ultimate failure or success of an expert system. Expert system development is a costly undertaking that requires much effort by the domain expert as well as the knowledge engineer. Therefore, one would like to be confident that system deployment yields benefits in terms of cost-reduction and/or quality-improvement. In other words, the system should be cost-effective. Once the problem has been selected, it becomes necessary to identify the possible sources from which to acquire the knowledge that may aid in expert system construction. The available knowledge sources that can be

distinguished are *domain literature*, *expert knowledge*, and *statistical data*. When adequate knowledge sources have been identified, we may proceed with the design of the components that make up the expert system. Among other tasks, this requires the design of the knowledge base and the selection of an appropriate inference algorithm. Once the expert system components have been designed, it is necessary to choose the development tool(s) that allow the construction of the actual system. Construction of the knowledge base is regarded to be the most difficult activity and vital to any expert system for clinical decision support. Nevertheless, other components are equally important for obtaining a deployable system. Once the expert system has been constructed, a constant process of testing, refinement, and maintenance ensures the continued use and improvement of the resulting system. Note that these steps are not sequential, but rather follow what is known as the spiral life cycle of system development (Boehm, 1988), where we have a continuous cycle of design, development, operation and evaluation. In the context of expert system construction, the spiral life cycle is considered to be a model of the knowledge engineering process: as construction progresses from an initial prototype to an increasingly complete system, the knowledge engineer's understanding about the domain and the domain expert's understanding of knowledge engineering practice deepens (Mahoney and Laskey, 1996). For more details about expert system development, we refer to (Weiss and Kulikowski, 1984; Turban, 1992; Castillo et al., 1997; Schreiber et al., 2000).

1.3 Traditional expert systems

Medical expert systems have been under development since the early 1960s (Warner et al., 1961), but they gained in popularity with the development of Mycin at Stanford University in the mid 1970's (Buchanan and Shortliffe, 1984). Mycin is an expert system that assists in the diagnosis and treatment of infectious diseases. It was one of the first expert systems to demonstrate impressive levels of performance, and other medical expert systems soon followed. Examples are Oncocin, a successor of Mycin that was used for protocol management in oncology (Shortliffe et al., 1981), Internist-1, and its successor QMR, for diagnosis in internal medicine (Miller and Pople, 1982; Miller et al., 1986), Puff and Centaur for determining the presence and severity of lung diseases (Aikins et al., 1983; Aikins, 1983), CasNet/Glaucoma for the diagnosis and treatment of glaucoma (Weiss et al., 1978b; Kulikowski and Weiss, 1982), PIP, a program that generates hypotheses about disease processes in patients with renal disease (Pauker et al., 1976), and Abel, for diagnosing acid-based and electrolyte disorders (Patil, 1981).

The earliest expert systems were inherently *rule-based* but it was quickly recognized that one cannot escape the need to represent knowledge of an uncertain nature. There are many examples of uncertain knowledge in the medical domain, such as symptoms that *may* be caused by a specific disease, a test indicating that a disease has

some *probability* of being present, and the *possibility* that a patient may be cured if a particular treatment is administered. Consequently, various methods for reasoning under uncertainty have been developed. At that time, it was claimed that probability theory was inadequate for representing uncertainty, based on the belief that the assignment of probabilities to events requires information that is not normally available (McCarthy and Hayes, 1969). This belief was based on the following two ideas (Jackson, 1990):

1. For a long time, the *frequentist* interpretation of probability theory, which dictates that probabilities should be computed as the long-run relative frequencies of events, has been dominant. This means that probabilities must be derived from empirical data, which is often scarce.

2. Probability theory requires the specification of a *joint probability distribution* that determines the probability for each elementary event in the domain. Systems that *had* been based on probability theory suffered from the fact that they either made unrealistically strong independence assumptions or became intractable as the number of domain variables increased (Gorry, 1973; Fryback, 1978).

The scarcity of empirical data, along with the fact that a huge number of probabilities is needed to fully specify a joint probability distribution, led to the initial dismissal of probability theory. As a result, various heuristic approaches were taken to integrate probabilistic knowledge in medical expert systems. Examples of such heuristic approaches are the certainty factor model, which was used in Mycin, and the scoring system that was employed in Internist-1 and QMR. However, as time progressed, it became evident that these approaches were ad hoc when it came to reasoning under uncertainty. This follows from the fact that certain desirable properties a measure of belief should adhere to are not fulfilled (Heckerman and Nathwani, 1992a). It has been demonstrated, for instance, that probabilistic reasoning in systems that employ such approaches is unsound if particularly strong independence assumptions between domain variables fail to hold (Horvitz et al., 1986; Horvitz and Heckerman, 1986; Lucas, 2001), and that illogical results are obtained, such as the dependence of a diagnosis on the order in which findings are entered (Cheeseman, 1985).

1.4 A rational approach

In contrast to the ad hoc approaches discussed in the previous section, there are rational arguments for using probability theory to express uncertainty. It has been shown that the axioms of probability theory follow as a logical consequence from the basic desiderata for any measure of belief (Cox, 1946; Jaynes, 2003). Furthermore, if one

does not follow the rules of probability theory, then one is willing to accept a *Dutch book*; a bet which leads to a guaranteed loss (Kyburg and Smokler, 1964).

One important development in the use of probability theory as the basis for representing uncertainty has been the shift from the frequentist interpretation of probability theory to the *subjectivist* or *Bayesian* interpretation to probability theory (named after reverend Thomas Bayes, the 19th century probability theorist), which views probabilities as a measure of belief. Under this interpretation, we may use available domain knowledge, together with available empirical data, in order to quantify our models. This allows the use of domain experts as a source of information when quantifying a probabilistic model.

Work by Pearl and colleagues in the 1980s eventually led to a breakthrough in the use of probability theory as a formalism for reasoning under uncertainty (Pearl, 1988; Lauritzen and Spiegelhalter, 1988). By taking into account conditional independence between random variables, a joint probability distribution can often be represented more compactly as a product of local conditional probability distributions. This representation takes the form of a graph whose vertices stand for the random variables that constitute the domain, and whose edges represent the independence structure that holds between random variables. Pearl devised an algorithm that allows for efficient probabilistic inference when the resulting graph forms a polytree (a directed graph that does not contain undirected cycles) (Kim and Pearl, 1983).

Bayesian networks and Markov networks are more general models, where the underlying graph is an acyclic and directed graph or an undirected graph respectively. Even though reasoning under uncertainty (also known as probabilistic inference) in these more general models is NP-hard in the exact (Cooper, 1990), as well as the approximate case (Dagum and Luby, 1993), over the years, a great deal of algorithms have been developed that perform well in practice (e.g., (Lauritzen and Spiegelhalter, 1988; Zhang and Poole, 1994)).[1] Together with the Bayesian interpretation of probability theory these models have proven to be a sound and practical framework for reasoning under uncertainty.

Although the described models allow for reasoning under uncertainty, often the focus lies not only on estimation of the posterior probability of events, but also on optimal decision-making. Decision theory (Wald, 1950) is an axiomatic theory of decision-making, which uses probability theory to represent uncertainty, allows the incorporation of interventions as made by a decision-maker, and expresses preferences among outcomes in terms of utilities. The soundness of this theory is motivated by the work of Von Neumann and Morgenstern, who have shown that, if a decision-maker adheres to five rational principles, then decision-making reduces to the maximization of expected utility (Von Neumann and Morgenstern, 1947). Early examples of research that views the decision-theoretic approach as normative for

[1] It has been shown (Kong, 1991) that, already in the 1970s, genetic linkage analysis researchers have solved special cases of probabilistic inference in Bayesian networks (Elston and Stewart, 1971).

(clinical) decision support are (Ben-Bassat et al., 1980; Charniak, 1983; Cooper, 1984; Spiegelhalter and Knill-Jones, 1984; Andreassen et al., 1987). *Influence diagrams* (Howard and Matheson, 1984a) augment Bayesian networks with decision variables and utility functions to allow for *planning* (finding optimal decision-making strategies).

We collectively refer to models that utilize a graph in order to represent (stochastic) independence as (*probabilistic*) *graphical models*. In the context of expert systems for clinical decision support, a graphical model can be regarded as the knowledge base, while the inference engine is formed by a suitable inference algorithm. In this thesis, we will mainly focus on Bayesian networks and influence diagrams.

1.5 Graphical model construction

In past years, much attention has been devoted to the development of algorithms that learn the structure and parameters of a graphical model from data (Cooper and Herskovits, 1992; Buntine, 1994; Heckerman et al., 1995). These algorithms can be distinguished into *search-and-score* based methods, which search the space of graphs, and try to optimize some measure of structure optimality (e.g., (Larrañaga et al., 1996; Chickering, 2002)), and *constraint-based* methods, which construct a graph based on conditional independence tests (e.g., (Spirtes et al., 1993; Cheng et al., 2002)). Both the graph topology and the joint probability distribution of a graphical model can be learnt from data, provided that the dataset is sufficiently large and of acceptable quality.

An alternative to this data-driven approach is to acquire knowledge from domain experts by means of protocol analysis and other knowledge elicitation techniques (Schreiber et al., 2000). This knowledge can subsequently be used to manually construct a graphical model. NESTOR (Cooper, 1984), a system for the differential diagnosis of seven diseases that cause hypercalcemia, was one of the first systems that has been developed using this approach. It uses a graph containing 100 vertices and 200 edges in order to represent causal and probabilistic knowledge. Pathfinder (Heckerman and Nathwani, 1992a,b) is an early example of a graphical model that is successfully applied in clinical practice, and is used for the diagnosis of more than 60 lymph node diseases, based on more than 130 microscopic, clinical, laboratory, immunological, and molecular-biologic features. It has also been demonstrated that existing expert systems can be successfully translated into expert systems that are based on graphical models. For example, QMR-DT (Shwe et al., 1991) is a decision-theoretic reformulation of the Internist-1/QMR expert system for diagnosis in internal medicine, which we have already encountered in Section 1.2. The structure of the bipartite graph allowed for the efficient diagnosis of multiple diseases, by means of the *quickscore* algorithm (Heckerman, 1989). Finally, detailed (anatomical) knowledge can be captured in terms of a graphical model as has been convincingly demonstrated

by MUNIN (Olesen et al., 1989), a probabilistic network for the diagnosis of neuromuscular disorders.

Manual construction of a graphical model for clinical decision support requires the specification of the independence structure between domain variables as well as the estimation of a large number of parameters, and is well-known to be non-trivial (Druzdzel et al., 1995; van der Gaag and Helsper, 2002). Clinical experts are often unable to articulate the knowledge needed for constructing an expert system (Johnson et al., 1981) and parameter estimation by experts suffers from various kinds of cognitive biases (Kahneman et al., 1982) as demonstrated for the medical domain in (Berwick et al., 1981).[2]

In practice, datasets for realistic domains can be small and of poor quality, such that learning a graphical model from data yields unsatisfactory results (Wu et al., 2001; van Gerven and Lucas, 2004b). In those cases, the only remaining options are either to learn a restricted model from data by making strong assumptions about model structure (Friedman et al., 1997) or to construct the model by hand using available expert knowledge.

1.6 Aim of this thesis

As described in Section 1.4, graphical models can serve as normative models for decision making under uncertainty. Given that data is scarce for many medical domains, we are faced with either learning graphical models from small datasets, or manual construction based on available expert knowledge. There are few guidelines that take into account all aspects of graphical model construction.[3] Given the complexity of this task, often strong assumptions are made with respect to the structure and/or parameters of the graphical model, such as assuming mutual exclusiveness of diseases in Pathfinder and independence of findings given diseases in QMR-DT. These assumptions may not always be warranted for the problem at hand, affecting both the realism and usefulness of the resulting systems. As a result, few graphical models for clinical decision support have seen a successful implementation in practice. The main objective of this thesis is therefore:

> *To provide techniques for the construction of graphical models for clinical decision support that are realistic enough to be applied in practice, where the focus on real-world problems entails that the model is constructed from available expert knowledge or a limited amount of data.*

[2] The difficulty of acquiring accurate domain knowledge is commonly known as the *knowledge-acquisition bottleneck* (Cullen and Bryman, 1988).

[3] See (Abramson and Ng, 1993; Pradhan et al., 1994; Mahoney and Laskey, 1996; Druzdzel et al., 1999) for some exceptions.

1.7 Thesis outline

In order to achieve the objective of Section 1.6, this thesis proceeds as follows.

Chapter 2

In Chapter 2, we deal with the necessary preliminaries. We describe probability theory and graph theory as the mathematical foundations for graphical models. Subsequently, we focus on (inference in) probabilistic graphical models and (solving) influence diagrams.

Chapter 3

Construction of graphical models for clinical decision support often proceeds in an ad hoc fashion, which implies the need for a more principled approach. In Chapter 3 we develop such an approach by making a connection between the description of clinical tasks in terms of problem solving and particular choices of Bayesian network designs.

Chapter 4

The manual construction of a graphical model from available expert knowledge is a difficult and time-consuming task. Therefore, any tool that reduces model construction efforts is welcomed. In Chapter 4, we focus on causal independence models, which use deterministic interaction functions in order to reduce the number of parameters that need to be specified. We provide a qualitative analysis of the independence of causal influence, which allows us to determine the qualitative properties of a causal independence model with a given interaction function without the need to specify the probabilistic parameters in advance.

Chapter 5

Clinical decision support systems often require that a decision-making strategy is represented as part of the system, and an important goal is to automatically find an optimal strategy for decision problems that are characterized by uncertainty and which evolve over time. Chapter 5 proceeds with the development of a framework for dynamic decision making under uncertainty and the construction of a number of algorithms that approximate optimal strategies. The usefulness of the approach is demonstrated with the solution of a dynamic decision problem in oncology.

Chapter 6

In Chapter 6, we describe the construction and validation of a realistic dynamic Bayesian network for clinical decision support, where we focus on prognosis of patients that suffer from a low-grade midgut carcinoid tumors. This model has been created in collaboration with an expert physician at the Netherlands Cancer Institute, and is one of the largest dynamic Bayesian networks for clinical decision support to date.

Chapter 7

Chapter 7 focuses on Bayesian networks that are used for the purpose of probabilistic classification and which are learned from a limited amount of data. Three different techniques are examined and validated using clinical data:

1. The *maximum mutual information algorithm*, which learns a probabilistic classifier based on information-theoretic principles.

2. The *decomposed tensor classifier*, which uses a rank-K tensor approximation for the purpose of classification.

3. The *noisy-threshold classifier*, which employs a particular causal independence model as a probabilistic classifier.

Chapter 8

This thesis is concluded in Chapter 8 with a summary of the scientific contributions, a discussion of the strengths and limitations of the described research, and a general conclusion about the subject matter of this thesis.

Chapter 2

Preliminaries

In this chapter, we deal with the necessary preliminaries. We describe the mathematical foundations of graphical models and the algorithms that are used for probabilistic inference and the solution of decision problems.

2.1 Probability theory

As discussed, probability theory is used in order to represent and reason with uncertainty.[1] The measurement of uncertainty proceeds by defining a *sample space* Ω, which includes the mutually exhaustive and collectively exhaustive outcomes of an experiment, and a collection $\mathcal{A}$ of subsets of Ω that adheres to the following properties:

1. $\varnothing \in \mathcal{A}$;
2. if $A_1, A_2, \ldots \in \mathcal{A}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$;
3. if $A \in \mathcal{A}$ then $\bar{A} \in \mathcal{A}$, where $\bar{A}$ denotes the complement of A.

The set $\mathcal{A}$ is known as a σ-field and its elements are called *events*. The aim is to express the degree of uncertainty about events by means of a *probability measure*.

Definition 2.1. *A* probability measure $P\colon \mathcal{A} \to [0,1]$ *on* $(\Omega, \mathcal{A})$*, defining the* probability space $(\Omega, \mathcal{A}, P)$*, is a function that satisfies the following axioms:*

1. $P(\varnothing) = 0$.
2. $P(\Omega) = 1$.
3. *For any infinite sequence,* $A_1, A_2, \ldots \in \mathcal{A}$*, it holds that*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)\,.$$

[1] For a more complete treatment of probability theory we refer to (Grimmett and Stirzaker, 1992).

In general, we find it convenient to work with random variables, which describe experimental outcome in terms of real numbers for some probability space.

Definition 2.2. *A* random variable *is a function* $X\colon \Omega \to \mathbb{R}$ *such that* $\{\omega \in \Omega\colon X(\omega) \leq x\} \in \mathcal{A}$ *for each* $x \in \mathbb{R}$.

Sometimes, it is necessary to take time into account and to determine how a sequence of random variables evolves over time. We call this sequence a *random process*.

Definition 2.3. *A* random process X *is a family* $\{X(t)\colon t \in T\}$ *of random variables that take values in* Ω_X *and are indexed by some set* T. *If* $T \subseteq \mathbb{N}$, *then we call the process a* discrete-time process.

Each random variable has an associated distribution function.

Definition 2.4. *The* distribution function $F\colon \mathbb{R} \to [0,1]$ *of a random variable* X *is defined as* $F(x) = P(X \leq x)$.

In this thesis, we will mainly deal with *discrete* random variables (whose values are restricted to a countable subset $\Omega_X = \{x_1, \ldots, x_n\}$ of $\mathbb{R}$) and to a lesser degree with *continuous* random variables (whose values are given by $\mathbb{R}$). Uppercase letters X, Y, Z are used to denote random variables, and boldface uppercase letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are used to denote sets or vectors of random variables. We use lowercase letters x, y, z to denote values that random variables may take on and use $\mathbf{x} = (x_1, \ldots, x_n)$ for an element in the sample space $\Omega_{\mathbf{X}} = \Omega_{X_1} \times \cdots \times \Omega_{X_n}$ for a vector $\mathbf{X} = (X_1, \ldots, X_n)$ of random variables. For a discrete random variable, we define its *probability mass function* as follows.

Definition 2.5. *The* probability mass function *of a discrete random variables* X *(loosely referred to as the probability distribution of* X*) is the function* $f\colon \mathbb{R} \to [0,1]$ *such that* $f(x) = P(X = x)$.

A joint probability distribution is then defined as follows.

Definition 2.6. *The* joint probability distribution *(JPD) of a vector* $\mathbf{X} = (X_1, \ldots, X_n)$ *of discrete random variables is the function* $f\colon \mathbb{R}^n \to [0,1]$ *such that* $f(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$.

We abbreviate $P(\mathbf{X} = \mathbf{x})$ by $P(\mathbf{x})$, and also write it as $P(X_1 = x_1, \ldots, X_n = x_n)$, which is abbreviated by $P(x_1, \ldots, x_n)$. The *marginal probability distribution* for a random variable X_i can be obtained from the JPD as follows:

$$P(X_i = x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} P(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)\,.$$

If knowledge is obtained about the occurrence of some event then this can modify the probabilities that other events occur. This is captured by the notion of *conditional probability*.

Definition 2.7. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two disjoint subsets of random variables with* $P(\mathbf{y}) > 0$*. The* conditional probability distribution *(CPD) of* $\mathbf{X}$ *given that* $\mathbf{Y}=\mathbf{y}$ *is given by*

$$P(\mathbf{X}=\mathbf{x} \mid \mathbf{Y}=\mathbf{y}) = P(\mathbf{x} \mid \mathbf{y}) = \frac{P(\mathbf{x},\mathbf{y})}{P(\mathbf{y})}$$

which stands for the probability of observing $\mathbf{X}=\mathbf{x}$ *given evidence* $\mathbf{Y}=\mathbf{y}$*.*

Since conditional probabilities play a central role in the Bayesian interpretation of probability theory, they are used to define a joint probability distribution $P(\mathbf{X},\mathbf{Y})$, as in:

$$P(\mathbf{X},\mathbf{Y}) = P(\mathbf{X} \mid \mathbf{Y})P(\mathbf{Y}) = P(\mathbf{Y} \mid \mathbf{X})P(\mathbf{X})\,.$$

By rearranging terms, Bayes' rule follows immediately from this definition:

$$P(\mathbf{X} \mid \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{X})P(\mathbf{X})}{P(\mathbf{Y})} \tag{2.1}$$

for $P(\mathbf{Y}) > 0$. Interpreting $\mathbf{X}$ as a hypothesis and $\mathbf{Y}$ as the available evidence, Bayes' rule allows us to update our prior beliefs about $\mathbf{X}$ as evidence $\mathbf{Y}$ becomes available. For the Bayesian subjectivist, Eq. (2.1) is a normative rule for belief updating in the light of available evidence (Pearl, 1988). A *posterior* belief in the hypothesis $P(\mathbf{X} \mid \mathbf{Y})$ is obtained by multiplying the *prior* belief in the hypothesis $P(\mathbf{X})$ with the *likelihood* $P(\mathbf{Y} \mid \mathbf{X})$ of the hypothesis given the evidence and by normalizing this quantity using the *evidence* $P(\mathbf{Y})$.

2.2 Graph theory

Decision-theoretic models, as used in this thesis, rely heavily on graph-theoretical concepts. In this section we define the necessary concepts. For additional background material, we refer the reader to (Diestel, 2000).

Definition 2.8. *A* graph *is a pair* $G = (V, E)$*, where* V *is a finite set of* nodes *and* $E \subseteq V \times V$ *a set of* edges*. We also use* $V(G)$ *and* $E(G)$ *to denote nodes and edges of* G*.*

We say that an edge is *undirected* if $\{(v,v'),(v',v)\} \subseteq E(G)$, and we say that an edge is *directed* if $(v,v') \in E(G) \Rightarrow (v',v) \notin E(G)$. We define the following sets of nodes:

- We call $\nu_G(v) = \{v' \mid \{(v,v'),(v',v)\} \subseteq E(G)\}$ the *neighbors* of v and $|\nu_G(v)|$ the *degree* of v.
- We call $\rho_G(v) = \{v' \mid (v',v) \notin E(G), (v',v) \in E(G)\}$ the *children* of v and $|\rho_G(v)|$ the *out-degree* of v.

- We call $\pi_G(v) = \{v' \mid (v', v) \in E(G), (v', v) \notin E(G)\}$ the *parents* of v and $|\pi_G(v)|$ the *in-degree* of v. The *family* of v is given by $fa_G(v) = \{v\} \cup \pi_G(v)$.

We also define the following node sequences.

Definition 2.9. *A* route *in G, with* length $n-1$*, is a sequence $v_1, \ldots, v_n$ of nodes such that $(v_i, v_{i+1}) \in E(G)$ or $(v_{i+1}, v_i) \in E(G)$ for $1 \leq i < n$.*

Definition 2.10. *A route is called a* path *if $(v_{i+1}, v_i) \notin E(G)$ for $1 \leq i < n$.*

The *ancestors* $an_G(v)$ of a node v are those nodes v' for which there is a path between v' and v, and the *descendants* $de_G(v)$ of a node v are those nodes v' for which there is a path between v and v'. We call a route $v_1, \ldots, v_n$ of distinct nodes such that $v_1 = v_n$ a *loop*, and a path $v_1, \ldots, v_n$ of distinct nodes such that $v_1 = v_n$ a *cycle*. A *chord* of a loop is an edge between two nodes in a loop that is not contained in the loop. A graph is *connected* if there is a route from v to v' for all $v, v' \in V(G)$ with $v \neq v'$. A *directed* (*undirected*) graph consists only of directed (undirected) edges, and an *acyclic* graph contains no (directed or undirected) cycles. For a directed graph G, we also use the term *arcs* $A(G)$ to refer to edges $E(G)$. Some important classes of graphs are the following.

Definition 2.11. *An* acyclic directed graph *(ADG) is a directed graph that is acyclic.*

Definition 2.12. *A* tree *is a connected acyclic undirected graph, where nodes of degree one are called* leafs *and non-leaf nodes are called* internal nodes*.*

Definition 2.13. *A* rooted tree *is an acyclic directed graph with the edges pointing away from a distinguished node, called the* root *of the tree.*

Definition 2.14. *A* polytree *is a directed acyclic graph that has no undirected cycles when we drop the directions of the edges in the graph.*

Definition 2.15. *A* moral graph G^m *is the graph that is obtained from an acyclic directed graph G, by linking the parents of each node in G by edges, and by dropping the orientation of arcs in the graph.*

Definition 2.16. *A* triangulated graph *is an undirected graph such that all loops of length four or more have at least one chord.*

Given an undirected graph G, a *clique* of G is a set of nodes $\mathbf{C} \subseteq V(G)$ that is *complete* (all pairs of nodes in $\mathbf{C}$ are neighbors in G) and *maximal* ($\mathbf{C} \cup \{V\}$ with $V \in V(G) \setminus \mathbf{C}$ is not complete). An important property of any triangulated graph is the following.

Definition 2.17. *An ordering $(\mathbf{C}_1, \ldots, \mathbf{C}_m)$ of the cliques in G satisfies the* running intersection property *if $\mathbf{C}_i \cap (\mathbf{C}_1 \cup \cdots \cup \mathbf{C}_{i-1}) \subseteq \mathbf{C}_j$ for all $1 \leq i \leq n$ where $1 \leq j \leq i-1$.*

We may use the cliques of a graph in the definition of a secondary graph as follows.

Definition 2.18. *A* clique graph *of a graph G with cliques $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_m\}$ such that $V(G) = \mathbf{C}_1 \cup \cdots \cup \mathbf{C}_m$ is an undirected graph G' with $V(G') = \mathcal{C}$, such that $(\mathbf{C}_i, \mathbf{C}_j) \in E(G') \Leftrightarrow \mathbf{C}_i \cap \mathbf{C}_j \neq \varnothing$.*

For any triangulated graph, we may construct a clique graph known as a *junction tree* (Jensen, 1988), which is defined as follows.

Definition 2.19. *A* junction tree *is a clique graph whose nodes and edges form a tree that satisfies the running intersection property.*

The junction tree is prominent in probabilistic inference, as is demonstrated later on.

2.3 Bayesian networks and Markov networks

The connection between probability theory and graph theory is made by using graphs to represent *(conditional) independence* relations between random variables (Dawid, 1979).

Definition 2.20. *Let $\mathbf{X}$ be a set of random variables with JPD $P(\mathbf{X})$. Let $\mathbf{U}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{X}$ be disjoint subsets of $\mathbf{X}$. Then, $\mathbf{U}$ is said to be* conditionally independent *of $\mathbf{Y}$ given $\mathbf{Z}$, denoted by $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z}$, iff*

$$P(\mathbf{U} \mid \mathbf{Y}, \mathbf{Z}) = P(\mathbf{U} \mid \mathbf{Z})$$

whenever $P(\mathbf{Y}, \mathbf{Z}) > 0$.

This independence relation can be characterized by axioms, such as the following:

1. *Symmetry*:
   $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Leftrightarrow \mathbf{Y} \perp\!\!\!\perp_P \mathbf{U} \mid \mathbf{Z}$
2. *Decomposition*:
   $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{V} \mid \mathbf{Z} \Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Z}$
3. *Weak Union*:
   $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{V} \mid \mathbf{Z} \Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{V}$
4. *Contraction*:
   $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Y} \cup \mathbf{Z} \Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{V} \mid \mathbf{Z}$
5. *Intersection*:
   $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \cup \mathbf{V} \wedge \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Z} \cup \mathbf{Y} \Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{V} \mid \mathbf{Z}$ *iff* $\forall_{\mathbf{u}} \colon P(\mathbf{u}) > 0$

If the first four axioms are satisfied then the independence relation is called a *semi-graphoid* and if the fifth condition is satisfied as well, then the independence relation is called a *graphoid*. These axioms allow the derivation of other interesting lemmas such as the following.

Lemma 2.1. $\mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{U} \cup \mathbf{Z} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Y} \Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Z}$

Proof. We derive

$$\begin{aligned} \mathbf{U} \cup \mathbf{Z} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Y} &\Rightarrow \mathbf{V} \perp\!\!\!\perp_P \mathbf{U} \cup \mathbf{Z} \mid \mathbf{Y} && (\textit{symmetry}) \\ &\Rightarrow \mathbf{V} \perp\!\!\!\perp_P \mathbf{U} \mid \mathbf{Y} \cup \mathbf{Z} && (\textit{weak union}) \\ &\Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Y} \cup \mathbf{Z} && (\textit{symmetry}) \end{aligned}$$

and use this result to obtain

$$\begin{aligned} \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \wedge \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Y} \cup \mathbf{Z} &\Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \cup \mathbf{V} \mid \mathbf{Z} && (\textit{contraction}) \\ &\Rightarrow \mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{Z} && (\textit{decomposition}) \end{aligned}$$

which concludes the proof. □

The semi-graphoid axioms have been proposed as basic to the definition of informational dependency (Pearl and Paz, 1985) and although the semi-graphoid axioms allow for the derivation of many other interesting independence relations, Studený has shown that the independence relation is not finitely axiomatizable (Studený, 1989, 1992).

One way to represent a set of independence relations is by means of a graph G. Let $\mathbf{X}$ denote a set of random variables. We assume that there is a one-to-one correspondence between variables in $\mathbf{X}$ and nodes in $V(G)$, and write $G = (\mathbf{X}, E)$ when this correspondence is established. We use the notation $\mathbf{U} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}$ to denote the *separation* of $\mathbf{U}$ and $\mathbf{Y}$ by $\mathbf{Z}$ in G, where $\mathbf{U}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint subsets of $\mathbf{X}$. The notion of separation depends on the type of the graph G. If G is *undirected* then separation is intuitively defined as the blocking of each path between $\mathbf{U}$ and $\mathbf{Y}$ by $\mathbf{Z}$. In *directed* graphs on the other hand, separation is referred to as *d-separation*, which is defined as follows.

Definition 2.21. $\mathbf{Z}$ d-separates $\mathbf{U}$ *and* $\mathbf{Y}$, *denoted by* $\mathbf{U} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}$, *if for every route* $X, \ldots, Y$ *in* G, *with* $X \in \mathbf{U}$ *and* $Y \in \mathbf{Y}$, *there is a vertex* Z, *such that*

- *there are no two arcs in the route that point towards* Z, *and* $Z \in \mathbf{Z}$, *or*
- *there are two arcs in the route that point towards* Z, *and neither* Z *nor any of its descendants are in* $\mathbf{Z}$.

Let $\mathbf{X}$ denote a set of random variables and $\mathbf{U}$, $\mathbf{Y}$, $\mathbf{Z}$ disjoint subsets of $\mathbf{X}$. We say that a graph G, with $V(G) = \mathbf{X}$, is a *dependency map* (D-map) of $P(\mathbf{X})$ iff

$$\forall_{\mathbf{U},\mathbf{Y},\mathbf{Z}} : \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{U} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z} \,.$$

Likewise, we say that G is an *independency map* (I-map) of $P(\mathbf{X})$ iff

$$\forall_{\mathbf{U},\mathbf{Y},\mathbf{Z}} : \mathbf{U} \perp\!\!\!\perp_P \mathbf{Y} \mid \mathbf{Z} \Leftarrow \mathbf{U} \perp\!\!\!\perp_G \mathbf{Y} \mid \mathbf{Z}\,.$$

Finally, if G is both a D-map and an I-map of $P(\mathbf{X})$, then we say that G is a *perfect map* of $P(\mathbf{X})$. Most probability models have no perfect map representation. However, we can use I-maps to represent independence statements for any probability model. We say that an I-map is *minimal* if with the removal of any edge from the graph, the I-map property ceases to hold, i.e., the graph represents the largest possible number of independence statements. We now arrive at the following two definitions.

Definition 2.22. *A* Markov network $\mathcal{M} = (G, \Psi)$ *is a pair, where* G *is an* undirected graph *with nodes corresponding to a set of random variables* $\mathbf{X}$*, representing a minimal I-map of* $P(\mathbf{X})$*, and* $\Psi = \{\psi_i(\mathbf{c}_i) \colon \mathbf{C}_i \in \mathcal{C}\}$ *is a set of non-negative functions, known as* potentials*, defined for the cliques* $\mathcal{C}$ *of* G*.*

Definition 2.23. *A* Bayesian network $\mathcal{B} = (G, P)$ *is a pair, where* G *is an* acyclic directed graph *with nodes corresponding to a set of random variables* $\mathbf{X}$*, representing a minimal I-map of* $P(\mathbf{X})$*, and* $P = \{P(x \mid \pi_x) \colon X \in \mathbf{X}\}$ *is a set of conditional probability distributions, defined for random variables* $\mathbf{X}$ *in* G*.*

Markov networks and Bayesian networks capture different sets of independence relations, leading to different representations of a JPD in terms of a product of local factors. For a Markov network, with cliques $\mathcal{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_m\}$, the JPD factorizes as follows:

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^{m} \psi_i(\mathbf{c}_i) \tag{2.2}$$

where $Z = \sum_{\mathbf{x}} \prod_{i=1}^{m} \psi_i(\mathbf{c}_i)$ is the *partition function*, which acts as a normalizing constant. Unfortunately, it is difficult to quantify the potentials ψ_i in terms of quantities that are meaningful for a domain expert. Only if the Markov network is *decomposable* (if G is triangulated) do we have a meaningful factorization (Pearl, 1988). Let $(\mathbf{C}_1, \ldots, \mathbf{C}_m)$ denote an ordering of the cliques of G that satisfies the running intersection property of Def. 2.17, and define $\mathbf{S}_i = \mathbf{C}_i \cap (\mathbf{C}_1 \cup \cdots \cup \mathbf{C}_{i-1})$ and $\mathbf{R}_i = \mathbf{C}_i \setminus \mathbf{R}_i$. Then, a decomposable Markov network with $\Psi = \{P(\mathbf{r}_i \mid \mathbf{s}_i) \colon \mathbf{C}_i \in \mathcal{C}\}$ can be factorized as

$$P(\mathbf{x}) = \prod_{i=1}^{m} P(\mathbf{r}_i \mid \mathbf{s}_i)$$

which allows a specification in terms of conditional probability distributions. For a Bayesian network, we immediately obtain such a meaningful interpretation, since for a set of random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$, the JPD factorizes as:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i \mid \boldsymbol{\pi}_i) \tag{2.3}$$

where $\boldsymbol{\pi}_i \in \Omega_{\pi(X_i)}$ denotes a realization of the parents of X_i that is compatible with $\mathbf{x}$. For convenience, when dealing with a Bayesian network (G, P) for a set of random variables $\mathbf{X}$, we often omit G from our notation when clear from context, and call P the JPD of $\mathbf{X}$.

2.4 Probabilistic inference

Probabilistic graphical models generally reduce the number of free parameters that are needed to specify a JPD and allow for efficient probabilistic inference. Using probabilistic inference, various queries may be answered, such as conditional and marginal probabilities of a set of random variables $\mathbf{U} \subseteq \mathbf{X}$ given evidence $\mathbf{E} \subseteq \mathbf{X}$, $\mathbf{U} \cap \mathbf{E} = \varnothing$, the maximum a posteriori (MAP) hypothesis, which is the most probable instantiation of random variables given partial evidence: $\mathbf{E} \subset \mathbf{X} \setminus \mathbf{U}$, or the most probable explanation (MPE), which is the most probable instantiation of random variables given complete evidence: $\mathbf{E} = \mathbf{X} \setminus \mathbf{U}$.

Over the years, various *exact* and *approximate* inference methods have been developed, where exact methods typically require the graph structure underlying a graphical model to be sufficiently sparse. As mentioned, *belief propagation* (Pearl, 1988) is exact only when the graph is a polytree. The *junction tree algorithm* (Lauritzen and Spiegelhalter, 1988), in contrast, allows for the computation of conditional and marginal probabilities for arbitrary graphs. First, a junction tree and associated potentials are constructed from a Bayesian network (G, P) for a set of random variables $\mathbf{X}$ as in Algorithm 2.1.

Algorithm 2.1 Junction tree construction.

input: ADG G, conditional distributions P, random variables $\mathbf{X}$.
construct the moral graph G^m from G (moralization)
construct a triangulated graph G' by adding edges to G^m (triangulation)
construct a junction tree $T = (\mathcal{C}, \mathcal{E})$ from G' with cliques $\mathcal{C} = \{C_1, \ldots, C_m\}$
for $X \in \mathbf{X}$ **do**
  assign X to a clique $\mathbf{C} \in \mathcal{C}$ that contains X
end for
let $\mathbf{X}_i$ denote the set of random variables assigned to $\mathbf{C}_i$
for $i = 1$ to m **do**
  define $\psi_i(\mathbf{c}_i) = \prod_{X \in \mathbf{X}_i} P(x \mid \pi_X)$
end for
return T, $\Psi = \{\psi_i(\mathbf{c}_i) \colon \mathbf{C}_i \in \mathcal{C}\}$

Inference proceeds by means of evidence absorption and message passing in the junction tree, and produces posteriors for random variables $\mathbf{X} \setminus \mathbf{E}$, as is described by Algorithm 2.2.

For exact algorithms, continuous distributions are often restricted to be conditional Gaussian distributions (Lauritzen and Wermuth, 1989). Although arbitrary

Algorithm 2.2 Junction tree inference.

input: junction tree T, set of potentials Ψ, evidence $\mathbf{e}$.
for $i = 1$ to m **do**
  absorb evidence by setting $\psi_i(\mathbf{c}_i) = 0$ if $\mathbf{c}$ is inconsistent with $\mathbf{e}$
end for
construct separators $\mathbf{S}_{ij} = \mathbf{C}_i \cap \mathbf{C}_j$ for all neighbors $\mathbf{C}_i$ and $\mathbf{C}_j$ in $\mathcal{C}$
repeat
  for $i = 1$ to m **do**
    for all neighbors $\mathbf{C}_j \in \nu(\mathbf{C}_i)$ **do**
      if messages by $\nu(\mathbf{C}_i) \setminus \{\mathbf{C}_j\}$ have been received by $\mathbf{C}_i$ then send $\mathbf{C}_j$ the message $M_{ij}(\mathbf{s}_{ij}) = \sum_{\mathbf{c}_i \setminus \mathbf{s}_{ij}} \psi_i(\mathbf{c}_i) \prod_{k \neq j} M_{ki}(\mathbf{s}_{ki})$.
    end for
  end for
until all messages have been computed
for $i = 1$ to m **do**
  calculate $P(\mathbf{c}_i) = \psi_{\mathbf{C}_i}(\mathbf{c}_i) \prod_k M_{ki}(\mathbf{s}_{ik})$
end for
for $X \in \mathbf{X} \setminus \mathbf{E}$ **do**
  compute $P(x \mid \mathbf{e}) = \sum_{\mathbf{c}_k \setminus \{x\}} P(\mathbf{c}_k)$ for the smallest clique $\mathbf{C}_k$ with $X \in \mathbf{C}_k$
end for
return $\{P(x \mid \mathbf{e}) \colon X \in \mathbf{X} \setminus \mathbf{E}\}$

continuous distributions can be represented by means of techniques such as *mixtures of Gaussians* (Shenoy, 2006) or *mixtures of truncated exponentials* (Cobb et al., 2006), this is computationally more expensive. When extensive use is being made of arbitrary continuous distributions, or if the graph structure becomes too dense, then one may resort to various *deterministic* or *stochastic* approximate inference algorithms. Examples of deterministic approximate inference methods are *loopy belief propagation* (Murphy et al., 1999), which is the application of belief propagation to acyclic directed graphs, and *variational methods* (Jordan et al., 1999), which transform a probabilistic model into a less complex model in order to compute bounds on probabilities of interest. Examples of stochastic approximate inference methods are *importance sampling* (Geweke, 1989; Yuan and Druzdzel, 2005) and *Gibbs sampling* (Geman and Geman, 1984; Pearl, 1987).

In case we are dealing with stochastic (decision) processes, we use a so-called dynamic Bayesian network (DBN) in order to represent the temporal structure of the problem (Murphy, 2002) (as explained in detail in Chapters 3 and 5). In this case, probabilistic queries of interest may be *prediction* (computing posterior probabilities of unobserved random variables at some future time given evidence), *filtering* (computing posterior probabilities of unobserved random variables at the current time given evidence), *smoothing* (estimating posterior probabilities of unobserved random variables at some past time given evidence), or *Viterbi decoding* (computing the most probable explanation at the current time given evidence). These queries are answered

by means of specialized inference algorithms such as the exact *interface algorithm* (Murphy, 2002), which uses the junction tree algorithm as a subroutine, or approximate *particle filtering*, which comprises the group of sequential Monte Carlo methods for dynamic state estimation. Particle filtering in a DBN amounts to sampling particles $\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N)}$ with associated weights $w^{(1)}, \ldots, w^N$ from a distribution $P(\mathbf{X}(t) \mid \mathbf{X}(t\text{-}1) = \mathbf{x}, \mathbf{Y}(t) = \mathbf{y})$ that represent our belief state about $\mathbf{X}(t)$. Algorithm 2.3 shows how a particle is sampled from this distribution (adapted from (Koller and Lerner, 2001)).

Algorithm 2.3 Particle sampling in a dynamic Bayesian network.

input: previous state $\mathbf{x}$, current observations $\mathbf{y}$
$w = 1$, $\mathbf{x} = \varnothing$
let $X'_1, \ldots, X'_n$ be an ancestral ordering such that parents occur before children
for $i = 1$ to n **do**
  set $\mathbf{z} \in \Omega_{\pi(\mathbf{X}'_i)}$ compatible with $\mathbf{x}$ and $\mathbf{x}'$
  if $X'_i \notin \mathbf{Y}$ **then**
    sample x'_i from $P(X'_i \mid \pi(X'_i) = \mathbf{z})$
  else
    set x'_i to its value in $\mathbf{y}$
    set $w = w \cdot P(x'_i \mid \pi(X'_i) = \mathbf{z})$
  end if
end for
return $(\mathbf{x}', w)$

2.5 Influence diagrams

Although Bayesian networks allow for efficient probabilistic inference, their semantics does not incorporate the notions of decision-making and outcome preference. Influence diagrams (Howard and Matheson, 1984a) are designed exactly for this purpose and are convenient for the solution of (medical) decision problems (Owens et al., 1997; Nease and Owens, 1997). Influence diagrams are defined as follows.

Definition 2.24. *An* influence diagram *(ID) is a tuple* $(\mathbf{C}, \mathbf{D}, U, \mathbf{A}, P)$ *such that* $G = (\mathbf{N}, \mathbf{A})$ *is an ADG with nodes* $\mathbf{N} = \mathbf{C} \cup \mathbf{D} \cup \{U\}$ *and arcs* $\mathbf{A}$*, with*

- $\mathbf{C}$ *a set of random variables, which we refer to as* chance variables*,*
- $\mathbf{D}$ *a set of* decision variables *such that*
  - *each decision variable* $D \in \mathbf{D}$ *can take on a value from a set of choices* Ω_D*,*
  - *there is a total ordering* $\prec$ *of the decision variables implied by a path in* G *that contains all* $D \in \mathbf{D}$*,*

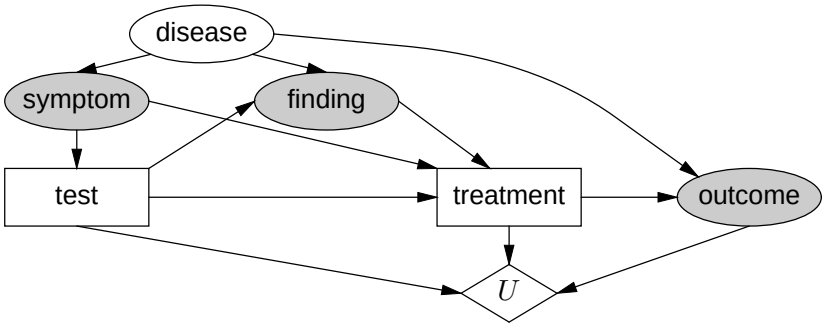


Figure 2.1: An influence diagram for patient treatment, where observable random variables are shaded and decision variables are represented by rectangles. Based on a symptom, we may choose whether or not to test a patient. This can result in a finding that some disease is present. Our treatment decision is based on the finding in conjunction with the symptom and the previous test decision. The actual outcome is determined by the treatment together with the state of disease. The test and treatment both have associated costs, and there is a utility associated with each of the different outcomes, as captured by the utility function U.

  - *if* $D' \prec D$ *then* $\pi(D') \subseteq \pi(D)$,

- U *a* utility function $U : \Omega_{\pi(U)} \rightarrow \mathbb{R}$ *such that* $\rho(U) = \varnothing$,

and where $P = \{P(C \mid \pi(C))midC \in \mathbf{C}\}$.

Figure 2.1 shows an example of an influence diagram. Chance nodes (depicted as ellipses) represent the stochastic component of the model. If $(X, C) \in A(G)$ then the conditional probability distribution associated with C may be influenced by X as in a Bayesian network. Decision nodes D (depicted as rectangles) represent the actions that may be performed by a decision maker. If $(X, D) \in A(G)$ then X represents information that is available to the decision maker prior to deciding upon D. X is also known as an *informational predecessor* and we normally depict only those informational predecessors of D that are not yet informational predecessors of decision nodes $D' \prec D$. The assumption that all past information is relevant to decision-making is known as the *no forgetting* principle. The utility node U (also known as value node and depicted as a diamond) represents the utility of being in a certain state, as defined by configurations of chance and decision variables. If $(X, U) \in A(G)$ then X takes part in the specification of U such that $U : \Omega_{\pi(U)} \rightarrow \mathbb{R}$. It is assumed that U has no children in the graph. Formally, the set P is not to be interpreted as a set of conditional probability distributions, since chance nodes may have decision nodes as parents, which do not normally have associated probability distributions. Each $P(C \mid \pi(C))$ is rather a family that specifies for each configuration $\mathbf{d} \in \Omega_{\mathbf{D}}$ a conditional probability distribution (Lauritzen and Nilsson, 2001).

We use the notation

$$P(\mathbf{C}\colon \mathbf{D}=\mathbf{d}) = \prod_{C\in\mathbf{C}} P(C \mid \pi(C)) \tag{2.4}$$

to represent the probability distribution of $\mathbf{C}$ given that the decision maker has set $\mathbf{D}$ equal to $\mathbf{d}$ (Cowell et al., 1999).

The ultimate goal of an influence diagram is to find the optimal decision making strategy for a given decision problem. A *stochastic policy* for decisions $D \in \mathbf{D}$ is defined as a probability distribution $P(D \mid \pi(D))$ that maps configurations of $\pi(D)$ to a distribution over alternatives for D. If $P(D \mid \pi(D))$ is degenerate then we say that the policy is *deterministic*. Let $\mathbf{V}$ denote $\mathbf{C} \cup \mathbf{D}$. A *strategy* is a set of policies $\Delta = \{P(D \mid \pi(D))\colon D \in \mathbf{D}\}$ which induces the following joint distribution over the variables in $\mathbf{V}$:

$$P_\Delta(\mathbf{V}) = P(\mathbf{C}\colon \mathbf{D}) \prod_{D\in\mathbf{D}} P(D \mid \pi(D))\,. \tag{2.5}$$

Using this distribution we can compute the expected utility of a strategy Δ as:

$$\mathrm{EU}(\Delta) = \sum_{\mathbf{v}} P_\Delta(\mathbf{v})U(\mathbf{v})\,. \tag{2.6}$$

The aim of any rational decision maker is then to maximize the expected utility by finding an optimal strategy:

$$\Delta^* \equiv \arg\max_{\Delta} \mathrm{EU}(\Delta)\,. \tag{2.7}$$

Influence diagrams are not the only way to represent decision problems (notable alternatives are *decision trees* (Quinlan, 1986, 1992), *valuation networks* (Shenoy, 1996), and *sequential decision diagrams* (Covaliu and Oliver, 1995)) but the compactness and intuitiveness with which (symmetric) decision problems are specified, are desirable properties of the influence diagram formalism (Bielza and Shenoy, 1999).

2.6 Solving an influence diagram

There are different ways to solve an influence diagram (i.e., finding the optimal strategy). The original solution method transforms an influence diagram into a corresponding decision tree and then solves the corresponding decision tree (Howard and Matheson, 1984a). This solution method does not necessarily require a total ordering of the decision nodes, although this results in enormous space requirements (Pearl, 1988). A popular algorithm for solving influence diagrams was presented in (Olmsted, 1983) and is based on the following four graph transformations:

- *Barren node removal*:
  Chance or decision nodes that do not have children may be removed from the graph.

- *Arc reversal*:
  The orientation of an arc between two chance nodes C and C' may be reversed if there is no other directed path between C and C', by letting the parents of C be inherited by C' and vice versa, and by recomputing the conditional probabilities for C and C' using Bayes' rule.

- *Conditional expectation*:
  A chance node C that directly precedes U may be removed by adding the parents of C to the parents of U and eliminating C by taking the conditional expectation.

- *Maximization*:
  A decision node D that directly precedes U may be removed by maximizing the expected utility, provided that barren nodes have been removed and predecessors of U are also predecessors of D.

The algorithm is guaranteed to find the optimal action for the first decision node after a finite number of transformations. A third solution method is based on the transformation of an influence diagram to a Bayesian network and to use probabilistic inference methods for evaluation (Cooper, 1988; Shachter and Peot, 1992). Due to this technique, we can represent decision-theoretic notions such as decisions and utilities in a Bayesian network, even though this is not explicitly provided by the semantics of Bayesian networks. We will make use of this observation when we deal with dynamic decision problems; i.e., when decision making extends over longer periods of time. Dynamic decision making is discussed in-depth in Chapter 5.

Chapter 3

Clinical Decision Support with Bayesian Networks

In the last decades, many techniques have been developed that can serve as the basis for automated clinical decision support. Some examples of these techniques are frame-based systems (Miller and Pople, 1982; Aikins, 1983), rule-based systems (Buchanan and Shortliffe, 1984), and probabilistic methods (Cooper, 1984; Spiegelhalter and Knill-Jones, 1984). In the latter category, Bayesian networks (also called belief networks) (Pearl, 1988) have become a popular tool for automated clinical decision support since they allow for the explicit representation of domain knowledge and sound probabilistic inference. Developing a Bayesian network as part of a system that supports clinical tasks such as diagnosis or treatment selection, implies bridging the gap between an informal description of the clinical task and its actual implementation in terms of a Bayesian network. It has been recognized before in the knowledge acquisition and modeling research community that it is often only feasible to bridge this gap in small steps, for example by using intermediate, semi-formal representations that somehow capture the essence of the task to be modeled. This is what is being offered by the idea of representing the clinical decision making process in terms of problem solving methods. However, to date, not much is known about how problem solving methods that capture clinical tasks relate to concrete implementations in terms of Bayesian networks.

In this chapter, we address the problem of how to get from a particular informally described clinical decision making task to the constructed Bayesian network that supports that task. We commence by providing abstract descriptions of some important tasks in clinical decision support in Section 3.1, and show how logical, Bayesian, and decision theoretic formulations of clinical decision support relate to these abstract descriptions. Subsequently, we show in Section 3.2 how the descriptions translate into concrete Bayesian network designs, where the implications of some common design assumptions will be discussed. In Section 3.3, we discuss concrete aspects of Bayesian network development that can be of practical use to the knowledge engineer.

3.1 Clinical problem solving

Insight into the nature of clinical decision making is, and should be, the starting point for the construction of models aimed at supporting the tasks involved in it. We start by adopting the view that clinical decision making can be described as a type of problem solving in a way related to previous work done in the knowledge modeling community (e.g. (Schreiber et al., 2000)).

3.1.1 Problem solving methods

All activities in clinical decision making can be described in terms of problem solving, where solving a problem is described in terms of domains, models, knowledge sources, and relationships between models and knowledge sources. This yields an abstract view of clinical decision making, which then can be elaborated on at a more detailed level, e.g., in terms of an underlying language L such as predicate logic, probability theory, or decision theory. We define a domain of discourse as follows.

Definition 3.1. *Let* $\Phi = (\mathcal{U}, \mathcal{A}, \mathcal{O})$ *be a* domain *of discourse, with the set* $\mathcal{U}$ *containing* unobservable *elements, the set* $\mathcal{A}$ *containing* action *elements, and the set* $\mathcal{O}$ *containing* observable *elements, where the sets are pairwise disjoint.*

The set $\mathcal{U}$ contains the domain concepts that cannot be observed by an external observer. E.g., the tumor size in a cancer patient is an unobservable element. The set $\mathcal{A}$ contains the actions that can be performed by a decision maker, such as chemotherapy or surgery for a patient. The set $\mathcal{O}$ contains the domain concepts that can be observed by an external observer, such as patient gender. It is assumed that the problem to be solved is given by a *problem description*, which is defined as follows.

Definition 3.2. *Let* $\mathcal{M}$ *define a* model *over a domain* Φ *and a set of* problem solutions Σ*, both in some language* L*. A* problem description *is defined as a tuple*

$$\mathcal{D} = (\Phi, \Sigma, \mathcal{M}, \mathbf{a}, \mathbf{o})$$

with $\mathbf{a} \subseteq \mathcal{A}$ *the set of actions and* $\mathbf{o} \subseteq \mathcal{O}$ *the set of observations.*

The set Σ represents the set of problem solutions, which may be elements in Φ, or more abstract elements defined for the problem at hand. For example, in a diagnostic problem, the set of problem solutions may be the set of unobservable disorders, such that $\Sigma = \mathcal{U}$, or a set of disorder classes, such as $\Sigma = \{\mathsf{benign\text{-}disease}, \mathsf{malignant\text{-}disease}\}$. The set $\mathbf{a}$ represents the set of actions that are selected by an external decision maker. This comprises both actions that have been performed in the past and actions that still need to be performed in the future. The set $\mathbf{o}$ represents the set of observations that are actually observed for a particular problem instance.

Problem solving in a particular domain may, or may not, involve explicit reasoning about time. This is the reason why in the following a distinction is made between *non-temporal* problem solving, where the temporal nature of the clinical task at hand is not explicitly taken into account, and *temporal* problem solving, where we do explicitly take into account the notion of time. In case of a non-temporal problem description, we make use of a non-temporal model, whereas in case of a temporal problem description, we make use of a temporal model where domain elements in Φ are assumed to be indexed by time. Non-temporal problem solving is defined as follows.

Definition 3.3. *A* (non-temporal) problem solution *of a non-temporal problem description* $\mathcal{D} = (\Phi, \Sigma, \mathcal{M}, \mathbf{a}, \mathbf{o})$ *is defined as a set* $\mathbf{s} \subseteq \Sigma$*, such that*

$$\mathcal{M} \cup \mathbf{a} \cup \mathbf{s} \models_N \mathbf{o}$$

for a (non-temporal) problem solving relation $\models_N$.

This means that the observations $\mathbf{o}$ can be explained in terms of a model $\mathcal{M}$ together with selected actions $\mathbf{a}$ and the solution $\mathbf{s}$. What this particular type of problem solving does, is determined by the content of the non-temporal problem solving relation $\models_N$. Temporal problem solving is defined similarly, as follows.

Definition 3.4. *A* (temporal) problem solution *of a temporal problem description* $\mathcal{D} = (\Phi, \Sigma, \mathcal{M}, \mathbf{a}, \mathbf{o})$ *is defined as a set* $\mathbf{s} \subseteq \Sigma$*, such that*

$$\mathcal{M} \cup \mathbf{a} \cup \mathbf{s} \models_T \mathbf{o}$$

for a (temporal) problem solving relation $\models_T$.

In the following, we use $\models$ to denote either a non-temporal or temporal problem solving relation. In general, it should hold that $\mathcal{M} \cup \mathbf{a} \cup \mathbf{s} \cup \mathbf{o} \not\models \bot$, meaning that the model is consistent with actions, solutions, and observations.

For clinical problem solving, we consider the relation between patient and physician. In this case, $\mathcal{M}$ can be distinguished into a *patient model* $\mathcal{M}^\pi$, which describes how the patient responds to decisions made by the physician, and a *physician model* $\mathcal{M}^\phi$, which describes how decision making by the physician is influenced by observations about the patient. The problem solution is then considered to be the physician's response for a particular problem description, which includes the model $\mathcal{M} = \mathcal{M}^\pi \cup \mathcal{M}^\phi$, observations $\mathbf{o}$ that represents observations about the patient's state, and interventions $\mathbf{a}$ that are explicitly chosen by the physician. The distinction between patient and physician is refined by organizing clinical concepts in the domain of discourse Φ, that are used in the definition of models $\mathcal{M}$, into the following categories (Weiss et al., 1978a):

Patient
- Risk factors (etiology, epidemiology, patient characteristics)
- Pathogenesis (concepts linking risk factors and disorders)
- Disorders (cause, localization, complications)
- Pathophysiology (concepts linking disorders and findings)
- Findings (clinical signs, test results, subjective symptoms)

Physician
- Tests (physical examination, laboratory tests)
- Treatments (drug therapy, surgery, radiotherapy, psychotherapy).

In the context of clinical problem solving, pathogenetic and pathophysiological concepts are assumed to be unobservable by the physician. Disorders are assumed to be particular pathophysiological concepts and form a special case, since prior to diagnosis, disorders are unknown to the physician, whereas after diagnosis it may be the case that disorders become observed. The actions that can be performed by the physician are given by tests, which may be used to gather information, and treatments, which may be used to influence the pathophysiological process. Risk factors and findings are assumed to be observable by the physician and the information they provide can be used as the basis for clinical problem solving. In case of temporal clinical problem solving, clinical concepts are indexed by times in $T \subseteq \mathbb{R}$. A problem solution $\mathbf{s} \in \Sigma$ may be a physician's conclusion about the patient's state, but often the solution involves decision making since most physicians agree that the majority of clinical questions for which support is needed deal with what the physician should *do* instead of what the physician should *know* (Shortliffe et al., 2001). Under the latter interpretation, a clinical problem description can be viewed as a *control problem*, requiring optimal manipulation of a (stochastic) process by an external decision maker. Figure 3.1 depicts this representation of clinical problem solving.

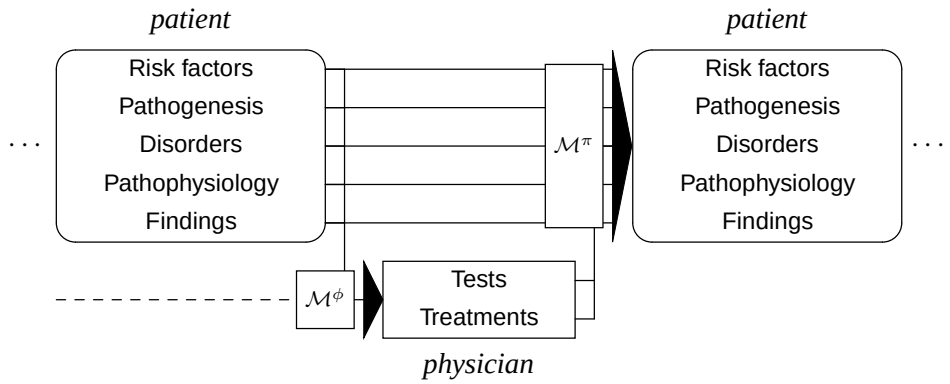


Figure 3.1: Clinical problem solving as a control problem, where the physician's decisions are based on the current patient state and the past patient state (as indicated by the dashed line). The decisions in turn influence the patient's future state.

In this chapter, we focus on problem solving for the primary tasks in clini-

cal patient management, which are taken to be *diagnosis*, *test selection*, *prognosis*, *treatment selection*, and *monitoring*. For these clinical tasks, we formulate non-temporal and temporal problem solving variants where the described categories form the domain of discourse Φ. In case of temporal problem solving, we use $t_c \in T$ to represent the *present*, $H = \{t\colon t < t_c, t \in T\}$ to represent the past, and $G = \{t\colon t > t_c, t \in T\}$ to represent the *future*. H^+ and G^+ are used to represent the past and future *including* the present time t_c. Table 3.1 provides for an overview of the choices of $\mathcal{U}$, $\mathcal{A}$, and $\mathcal{O}$ that are typically made for the various clinical tasks in case of a temporal problem description. For a non-temporal problem description, we use the same structure save the fact that time is omitted from the description.

Table 3.1: Choices of $\mathcal{U}$, $\mathcal{A}$, and $\mathcal{O}$ for a temporal problem description of a clinical task.

Task	Design choices
Diagnosis	$\mathcal{U} \subseteq \big(\mathsf{Pathogenesis} \cup \mathsf{Pathophysiology} \cup \mathsf{Disorders}\big) \times H^+$ $\mathcal{A} \subseteq \big(\mathsf{Tests} \cup \mathsf{Treatments}\big) \times H$ $\mathcal{O} \subseteq \big(\mathsf{Risk\ factors} \cup \mathsf{Findings}\big) \times H^+$
Test selection	$\mathcal{U} \subseteq \big(\mathsf{Pathogenesis} \cup \mathsf{Pathophysiology} \cup \mathsf{Disorders}\big) \times H^+$ $\mathcal{A} \subseteq \big(\mathsf{Tests} \cup \mathsf{Treatments}\big) \times H$ $\mathcal{O} \subseteq \big(\mathsf{Risk\ factors} \cup \mathsf{Findings} \cup \mathsf{Disorders}\big) \times H^+$
Prognosis	$\mathcal{U} \subseteq \mathsf{Pathophysiology} \times H^+$ $\mathcal{A} \subseteq \big(\mathsf{Tests} \cup \mathsf{Treatments}\big) \times T$ $\mathcal{O} \subseteq \big(\mathsf{Risk\ factors} \cup \mathsf{Findings} \cup \mathsf{Disorders}\big) \times H^+$
Treatment selection Monitoring	$\mathcal{U} \subseteq \mathsf{Pathophysiology} \times H^+$ $\mathcal{A} \subseteq \big(\mathsf{Tests} \cup \mathsf{Treatments}\big) \times H$ $\mathcal{O} \subseteq \big(\mathsf{Risk\ factors} \cup \mathsf{Findings} \cup \mathsf{Disorders}\big) \times H^+$

We now turn our attention to a description of the various clinical tasks.

Diagnosis

Diagnosis refers to the explanation of observations in terms of unobservable disorders. Since we do not need to model decision making explicitly for pure diagnosis, we may restrict ourselves to a patient model $\mathcal{M}^\pi$, which models the relation between disorders and observations, possibly influenced by selected actions. The problem solving relation $\models$ uses observations $\mathbf{o} \subseteq \mathcal{O}$ (and possibly actions $\mathbf{a} \subseteq \mathcal{A}$ whenever they induce changes in how disorders relate to observations) in order to predict disorders $\mathbf{s} \subseteq \Sigma$ from which the patient is suffering. In case of *non-temporal* diagnosis, it is assumed that the set of solutions is defined as $\Sigma \subseteq \mathsf{Disorders}$. In case of *temporal* diagnosis, the set of solutions is defined as $\Sigma \subseteq \mathsf{Disorders} \times H^+$. Note that actions are confined to the strict past, since current actions do not have an immediate effect on the diagnosis. The definition of Σ allows the expression of *when* disorders have developed, but often our interest is in *current* disorders only, such that $\Sigma \subseteq \mathsf{Disorders} \times \{t_c\}$.

Test selection

Test selection stands for the selection of tests by the physician for the purpose of information gathering. The model $\mathcal{M}$ consists of a physician model $\mathcal{M}^{\phi}$ that dictates which tests to choose in a given situation, and possibly a patient model $\mathcal{M}^{\pi}$ which allows for the representation of how unobserved quantities affect decision making. The problem solving relation $\models$ uses observations $\mathbf{o} \subseteq \mathcal{O}$ (and possibly actions $\mathbf{a} \subseteq \mathcal{A}$) in order to select tests in Σ that maximize the information gained and minimize patient risk. For *non-temporal* test selection, the set of solutions is given by $\Sigma \subseteq \mathsf{Tests}$. Note that disorders can be part of either the unobservable or the observable variables, depending on whether testing is performed for the purpose of diagnosis (in which case the disorder is unknown) or for the purpose of treatment (in which case the disorder is typically known). For *temporal* test selection, the set of solutions is given by $\Sigma \subseteq \mathsf{Tests} \times G^{+}$. We remark that for a diagnostic *process*, diagnosis and test selection is interleaved, since diagnosis depends on the information that is unveiled by selected tests. The same holds for the treatment process, where treatment and testing depend on one another.

Prognosis

Prognosis stands for the prediction of a prognostic outcome for a patient given observations, performed actions, and projected actions. The model $\mathcal{M}$ should contain the patient model $\mathcal{M}^{\pi}$, which describes projected patient response, and may additionally contain the physician model $\mathcal{M}^{\phi}$, which describes projected interventions by the physician. The problem solving relation $\models$ uses observations $\mathbf{o} \subseteq \mathcal{O}$ together with performed actions and projected actions $\mathbf{a} \subseteq \mathcal{A}$ in order to assign the patient to a prognostic solution in Σ. Prognosis is typically performed in the situation where the disorder from which a patient is suffering is known, and the set of solutions Σ may either be defined in terms of abstract concepts such as quality-adjusted life expectancy or concrete concepts such as health status, tumor size, etc. For *temporal* prognosis, the set of solutions Σ may again be defined in terms of abstract or concrete concepts that are indexed by time, such as patient survival in the coming five years, or tumor remission in the next year. Note that the action set $\mathcal{A}$ also contains future actions since this allows the physician to insist on a projected treatment. This is to be contrasted with the physician model $\mathcal{M}^{\phi}$, which represents future decision making as a whole and may depend on future (yet to be made) observations.

Treatment selection

Treatment selection stands for the selection of actions by the physician for the purpose of influencing the pathophysiological process. It is not much different from test selection since the only change is the purpose of the task, namely control instead of information gathering. The model $\mathcal{M}$ therefore consists of a physician model $\mathcal{M}^{\phi}$

that dictates which treatments to choose in a given situation, and possibly a patient model $\mathcal{M}^\pi$ which allows for the representation of how unobserved quantities affect decision making. The problem solving relation $\models$ uses observations $\mathbf{o} \subseteq \mathcal{O}$ (and possibly actions $\mathbf{a} \subseteq \mathcal{A}$) in order to select treatments from the set of possible treatments Σ, where the selected treatments should maximize patient benefit and minimize patient risk. It is assumed that during treatment, disorders are known, as is shown in Table 3.1. For *non-temporal* treatment selection we use solutions $\Sigma \subseteq$ Treatments and for *temporal* treatment selection we use solutions $\Sigma \subseteq$ Treatments $\times\, G^+$.

Monitoring

Monitoring stands for the prediction of the current pathophysiological state based on observations and actions. The prediction requires a patient model $\mathcal{M}^\pi$ and a problem solving relation $\models$ which uses observations $\mathbf{o} \subseteq \mathcal{O}$ (and possibly actions $\mathbf{a} \subseteq \mathcal{A}$) in order to predict the current (unobservable) pathophysiological state of the patient. In case of *non-temporal* monitoring, we have solutions $\Sigma \subseteq$ Pathophysiology, and in case of *temporal* monitoring, we have solutions $\Sigma \subseteq$ Pathophysiology $\times\, \{t_c\}$.

We have described the various clinical tasks in terms of abstract non-temporal and temporal problem solving, independent of the language L at hand. In the following, we describe logical, probabilistic, and decision-theoretic problem solving respectively, and also discuss what, according to these interpretations, constitutes a good problem solution.

3.1.2 Logical problem solving

In logical problem solving, we use standard first-order predicate logic as our language L. In order to make the notion of logical problem solving more concrete, we focus on a logical formulation of non-temporal diagnosis, called *abductive diagnosis*.

In abductive diagnosis (Console et al., 1989, 1991), we start with a domain $\Phi = (\mathcal{U}, \mathcal{A}, \mathcal{O})$ with $\mathcal{U} =$ Disorders, $\mathcal{A} = \varnothing$, and $\mathcal{O} \subseteq$ Findings; sets are interpreted logically as conjunctions of their elements. Here, disorders are given by so-called *defect literals* d and findings are given by so-called *finding literals* f. The non-temporal problem description is given by $\mathcal{D} = (\Phi, \Sigma, \mathcal{M}^\pi, \varnothing, \mathbf{o})$, where the set of solutions is given by $\Sigma = \mathcal{U}$ and the patient model $\mathcal{M}^\pi$ is a set of logical implications of the form:

$$d_1 \wedge \cdots \wedge d_n \rightarrow d, \quad d_1 \wedge \cdots \wedge d_m \rightarrow f$$

linking defects with defects and defects with findings respectively. Selected actions are not taken into account, and observations $\mathbf{o}$ are assumed to be given by the union of present and absent findings:

$$\begin{aligned} \mathbf{o}^+ &\subseteq \{f \in \text{Findings}\colon f \text{ is a positive literal}\} \\ \mathbf{o}^- &\subseteq \{\neg f \in \text{Findings}\colon f \notin \mathbf{o}^+, f \text{ is a positive literal}\} \end{aligned}$$

An *abductive diagnosis* of $\mathcal{D}$ is defined as a set of defects $\mathbf{s} \subseteq \Sigma$, such that:

1. $\mathcal{M}_N^\pi \cup \mathbf{s} \models \mathbf{o}^+$ (covering condition)
2. $\mathcal{M}_N^\pi \cup \mathbf{s} \cup \mathbf{o}^- \not\models \bot$ (consistency condition)

Hence, in abductive diagnosis, the hypothesis $\mathbf{s}$ must predict all present findings and should not predict any absent finding. Therefore, we have a non-temporal problem solving relation of the following form:

$$\mathcal{M}_N^\pi \cup \mathbf{a} \cup \mathbf{s} \models_N \mathbf{o} \Leftrightarrow \mathcal{M}_N^\pi \cup \mathbf{s} \models \mathbf{o}^+ \wedge \mathcal{M}_N^\pi \cup \mathbf{s} \cup \mathbf{o}^- \not\models \bot$$

where $\mathbf{s}$ is a possible problem solution.

From a logical point of view, any problem solution that is derived using logical deduction is *optimal* in the sense that solutions are indistinguishable. Often, however, in a particular logical framework, extra optimality criteria are added that allow the selection of an optimal solution from a set of possible solutions (e.g. (Lucas, 1998)). In abductive diagnosis, we often require that the solution is minimal with respect to set inclusion. For example, suppose we have a model

$$d_1 \rightarrow f, \quad d_2 \rightarrow f, \quad d_1 \wedge d_2 \rightarrow f$$

Then, upon observing f, we deduce that d_1 and d_2 are optimal problem solutions, whereas $d_1 \wedge d_2$ is not.

Although logical approaches to clinical problem solving such as abductive diagnosis have proven successful, there are also problems which logical problem solving cannot handle. In particular, the resulting optimal problem solutions may lead to non-optimal behavior. For example, it may well be the case that an observed finding f provides much stronger evidence for d_2 than for d_1, which may lead us to favor d_2 over d_1, which is not easily expressed in a logical framework. Bayesian problem solving, as discussed in the following section, solves this problem by expressing preferences for optimal solution, in terms of a measure of belief.

3.1.3 Bayesian problem solving

In Bayesian problem solving, the language L is chosen to be probability theory. We start with a problem description $\mathcal{D} = (\Phi, \Sigma, \mathcal{M}, \mathbf{a}, \mathbf{o})$ where Φ contains assignments to random variables of the form $X = x$ for non-temporal problems, and assignments to random processes of the form $X(t) = x$ for temporal problems. The model $\mathcal{M}$ represents a probability model that allows the association of a posterior probability $P(\mathbf{s} \mid \mathbf{a}, \mathbf{o}) \in [0, 1]$ with any problem solution $\mathbf{s} \subseteq \Sigma$, expressing our degree of belief in $\mathbf{s}$ *given* actions $\mathbf{a}$ and observations $\mathbf{o}$. Then, to say that a problem solution is possible, is equivalent to stating that $P(\mathbf{s} \mid \mathbf{a}, \mathbf{o}) > 0$.

Bayesian problem solving gives us a stronger optimality criterion than logical problem solving. In a truly Bayesian setting, we would express a problem solution

as a posterior distribution over $\mathbf{s} \subseteq \Sigma$, but if we are forced to choose a particular solution $\mathbf{s} \subseteq \Sigma$ then Bayesian problem solving dictates that we should use the most probable explanation (MPE) (or maximum a posteriori (MAP) hypothesis in case of incomplete evidence) as our optimal problem solution. The MPE criterion states that out of all the world models consistent with the evidence, we should choose the one with highest overall probability (Pearl, 1988):

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} P(\mathbf{s} \mid \mathbf{a}, \mathbf{o}) .$$

Using the MPE criterion, we define an optimal problem solution as follows:

$$\mathcal{M} \cup \mathbf{a} \cup \mathbf{s} \models \mathbf{o} \Leftrightarrow \mathbf{s} = \mathbf{s}^*$$

where $\mathbf{s}^*$ is an MPE problem solution. For the example of Section 3.1.2, we would choose the configuration of $\mathbf{s} = \{D_1 = d_1, D_2 = d_2\}$ for which $P(\mathbf{s} \mid f)$ is maximal. Note that the minimality criterion, which was used as an additional constraint in Section 3.1.2, is implied by Bayesian problem solving since it follows from the rules of probability theory that $P(\mathbf{x} \mid \mathbf{y}) \geq P(\mathbf{x}' \mid \mathbf{y})$ if $\mathbf{x} \subseteq \mathbf{x}'$. Still, Bayesian problem solving may lead to non-optimal behavior in case one possesses payoff information for the different solutions. For example, if misdiagnosing D_2 leads to more negative consequences than misdiagnosing D_1 (such as higher death risk), we may still be inclined to diagnose D_2, even if it holds that $P(D_1 = \textit{yes}, D_2 = \textit{no} \mid f) \gg P(D_1 = \textit{no}, D_2 = \textit{yes} \mid f)$. We handle this with decision-theoretic problem solving, as discussed in the next section.

3.1.4 Decision-theoretic problem solving

Decision-theoretic problem solving, where the language L is decision-theory, subsumes Bayesian problem solving and dictates that, in the presence of payoff information, an optimal problem solution is given by the maximum expected utility (MEU) criterion (Von Neumann and Morgenstern, 1947). The MEU criterion represents payoff information in the form of utilities $U(\mathbf{x})$ that express the reward gained (or cost experienced) for different solutions, and states that the best solution is the one which maximizes reward (or minimizes cost):

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} \sum_{\mathbf{x}} U(\mathbf{x}) P(\mathbf{x} \mid \mathbf{a}, \mathbf{o}) .$$

for $\mathbf{x}$ compatible with $\mathbf{a}, \mathbf{o}$, and $\mathbf{s}$. Using the MEU criterion, we define an optimal problem solution as follows:

$$\mathcal{M} \cup \mathbf{a} \cup \mathbf{s} \models \mathbf{o} \Leftrightarrow \mathbf{s} = \mathbf{s}^*$$

where $\mathbf{s}^*$ is a MEU problem solution. Returning to the diseases and findings of Section 3.1.2, suppose we have

$$\begin{array}{lcll} P(D_1 = \textit{yes}, D_2 = \textit{no} \mid f) & = & 0.99 & U(D_1 = \textit{yes}, D_2 = \textit{no}) = 1 \\ P(D_1 = \textit{no}, D_2 = \textit{yes} \mid f) & = & 0.01 & U(D_1 = \textit{no}, D_2 = \textit{yes}) = 100 \end{array}$$

Then, we would choose d_2 as our diagnosis, even though it is not the MPE problem solution. Note that in case we have no payoff information or if utilities are equal for all solutions then decision-theoretic problem solving reduces to Bayesian problem solving. If, additionally, uncertainty does not play a role then decision-theoretic problem solving reduces to logical problem solving.

3.2 Bayesian network designs for clinical tasks

In Section 3.1, we have shown how clinical tasks can be solved using different problem solving strategies, but we have not yet addressed the properties of the model $\mathcal{M}$ that is used for problem solving. In this section, we will focus on decision-theoretic problem solving (with Bayesian and logical problem solving as special cases), and show how $\mathcal{M}$ can be described in terms of particular Bayesian network designs. As before, we distinguish non-temporal and temporal problem solving.

3.2.1 Non-temporal problem solving with Bayesian networks

Let $\mathbf{X}$ be a set of random variables, representing relevant domain variables.[1] A Bayesian network (G, P) consists of an acyclic directed graph G that represents the independence structure between domain variables and a joint probability distribution (JPD) P for random variables in $\mathbf{X}$. A Bayesian network can often represent the JPD compactly, since G factorizes the JPD according to:

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X \mid \pi(X)) \tag{3.1}$$

where $\pi(X)$ denotes the parents of X in G. This factorization generally reduces the number of parameters that need to be estimated and allows for more efficient probabilistic inference.

If a Bayesian network is used for the purpose of non-temporal problem solving, then the aim is to define a JPD for variables in $\mathbf{X} \subseteq \mathcal{U} \cup \mathcal{A} \cup \mathcal{O} \cup \Sigma$, where probabilistic independence between domain variables is modeled by the absence of arcs in G. The design of a Bayesian network is then determined in part by (1) the nature of the clinical task, (2) the selected clinical categories, and (3) independence relations between clinical categories that are assumed to hold. For example, if the

[1] Decisions and utility functions can be transformed into random variables when required (Cooper, 1988; Shachter and Peot, 1992).

task is (non-temporal) test selection, then we need to represent at least the solutions $\Sigma \subseteq$ Tests and a physician model $\mathcal{M}^{\phi}$ which specifies how tests are selected. More complex designs may distinguish more clinical categories, incorporate more domain variables, and/or use less restrictive independence assumptions at the level of clinical categories and/or domain variables. For a real-world clinical problem, choosing the right design requires finding a balance between many constraints, which allows for the easy specification of few parameters at the expense of model expressiveness, and few constraints, which allows for an expressive model at the expense of the more difficult specification of many parameters.

A rigorous restriction is to disregard hidden variables and actions and to focus solely on how a solution S is influenced by a set of observations $\mathbf{O}$. There are two common approaches to the implementation of this restriction, as shown in Figs. 3.2 and 3.3.

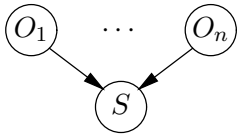


Figure 3.2: A discriminative model.

The discriminative model in Fig. 3.2 predicts the state of S directly from the states of $\mathbf{O} = \{O_1, \ldots, O_n\}$ through the associated conditional probability distribution $P(S \mid \mathbf{O})$. For discrete random variables, the number of parameters that need to be estimated for this model, equals $(|\Omega_S| - 1) \cdot \prod_{i=1}^{n} |\Omega_{O_i}|$. This normally remains prohibitive in practice, since the number of observations n, and/or the state-spaces Ω_{O_i} and Ω_S can be large. One way to solve this problem, is to constrain the form of $P(S \mid \mathbf{O})$. If, for instance, it is assumed that the influences of observations O_i on the outcome S combine linearly, then we can use a special form such as the softmax regression model:

$$P(S = s_j \mid \mathbf{O} = \mathbf{o}) = \frac{e^{a_{j1}(o_1)+\cdots+a_{jn}(o_n)}}{\sum_{k=1}^{m} e^{a_{k1}(o_1)+\cdots+a_{kn}(o_n)}} .$$

For continuous observations with $a_{ij}(o_j) = \alpha_{ij} \cdot o_j$ and a binary valued outcome variable S, this model reduces to the well-known logistic regression model, which is used extensively in medicine. Another approach that constrains the form of $P(S \mid \mathbf{O})$ would be to assume that observations act independently and combine deterministically, as is the topic of Chapter 4.

The generative model of Fig. 3.3 takes a different approach. Instead of constraining the form of $P(S \mid \mathbf{O})$, it uses Bayes' theorem together with additional independence assumptions, in order to make the computation of $P(S \mid \mathbf{O})$ feasible. Recall that according to Bayes' theorem, it holds that

$$P(S \mid \mathbf{O}) \propto P(S)P(\mathbf{O} \mid S).$$

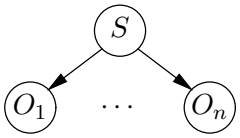


Figure 3.3: A generative model.

By introducing the assumption that observations $O, O' \in \mathbf{O}, O \neq O'$ are conditionally independent given the outcome, we arrive at the generative model of Fig. 3.3, with associated conditional probability distribution

$$P(S \mid \mathbf{O}) \propto P(S) \prod_{O \in \mathbf{O}} P(O \mid S)\,. \tag{3.2}$$

Since the generative model assumes independence of observations given the outcome, it is known as the *naive* Bayes model. It has the advantage that it only requires the estimation of $|\Omega_S - 1| + \sum_{O \in \mathbf{O}} (|\Omega_O| - 1) \cdot |\Omega_S|$ parameters. De Dombal's system for the diagnosis of acute abdominal pain employed the naive Bayes model and was one of the first successful implementations of Bayesian probability theory in medicine (de Dombal et al., 1972).

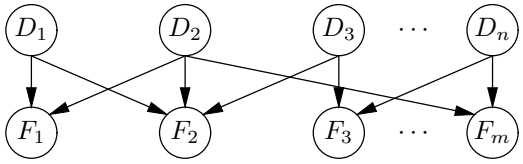


Figure 3.4: The QMR-DT system is a bipartite graph, modeling the association between disorders D_i and findings F_j.

A generalization of the naive Bayes model to multiple class variables has been used in the definition of QMR-DT (Shwe et al., 1991). It is a Bayesian reformulation of the Internist-1/QMR expert-system for differential diagnosis in internal medicine (Miller and Pople, 1982; Miller et al., 1986) and is shown in Fig 3.4. The graph is constrained to be a bipartite graph relating disorders D_i (such that $\Sigma = \mathsf{Disorders}$) and findings F_j (such that $\mathcal{O} = \mathsf{Findings}$). Note that additional independence assumptions are defined at the level of clinical categories, since QMR-DT assumes that findings are conditionally independent given the disorders, as in the naive Bayes model.

Promedas (Kappen and Neijt, 2002) covers a large diagnostic repertoire of internal medicine and extends the QMR-DT architecture by defining risk factors R_k that condition the occurrence of disorders (such that $\mathcal{O} = \mathsf{Risk\ factors} \cup \mathsf{Findings}$). Additionally, disorders may condition other disorders since the occurrence of one disorder may influence the occurrence of another disorder. Note that additional independence assumptions are again defined at the level of clinical categories, since risk factors are

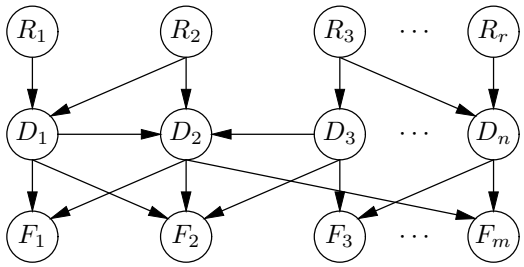


Figure 3.5: Promedas models associations between risks R_k, disorders D_i, and findings F_j.

marginally independent and findings are conditionally independent given the disorders (Fig. 3.5).

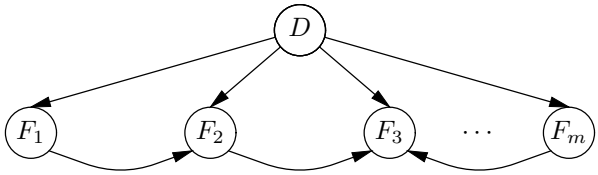


Figure 3.6: Pathfinder consists of a disorder node D containing over 60 mutually-exclusive disorders that condition findings $F_1, \ldots, F_m$ with $m > 130$.

The assumption of independence between observations given the outcome that is made by the above structures is often unrealistic and many alternative structures therefore focus on lifting the independence assumptions of the naive Bayes model (Spiegelhalter and Knill-Jones, 1984). Pathfinder was one of the first large graphical models for medical decision support (Heckerman and Nathwani, 1992a,b) and does not consider findings to be conditionally independent given the disorder, although it does assume that disorders are mutually exclusive (Fig. 3.6). Pathfinder is used for the diagnosis of more than 60 lymph node disorders, using more than 130 microscopic, clinical, laboratory, immunological, and molecular-biologic features. Its commercialization, known as IntelliPath, has been used by physicians, both in practice and in education (Heckerman, 1990).

Even though Bayesian networks that are based on restricted structures may perform well in clinical tasks such as differential diagnosis, they often make unrealistic assumptions which affects both the accuracy of computed posterior probability distributions and the ability to understand how random variables interact in the domain. The developers of QMR-DT remark, for instance, that performance suffered from the lack of anatomical knowledge, the absence of the representation of intermediate pathophysiological states, and the lack of dependencies between diseases (Shwe et al., 1991). In practice, one often needs detailed information about the causal mechanisms that are responsible for observed findings. The use of causality as a guiding principle when building a Bayesian network for clinical decision support is advan-

tageous, since knowledge concerning pathophysiology and the effect of treatment is normally described in the medical literature in terms of causes and effects (Lucas, 1995). This has been used as a modeling strategy in some of the early medical expert systems (Kulikowski and Weiss, 1982; Patil et al., 1982; Miller and Pople, 1982; Pople, 1982), and a number of Bayesian networks have recently been developed that capture the causal structure of restricted medical domains to various degrees of realism (e.g., (Andreassen et al., 1987; Díez et al., 1997; Kahn Jr et al., 1997; Wasyluk et al., 2001; Lacave and Díez, 2003; van der Gaag et al., 2001)). Causal models allow for an accurate representation of domain knowledge, and also facilitate the explanation of drawn conclusions, which may increase the acceptance of decision support in medicine, both by the physician and by the patient (Teach and Shortliffe, 1984; Suermondt and Cooper, 1993; Lacave and Díez, 2002). As our discussion about decision-theoretic problem solving suggests, even if systems such as QMR-DT would be capable of estimating the posterior probability of disease given findings with a reasonable accuracy, then, in general, this is insufficient for guiding treatment, since clinical decision support often requires the suggestion of appropriate action (Long, 1996). In other words, automatically obtaining a differential diagnosis is beneficial in the sense that the physician is less likely to misdiagnose, but does not always give insight into the optimal treatment given the differential diagnosis. Hence, it is often necessary to represent decision-making as well.

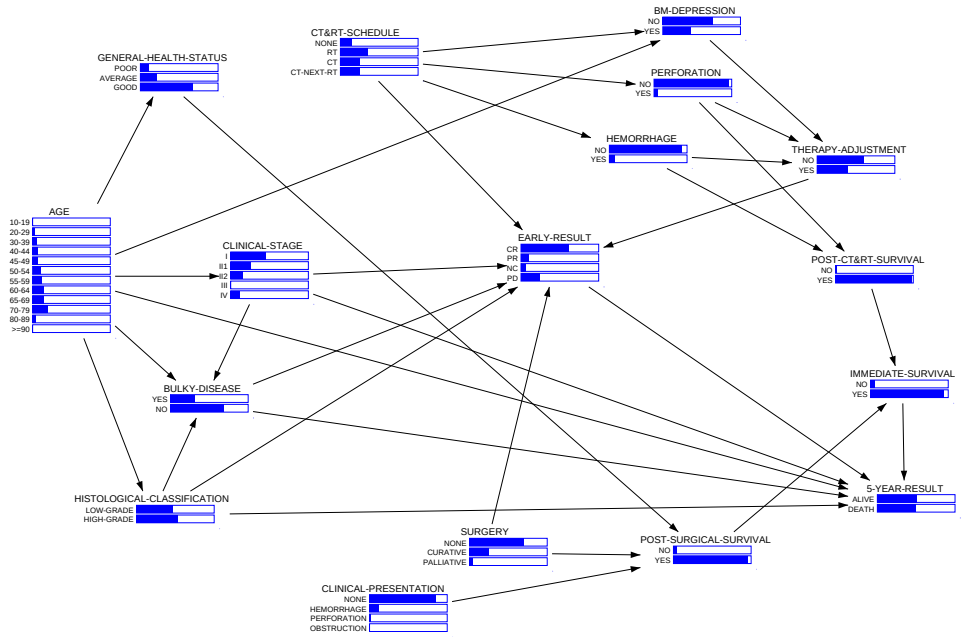


Figure 3.7: A causal model for prognosis of non-Hodgkin lymphoma.

The Bayesian network in Fig. 3.7 is an example of a non-temporal Bayesian network for prognosis of non-Hodgkin lymphoma that incorporates causal knowledge, with arc orientation denoting the flow from causes to effects (Lucas et al., 1998). It depicts for instance that general-health-status is influenced by age and shows that bulky disease is determined by age, the tumor's clinical-stage, and the tumor's histological-classification. It also incorporates decision making through the representation of the influence of treatment variables ct&rt-schedule (chemotherapy and radiotherapy schedule) and surgery on prognosis. Note that the Bayesian network does not represent the decision making strategy through a model $\mathcal{M}^{\phi}$, but rather requires the physician to impose a strategy through the selection of actions in $\mathcal{A}$. A prognosis is performed by selecting actions and observations, which gives a posterior distribution on the outcome 5-year result.

3.2.2 Temporal problem solving with Bayesian networks

If time is involved, domain variables are taken to be random processes, where $X(t)$ denotes a random process X at time $t \in T$. A Bayesian network (G, P) defined for a set $\mathbf{X}$ of random processes, is called a *dynamic Bayesian network (DBN)*, where G factorizes the JPD according to:

$$P(\mathbf{X}) = \prod_{X(t) \in \mathbf{X}} P(X(t) \mid \pi(X(t))) \,. \tag{3.3}$$

Since a DBN may be defined for a possibly infinite sequence of times $t \in T$, often, a number of standard assumptions are made. It is natural to assume that influences between random processes cannot be oriented against the arrow of time; i.e., $(X(t), Y(u)) \notin A(G)$ if $u < t$ with $t, u \in T$. We also find it useful to focus on discrete time $\{t_0, t_1, \ldots, t_h\} \subset \mathbb{R}_0^+$ with $t_{i+1} > t_i$ for $0 \leq i < h$, representing the decision moments for the clinical task. Here, t_0 denotes the *initial time*, such as for instance the time of birth or the time of admission to the hospital, and t_h denotes a (possibly countably infinite) *horizon*, which can be a fixed period (e.g., five years after admission) or an as yet undefined time of death (formally representing an infinite-horizon process). We also define a fixed *interval* $\delta_t = t_{i+1} - t_i$ for $t_i, t_{i+1} \in T$, which is chosen for the problem at hand. For example, in case of monitoring this period could be measured in seconds, hours, months, or even years. Given an infinite-horizon process, specification of a discrete-time DBN may still be prohibitive. In order to allow for a compact specification the following assumptions are commonly made:

- The DBN is *(first-order) Markovian*:

  $$\mathbf{X}(t+1) \perp\!\!\!\perp_P \mathbf{X}(t-1) \mid \mathbf{X}(t)$$

  such that the future is independent of the past given the present.

- The DBN is *time-invariant*:
  - The same independence relations hold at each point in time:

$$\mathbf{U}(t) \perp\!\!\!\perp_P \mathbf{V}(u) \mid \mathbf{W}(s) \Leftrightarrow \mathbf{U}(t+c) \perp\!\!\!\perp_P \mathbf{V}(u+c) \mid \mathbf{W}(s+c)$$

    for $\mathbf{U}, \mathbf{V}, \mathbf{W} \subseteq \mathbf{X}$ and $t, u, s, t+c, u+c, s+c \in T$. I.e., domain structure is fixed.
  - The model is *homogeneous*, such that

$$P(\mathbf{U}(t+c) \mid \mathbf{V}(t)) = P(\mathbf{U}(t'+c) \mid \mathbf{V}(t'))$$

    for $\mathbf{U}, \mathbf{V} \subseteq \mathbf{X}$ and $t, t', t+c, t'+c \in T$. I.e., transition probabilities are fixed.

Given these assumptions, the control structure of Fig. 3.1 can be completely specified by means of a *prior model* $\mathcal{B}_0$, representing the situation $P(\mathbf{X}(t_0))$, and a *transition model* $\mathcal{B}_t$, representing the change in state $P(\mathbf{X}(t) \mid \mathbf{X}(t\text{-}1))$ for $t > t_0, t \in T$, that takes place by moving forward in discrete time until the horizon t_h is reached. The pair $(\mathcal{B}_0, \mathcal{B}_t)$ is often used in the formulation of a dynamic Bayesian network (Dean and Kanazawa, 1989).

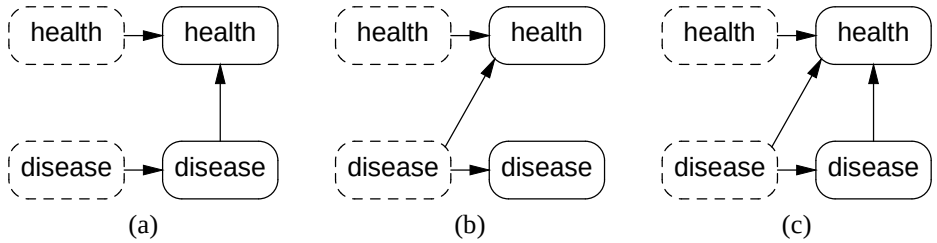


Figure 3.8: Three transition models for health and disease, where dashed objects represent the situation at time $t-1$, and solid objects represent the situation at time t. Solid arcs between objects denote possible dependence between the random variables that constitute the objects, and the absence of arcs denotes a statement of (conditional) independence.

There are multiple ways to indicate (in)dependence between random variables for a transition model. Consider for instance the way health is influenced by a disease in Fig. 3.8. Figure 3.8 (a) depicts an immediate influence of disease on health, which has the advantage that health can be predicted from the disease status, without taking into account temporal interactions. Figure 3.8 (b) depicts a lagged influence of disease on health. This has the advantage that future health is predicted from the current disease status, which can be more natural to the physician. Figure 3.8 (c) depicts the combined influence of past disease and present disease on health, and provides the most precise representation.

Example 3.1. Suppose that the disease is present at time t-1 and absent at time t. In case of Fig. 3.8 (a) we compute $P_a(\mathsf{health}(t) \mid \mathsf{disease}(t) = \textit{absent}, \mathsf{health}(t\text{-}1))$, whereas for Fig. 3.8 (b) we compute $P_b(\mathsf{health}(t) \mid \mathsf{disease}(t\text{-}1) = \textit{present}, \mathsf{health}(t\text{-}1))$. It is likely that the model of Fig. 3.8 (a) overestimates patient health since is does not take into account that the patient was still diseased at the previous point in time, whereas it is likely that the model of Fig. 3.8 (a) underestimates patient health since is does not take into account that the patient is cured at this point in time. In contrast, Fig. 3.8 (c) may take these effects into account, by representing $P_c(\mathsf{health}(t) \mid \mathsf{disease}(t), \mathsf{disease}(t\text{-}1), \mathsf{health}(t\text{-}1))$ as a weighted average of P_a and P_b. Note that, if disease progression is sufficiently slow, then the quality of the approximations would increase as δ_t decreases.

Once a transition model is completed, the prior model needs to be specified. In the prior model, we use variables $X(0)$ for all variables $X(t)$ in the transition model. Furthermore, since direct influences should hold at the initial time as well, arcs $(X(0), Y(0))$ are added to the prior model for each arc $(X(t), Y(t))$ in the transition model. Finally, for each variable Y that has an arc $(X(t\text{-}1), Y(t))$ in the transition model, we determine whether there are variables $Z_1(0), \ldots, Z_k(0)$ in the prior model that condition the distribution for $Y(0)$. The arcs $(Z_i(0), Y(0))$ do not necessarily reflect causation but rather associations between random variables that have arisen due to causal interactions in the past. Algorithm 3.1 summarizes the strategy for constructing a prior model.

Algorithm 3.1 Construction of a prior model.

1. Add a variable $X(0)$ to the prior model $\mathcal{B}_0$ for each variable $X(t)$ in the transition model $\mathcal{B}_t$.
2. Add an arc $(X(0), Y(0))$ to the prior model $\mathcal{B}_0$ for each arc $(X(t), Y(t))$ in the transition model $\mathcal{B}_t$.
3. For all variables Y with arcs $(X(t\text{-}1), Y(t))$ in the transition model $\mathcal{B}_t$ such that $X \neq Y$, determine if there are variables $Z_i(0)$ in the prior model $\mathcal{B}_0$ that condition the prior distribution of $Y(0)$ and add arcs $(Z_i(0), Y(0))$ to $\mathcal{B}_0$.

The temporal nature of a problem is often essential to clinical decision-making (Augusto, 2005). During diagnosis, to know the temporal order and duration of symptoms can influence the diagnostic conclusions, the selection of treatments or tests may depend on the time at which the selection is made, during prognosis, the disease dynamics is described as the unfolding of events over time, and during monitoring, we need to track the patient's pathophysiological status over time. The benefits of temporal modeling of clinical problems have become clear in practice, as illustrated by the work of Long (Long, 1996), who used a representation based on Bayesian networks and time intervals (Allen, 1984) for diagnosing heart disease, which eliminated errors that were made by a non-temporal model. Similarly, it was

found in (Charitos et al., 2005) that a redefinition of a static Bayesian network for the diagnosis of ventilator-associated pneumonia (VAP) in terms of a dynamic Bayesian network that allows for temporal reasoning, increased diagnostic performance. Some other examples of dynamic Bayesian networks in medicine are presented in Refs. (Dagum and Galper, 1993; Andreassen et al., 1994; Hernando et al., 1996).

Once the Bayesian network design has been chosen, we proceed with the actual construction of the Bayesian network. In the next section, we describe how Bayesian networks for clinical decision support are constructed in practice.

3.3 Bayesian network construction

Bayesian network construction may be distinguished into variable definition, structure specification, factor association, and parameter estimation. In this section, these basic steps will be discussed.

3.3.1 Variable definition

Variable definition refers to the identification of domain variables, and the determination of their *name*, *category*, *type*, and *states*. The name of a variable should be unambiguous, intuitive to the domain expert(s), and conforming to domain terminology. The category of the variable can be distinguished into chance, decision, or utility. The type of the variable is either discrete or continuous, and if it is discrete, then the mutually exclusive states of the variable should be determined. With respect to determining which variables are relevant, it is useful to take into account a number of heuristics. It is good practice to start with a simple initial model and to refine it by gradually introducing additional variables in order of importance until the model is accurate enough. Too complex models will result in the estimation of huge numbers of probabilities during parameter estimation and often obscures how the model operates (Druzdzel et al., 1999). One way to quickly zoom in on relevant variables is the *overkill test* (Abramson and Ng, 1993), which aims to identify questions that would get the expert to provide all relevant information and suppress all irrelevant information. Once variables and states are identified, they should pass the *clarity test*; i.e., it should be explicitly questioned whether the definition is precise enough to allow for later estimation of (conditional) probabilities (Druzdzel et al., 1999). One way to ensure the definition of quantifiable variables is to use concepts that follow formal domain standards.

As the number of domain variables grows, the graphical model structure can become overwhelmingly complex. This problem has been recognized by the research community and various approaches have been used to reduce this complexity. *Object-oriented Bayesian networks* (OOBNs) (Koller and Pfeffer, 1997) use an object-oriented approach analogous to the object-oriented approach in software engineering. An object in a Bayesian network is associated with a network fragment

that represents a collection of attributes that may themselves be defined in terms of such fragments. A class in a Bayesian network is then simply a fragment that is not associated with an object. This object-oriented approach has several advantages:

- *Generalization*: Classes allow network fragments to be reused for multiple objects.
- *Encapsulation*: The internal details of a class are encapsulated within that class.
- *Reusability*: The inheritance hierarchy over classes provides for an is-a hierarchy over objects that supports reusability.
- *Modularization*: The ability to enclose objects within objects allows for a part-of hierarchy over objects.

Similar ideas of object-orientation can be found in the work on network fragments as defined by Laskey and Mahoney (Laskey and Mahoney, 1997). One difference between OOBNs and network fragments is the way in which the combination of Bayesian network structures is handled. In the former case, a random variable always belongs to one particular object, and objects are combined by defining interfaces to variables internal to an object. In the latter case, a random variable may belong to multiple network fragments, where fragments are combined by defining suitable combination functions that combine the distributions for random variables that belong to multiple fragments. The idea of modularization has also been exploited in work on hierarchical model-based diagnosis (Srinivas, 1994). Here, a top-down approach of model construction is advocated where increasingly detailed subsystems are added to a hierarchical structure. This makes it possible to focus on the global aspects of model architecture in early stages of model construction deferring the modeling of details to later stages. In our research, we have found object-oriented Bayesian networks particularly useful in order to structure our domain models, as demonstrated by Fig. A.2 in Appendix A.

3.3.2 Structure specification

The construction of a Bayesian network for clinical decision support is a difficult undertaking, and the most important directive is to *keep it simple*. Simple models can gradually be extended to more complex models by adding detail to small domain fragments and evaluating the functionality of this fragment. Starting with complex models on the other hand makes it virtually impossible to evaluate functionality, since distant variables may interact in a complex way. A useful starting point when constructing a model for clinical decision support, is to first construct the patient model, which represents disease progression *without interventions*, and to subsequently construct the physician model, which represents the interventions made by the decision maker.

If a DBN is used, then we also distinguish the prior model and transition model, and need to choose an initial time, an interval, and a horizon for the model. These choices should be motivated both by the properties of the domain (i.e., we need to be able to model the processes on the time-scale in which we are interested) and by considerations with regard to available domain knowledge (i.e., domain experts need to be able to express the knowledge that is required to specify the model).

Refining a patient model $\mathcal{M}^{\pi}$

One way to refine the structure of a patient model is by means of *extension*, which is the notion that we (partially) explain the influence of a variable X on another variable Y by introducing an intermediate variable Z such that (X, Z) and (Z, Y) are arcs in the graph (Fig. 3.9). For instance, suppose that the disease we are dealing with in Fig. 3.8 is acquired immuno-deficiency syndrome (AIDS). Then, assuming that health does not influence the disease, we might introduce pneumonia as a complication, that partially explains the influence between aids and health. If the influence is totally explained, then the original direct influence should be removed from the model. If the influence of multiple direct parents $\mathbf{X}$ of Y on Y are totally explained by Z then extension is also known as parent divorcing (Olesen et al., 1989).

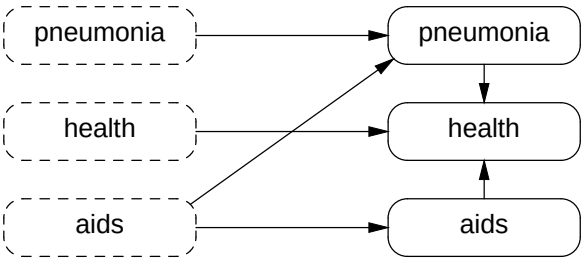


Figure 3.9: Model refinement by means of extension, where the arc between aids and health may be removed if pneumonia fully explains their dependence.

Another way to refine the model is by means of *decomposition*, which is the notion that we decompose a variable X into constituents $X_1, \ldots, X_n$. For instance, the complication pneumonia of Fig. 3.9 could be decomposed into the variables microbe and location, since the cause of infection (microbe), as well as the location of infection in the lungs, are important components of pneumonia (Fig. 3.10).

A third way to refine a model is *state revelation*, which adds observable variables $O \in \mathcal{O}$ to the model, that (partially) reveals the state of unobservable variables. Consider for instance the further refinement of pneumonia in Fig. 3.11.

Extension, decomposition, and state revelation are methods to incrementally add random variables to the model. Once sufficient detail has been added, it becomes useful to focus on *context-specific independencies* that may hold between the states of random variables (Boutilier et al., 1996b).

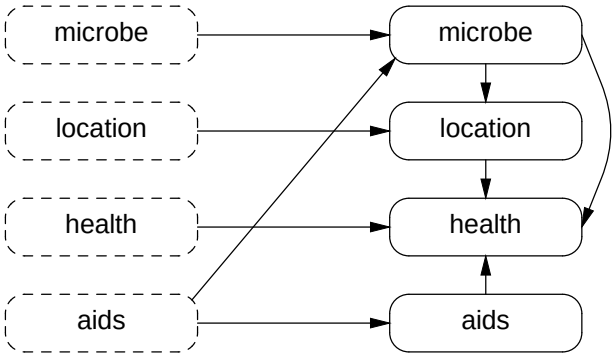


Figure 3.10: Model refinement by decomposing pneumonia into microbe and location.

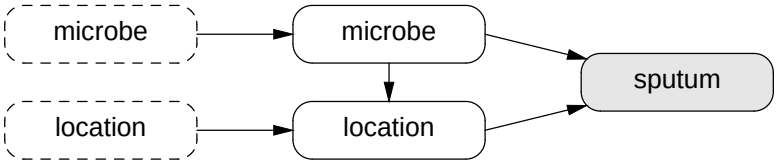


Figure 3.11: Pneumonia is characterized by the unobservable variables microbe and location, but the state may be partially revealed by analyzing a sputum sample.

Definition 3.5. *Let* $\mathbf{U}, \mathbf{V}, \mathbf{W} \subseteq \mathbf{X}$ *be disjoint subsets of a set of random variables* $\mathbf{X}$*, and let* ϕ *be a Boolean formula over variables in* $\mathbf{W}$*, where literals are of the form* $(W = w)$ *or* $\neg(W = w)$*. Then,* $\mathbf{U}$ *is said to be* contextually independent *of* $\mathbf{V}$ *given* $\mathbf{W}$ *and* ϕ*, denoted by* $\mathbf{U} \perp\!\!\!\perp_P \mathbf{V} \mid \mathbf{W}, \phi$*, iff*

$$P(\mathbf{U} \mid \mathbf{V}, \mathbf{W}, \phi) = P(\mathbf{U} \mid \mathbf{W}, \phi)$$

whenever $P(\mathbf{V}, \mathbf{W}, \phi) > 0$*.*

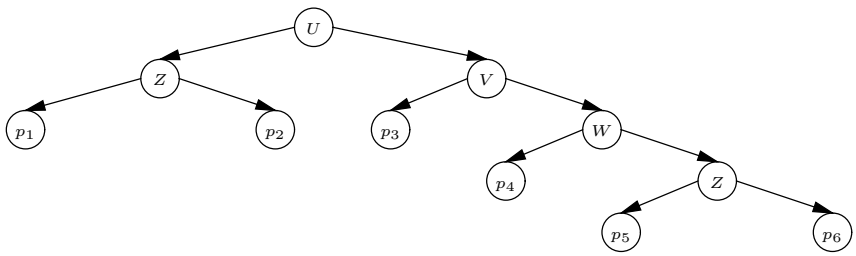


Figure 3.12: A probability tree for $P(Y \mid U, V, W, Z)$ with probabilities $p_1, \ldots, p_6$.

As an example of context-specific independence, consider the conditional probability distribution of $P(Y \mid U, V, W, Z)$. If variables are binary then we need to specify 2^5 conditional probabilities. However, it may well be the case that a considerable amount of structure is present in the table, which can be represented in

terms of a probability tree (Fig. 3.12) that expresses the following context-specific independencies:

- $Y \perp\!\!\!\perp_P \{V, W\} \mid \{U, Z\}, u \wedge z$
- $Y \perp\!\!\!\perp_P \{W, Z\} \mid \{U, V\}, \neg u \wedge v$
- $Y \perp\!\!\!\perp_P Z \mid \{U, V, W\}, \neg u \wedge \neg v \wedge w$

Using these context-specific independencies, we need to specify just the six probabilities shown in Fig. 3.12, which is a substantial improvement. Context-specific independence can be represented within a Bayesian network by means of a recursive decomposition using multiplexer nodes (Boutilier et al., 1996b) or by means of an additive factorization that employs hidden variables (van Gerven, 2006). These representations not only allow for a more compact specification but additionally for more efficient probabilistic inference.

Model refinement should always be well-motivated, where valid reasons are (1) when the refinement has a significant impact on the posterior distributions of our query variable(s), (2) when the refinement alters model structure in such a way that it increases model intelligibility, or (3) when the refinement leads to a more compact factorization of the JPD. Algorithm 3.2 summarizes the strategy for constructing a transition model.

Algorithm 3.2 Construction of a Bayesian network.

1. Start with a basic model that includes the desired solutions in Σ.
2. Try to refine the model by decomposing a variable X into constituents $X_1, \ldots, X_n$.
3. Try to extend the model by adding a variable Z, such that for all $X \in \mathbf{X}$, arcs $(X, Y) \in A(G)$ are (partially) explained by (X, Z) and (Z, Y) in $A(G)$.
4. Try to add observable variables $O \in \mathcal{O}$ to the model that (partially) reveal the state of unobserved variables in $\Sigma \cup \mathcal{U}$.
5. While the model is incomplete, return to step 2.
6. Try to reduce the number of free parameters by taking context-specific independencies into account.

Specifying a physician model $\mathcal{M}^\phi$

Once a patient model has been completed, we proceed with the specification of decision making by the physician. The construction of a physician model should always start with the definition of the treatment protocol that is used in clinical practice. The treatment protocol describes exactly which treatment is applied in which situation, and if dependencies between treatments exist. Treatments may be represented in two alternative ways:

- One random variable for each treatment X, where Ω_X represents the different possibilities for X, and where X is conditioned on all its required preconditions.

- One random variable X, whose states represent all *possible* mutually exclusive treatment combinations, and where X is conditioned on all the required preconditions of every treatment combination.

For example, for pneumonia, we may choose to represent each antibiotic as a separate random variable, or to represent each possible antibiotic combination as that state of a random variable. Assuming that we decide for the last option, disregarding patient characteristics, a protocol may be determined by the structure in Fig. 3.13.

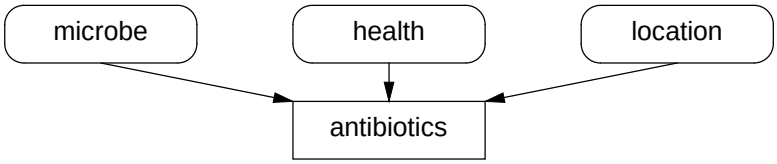


Figure 3.13: Representation of a protocol for treatment of pneumonia.

It is assumed in the above that the states of microbe, health, and location are fully observable, since otherwise, the protocol is represented by a stochastic policy, which may lead to arbitrarily poor results (Singh et al., 1994). If the state of a relevant variable is unobservable then we condition instead on indirect observations of these states; in our case, by explicitly representing the sputum history, as in Fig. 3.14.

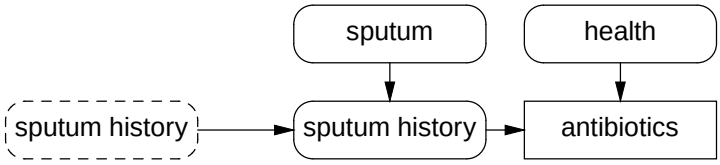


Figure 3.14: An alternative representation of a protocol for treatment of pneumonia.

For more elaborate treatments, such as the prolonged administration of medication, modeling of the protocol can quickly become more complex. For example, if treatments can only be started when other treatments have failed, then this failure should be represented explicitly in the model, and if treatments can only be given for a maximum amount of time, then the treatment history should become part of the model.

The effect of treatment can often be distinguished into a positive effect on the target of treatment, and a negative effect on patient health. Figure 3.15 depicts these effects for antibiotics treatment. Once disease progression is represented in enough detail by the patient model, and the treatment protocol is captured by the physician

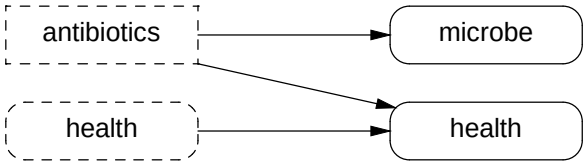


Figure 3.15: Positive and negative effects of antibiotics treatment.

model, we have completed the specification of a Bayesian network structure for clinical decision support.

3.3.3 Factor association

Factor association refers to the association of a factor with each random variable that defines the functional form of how the outcome of the random variable depends on its parent variables. This is an important step since it determines the number of parameters that need to be estimated subsequently. We restrict the discussion to discrete random variables and assume that continuous quantities have been discretized a priori. For discrete random variables, a factor can be thought of as a (conditional) probability table (CPT), which is a mapping $\gamma\colon \Omega_Y \times \Omega_{\mathbf{X}} \rightarrow [0,1]$ such that $\sum_y \gamma(y,\mathbf{x}) = 1$, for a random variable Y and a (possibly empty) set of parents $\mathbf{X}$.

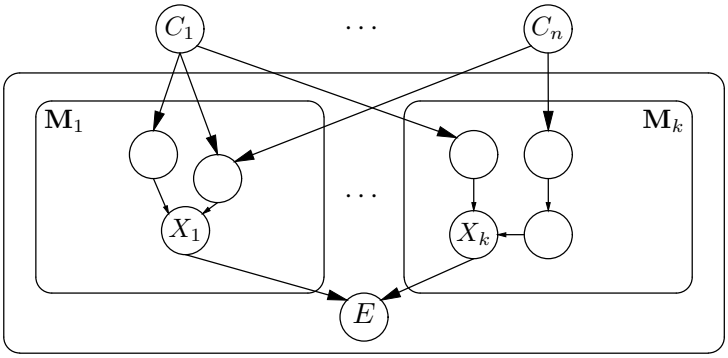


Figure 3.16: A causal interaction model, where causes C_j, $1 \leq j \leq n$ may take part in multiple mechanisms that lead to the effect. Each mechanism $\mathbf{M}_i$, $1 \leq i \leq k$ has an associated intermediate variable X_i, that partially determines the effect E through a deterministic function $f\colon \Omega_{\mathbf{X}} \rightarrow \Omega_E$. In an object-oriented approach, the internal details of how mechanisms interact to produce the effect can be private to an object and hidden for other objects.

One way to reduce the size of this mapping is to determine how causes (parent variables) interact in order to produce the effect (child variable). Meek and Heckerman formalized this idea in terms of *causal interaction models* (Meek and Heckerman, 1997), where a cause may be associated with several mechanisms and multiple causes may be associated with a single mechanism, and combine deterministically

using a deterministic interaction function f (Fig. 3.16). Although causal interaction models allow arbitrarily complex mechanisms and interaction functions, the most widely employed causal interaction models are *causal independence models*, which assume that mechanisms $\mathbf{M}_i$ are given by the intermediate variables X_i, and for which it holds that:

$$P(e \mid \mathbf{c}) = \sum_{\mathbf{x}:\, f(\mathbf{x})=e} \prod_{i=1}^{n} P(x_i \mid c_i)$$

The theory of causal independence adopts specific independence assumptions to model the interactions between a set of cause variables and an effect variable, and using this approach, the number of parameters that need to be estimated decreases from exponential to linear in the number of variables. The most widely employed causal independence model is the *noisy-max* model, which specializes to the *noisy-or* model for binary random variables (Good, 1983; Pearl, 1988; Henrion, 1989; Díez, 1993). This model expresses that the presence of one or more causes $C_1, \ldots, C_n$ is sufficient to give rise to the occurrence of an effect E, and has been used for instance in the QMR-DT system, and the Promedas system. As an example of a noisy-or model, consider a disease D that may have multiple causes $\mathbf{C} = \{C_1, \ldots, C_n\}$, and where state spaces are given by $\{\textit{true}, \textit{false}\}$. Then

$$P(D = \textit{true} \mid \mathbf{C}) = 1 - \prod_{i=1}^{n} P(X_i = false \mid C_i)$$

and requires the specification of $2n$ instead of 2^n free parameters. Since it is often assumed that absent causes do not contribute to the effect, we obtain a further reduction to just n free parameters. Another frequently used CI model is the *noisy-and* model; it expresses that all causes must be present in order to give rise to the effect. It has, for example, been used to model the joint effect of antibiotics on bacteria causing ventilator-associated pneumonia in patients (Lucas et al., 2000).

3.3.4 Parameter estimation

Once factors have been attached, the final task is to estimate the parameters that complete the distributions. One way to estimate parameters is to learn them from data. However, in practice, datasets can be too small or of too poor quality to yield accurate estimates for the desired quantities (Korver and Lucas, 1993; Jensen, 1995). Small datasets can be a consequence of the prohibitive costs of obtaining the data, undisclosed data due to data privacy issues, or a *data-poor domain*; a domain whose properties forbid the accumulation of enough data. This phenomenon can be observed, for instance, when the prevalence of a disease one is about to model is low. Data quality may be compromised in a number of ways such as missing values, measurement errors, selection effects, and data which is not independently sampled and identically distributed (i.i.d). An example of a violation of the i.i.d. assumption can

be found in (Lucas, 2004), where the evolution of a treatment protocol introduces a systematic bias in the data.

An alternative way to estimate parameters when data is lacking is to acquire them from available domain literature. Although much probabilistic information can be obtained in this way, often the information is incomplete (Druzdzel et al., 1995). For example, although the probability of a symptom in the presence of a disease is often mentioned, the probability of a symptom in the absence of a disease is not.

Finally, parameters may be estimated by eliciting them from experts. This is a subject that should be treated with care, where both statistical (Savage, 1971) and psychological aspects (Kahneman et al., 1982; Baron, 1994) should be taken into account. The subject of probability elicitation is treated in detail in (O'Hagan et al., 2006; Jenkinson, 2005). In the context of expert systems, research has focused on the fast elicitation of many probabilities. Reference lotteries, for instance, are very time consuming and less appropriate due to the large number of parameters involved (van der Gaag et al., 1999). A good strategy seems to be the combined use of linguistic and numerical anchors (Renooij and Witteman, 1999) for the assessment of probabilities (Fig. 3.17).

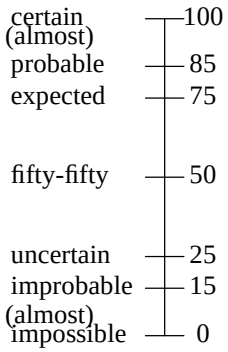


Figure 3.17: Assessment of probabilities by means of linguistic and numerical anchoring.

Other approaches to parameter estimation are the estimation of distributions based on qualitative constraints (Feelders and van der Gaag, 2006) or the completion of partially specified models based on maximum entropy arguments (Wiegerinck and Heskes, 2001; Wiegerinck, 2005). A modeling strategy that is much used in practice is to build the structural part of the underlying graphical model based on expert knowledge and domain literature, whereas parameters are estimated from statistical data (Druzdzel and Díez, 2003).

With respect to decision making, we need to model how decisions influence domain variables (being part of $\mathcal{M}^{\pi}$), as well as how the decision making strategy is influenced by domain variables (being part of $\mathcal{M}^{\phi}$). The strategy is often given by some deterministic policy as is dictated for instance by medical guidelines. Stochasticity in the policy can be useful in some cases, such as when the time of treatment is

uncertain, when we choose randomly between treatments that have the same expected utility, or when conditioning variables remain unmodeled. If utilities that capture outcome preference need to be assessed then we may resort to decision analytical techniques such as *direct scaling* and the *standard reference gamble* (Sox et al., 1988). The robustness of the assessed probabilities and utilities can be determined by means of *sensitivity analysis* (Morgan and Henrion, 1990), which amounts to the systematic variation of probabilities and an analysis of its effects; it is discussed in the context of Bayesian networks in (Coupé et al., 1999).

3.4 Summary

Bayesian network construction for clinical decision support is a difficult task, especially if causality, decision making, and the dynamics of a problem need to be taken into account. To date, guidelines for the construction of such Bayesian networks have remained scarce, and the aim of this chapter is to contribute to these guidelines. We have described clinical tasks in terms of problem solving methods and discussed the different design choices that can be made for Bayesian networks that are used for clinical decision support. Subsequently, we have discussed the steps that need to be taken when constructing realistic Bayesian networks that capture disease dynamics in terms of a patient model and a treatment protocol in terms of a treatment model. It is hoped that this work aids the knowledge engineer who is faced with the construction of a Bayesian network for clinical decision support.

Chapter 4

A Qualitative Characterization of Causal Independence

In designing Bayesian networks, developers try to create acyclic directed graphs that are as sparse as possible, as the size of a conditional probability table is exponential in the number of associated variables. Creating sparse graphs not only saves space, but may also speed up probabilistic inference. Unfortunately, the creation of sparse graphs for a given problem may not always be possible. However, by imposing extra independence assumptions, supplemented by assumptions of functional dependence, it may be possible to reduce the number of conditional probabilities that need to be assessed. The theory of *causal independence* is especially suited for this purpose (Heckerman and Breese, 1996).

The theory of causal independence adopts specific independence assumptions to model the interactions between a set of cause variables and an effect variable; using this approach, the number of parameters that need to be estimated decreases from exponential to linear in the number of variables. The noisy OR model, that expresses that the presence of one or more causes is sufficient to give rise to the occurrence of the effect, is an example of a causal independence model that is widely used in practice (Good, 1983; Henrion, 1989; Díez, 1993). It has been used in the QMR-DT system, which includes knowledge of approximately 600 diseases and approximately 4000 findings (Shwe et al., 1991), the Promedas system, which aims to cover a large diagnostic repertoire of internal medicine (Kappen and Neijt, 2002), and in DIAVAL, an expert system for electrocardiography that uses a generalization of the noisy OR for non-binary random variables (Díez et al., 1997). Another, frequently used causal independence model is the noisy AND model; it expresses that all causes must be present in order to give rise to the effect. It has, for example, been used to model the joint effect of antibiotics on bacteria causing ventilator-associated pneumonia in patients (Lucas et al., 2000).

The noisy OR and noisy AND models are special cases of causal independence models based on Boolean functions; any of the 2^{2^n} possible n-ary Boolean functions

This chapter is based on (van Gerven et al., 2005, 2006c).

can be used to model deterministic interactions between cause and effect variables. Given the favorable properties of causal independence models, it is unfortunate that only very few of these are used in practice: only the mentioned noisy OR and noisy AND are popular amongst developers. This is caused by the fact that it is often unclear with what behavior a particular causal independence model is endowed when choosing a particular Boolean function. In (Lucas, 2005) this problem was addressed by exploiting *qualitative probabilistic network* (QPN) theory to characterize the behavior of causal independence models in terms of *influences* and *synergies* (Wellman, 1990). Such a qualitative characterization may then be matched with the behavior that is dictated by the domain, as suggested in Figure 4.1. The qualitative pattern associated with a particular causal independence model is termed a *qualitative causal pattern*.

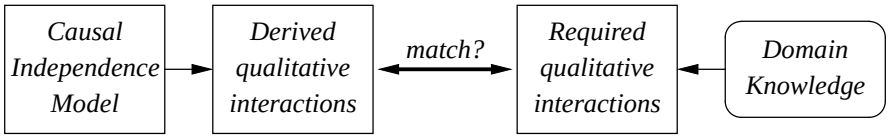


Figure 4.1: Comparing the observed qualitative behavior of a causal independence model with the desired qualitative behavior as specified by a domain expert.

The idea that QPN theory might be suitable for analyzing the behavior of causal independence models was recognized by Wellman, who states that: *"...prototypical patterns of systematic interaction might alleviate the burden of specifying qualitative synergies"* and *" ...we should expect non-ambiguous synergy results from canonical models because any representation that specifies an n-way influence in terms of $O(n)$ parameters must employ some systematic assumption about interactions"* (Wellman, 1990). However, (Lucas, 2005) offers the first systematic approach to analyzing causal independence models in terms of QPN theory. This was done in particular for decomposable causal independence models, i.e., causal independence models which are characterized in terms of binary functions. There are 16 binary Boolean functions, which can be used to compose a subset of n-ary Boolean function, and which can be classified in terms of presence or absence of the properties of *associativity* and *commutativity*. The previously discussed noisy-OR model is based on the Boolean OR, which is both commutative and associative. Although this offers an analysis of a useful subset of Boolean functions, a general characterization of the behavior of Boolean functions is not provided by Ref. (Lucas, 2005).

This chapter offers a substantial generalization of previously published results as it develops a general theory of qualitative causal patterns. The theory identifies:

1. The qualitative behavior that holds for a given causal independence model.

2. Properties of causal independence models that hold given a qualitative specification.

The theory developed in this chapter is useful in Bayesian network design, as it provides a tool for matching desired qualitative behavior of causal independence models with the appropriate structural and quantitative parameters. Furthermore, a more widespread use of causal independence models in Bayesian networks will facilitate the intelligibility of network behavior, allow the construction of denser networks and ease the estimation of network parameters.

The structure of this chapter is as follows. In Section 4.1 we review some necessary preliminaries, drawing upon Bayesian network, causal independence and QPN theory. Subsequently, we study some general properties of causal independence models in Section 4.2. These properties are then used to identify the qualitative behavior for different Boolean functions in Section 4.3. Finally, in Section 4.4 we round off with a discussion of the obtained results.

4.1 Preliminaries

In this section we will subsequently discuss Bayesian networks, causal independence models, the running example of this chapter, and QPN theory. Throughout, it is assumed that all random variables are binary. We will use x to denote $X = \top$ (logical truth) and $\bar{x}$ to denote $X = \bot$ (logical falsehood). If the value of variable X is either true of false, but unspecified, then this is indicated by $X = \hat{x}$, or simply by $\hat{x}$.

4.1.1 Causal Independence Models

Causal independence is the notion that causes are independently contributing to the occurrence of an effect through some pattern of interaction, represented as a set of local conditional probability distributions of a Bayesian network (Heckerman and Breese, 1996). The associated Bayesian network structure is depicted in Figure 4.2, where variables C_k indicate cause variables, M_k intermediate variables and E is an effect variable. Let $\mathbb{B} = \{\bot, \top\}$. We use $\mathbf{c} \in \mathbb{B}^n$, possibly with a subscript, to denote an element of $\mathbb{B}^n$ for vectors $\mathbf{C} = (C_1, \ldots, C_n)$; similarly, we use $\mathbf{m} \in \mathbb{B}^n$ for elements of $\mathbf{M} = (M_1, \ldots, M_n)$. These are called *configurations*. To reduce the use of numeric indices, we associate with each cause variable C an intermediate variable M_C. The notion of causal independence is captured by the requirement that an intermediate variable $M_C \in \mathbf{M}$ is dependent of cause variable C and independent of the other cause variables $\mathbf{C} \setminus \{C\}$. According to the independence structure shown in Figure 4.2, it holds that:

$$\begin{aligned} P(e \mid \mathbf{c}) &= \sum_{\mathbf{m}} P(e \mid \mathbf{m}) P(\mathbf{m} \mid \mathbf{c}) \\ &= \sum_{\mathbf{m}} P(e \mid \mathbf{m}) \prod_{i=1}^{n} P(\hat{m}_i \mid \hat{c}_i). \end{aligned} \tag{4.1}$$

An intermediate variable M_C can be interpreted as modulating the contribution of a cause C to the effect E and often specific assumptions are made about this contribution. Here we formalize this by the notions of *consequentiality* and *accountability*. Consequentiality states that the truth of a cause variable increases our belief that the associated intermediate variable is true as well. Formally, we require that $P(m_C \mid c) > 0$. Accountability states that the truth of an intermediate variable must imply the truth of its associated cause variable; formally, $P(m_C \mid \bar{c}) = 0$. The conditional probability distribution $P(E \mid \mathbf{M})$ used in Eq. (4.1) is assumed to be deterministic in causal independence models, and, thus, can be taken as representing a function $f\colon \mathbb{B}^n \to \mathbb{B}$, such that $P(e \mid \mathbf{m}) = 1$ if $f(\mathbf{m}) = \top$ and $P(e \mid \mathbf{m}) = 0$ otherwise. A causal independence model, or CI model, is now defined formally as follows:

Definition 4.1. *A* causal independence model $\mathcal{C}$ *is a tuple* $(\mathbf{C}, \mathbf{M}, E, f, \mathbf{P})$*, where* $\mathbf{C}$ *is a set of* cause variables*,* $\mathbf{M}$ *is a set of* intermediate variables*,* E *is an* effect variable*,* f *is an* interaction function*, and* $\mathbf{P}$ *is a set* $\{P(M_C \mid C) \mid C \in \mathbf{C}\}$ *of parameters, with* $M_C \in \mathbf{M}$*, for each* $C \in \mathbf{C}$ *and vice versa, such that:*

$$P(e \mid \mathbf{c}) = \sum_{f(\mathbf{m})=e} \prod_{i=1}^{n} P(\hat{m}_i \mid \hat{c}_i). \tag{4.2}$$

By $f(\mathbf{m}) = e$ is denoted the situation where both $f(\mathbf{m}) = \top$ and $E = \top$ hold. The probability $P(m_C \mid c)$ will often be abbreviated to $P(m \mid c)$. In literature different interpretations of causal independence exist, often taking the form of restrictions on an interaction function f that underlies the model (Cozman, 2004; Heckerman and Breese, 1996). Here, we assume that an interaction function can be any Boolean function $f : \mathbb{B}^n \to \mathbb{B}$.

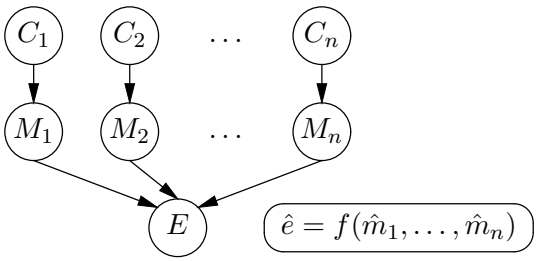


Figure 4.2: Causal independence model.

A causal independence model $\mathcal{C} = (\mathbf{C}, \mathbf{M}, E, f, \mathbf{P})$ can act as the basis for the specification of a Bayesian network $\mathcal{B} = (G, P)$, with ADG $G = (\mathbf{V}, \mathbf{E})$, as depicted in Figure 4.2, and joint probability distribution P, where G respects all the dependences represented by the joint probability distribution P. The vertices in G are given by

$$\mathbf{V} = \mathbf{C} \cup \mathbf{M} \cup \{E\}$$

such that the sets $\mathbf{C}, \mathbf{M}$ and $\{E\}$ are disjoint, and the arcs in G are given by

$$\mathbf{E} = \{(C, M_C) \mid C \in \mathbf{C}\} \cup \{(M, E) \mid M \in \mathbf{M}\}.$$

In addition to the parameters $P(M_C \mid C)$ and the interaction function f, we also need to specify a prior joint probability distribution $P(\mathbf{C})$ to obtain a complete specification of the Bayesian network $\mathcal{B}$.

In the sequel, we will often use the notation $P\,[f]$ to refer to the probability distribution $P(E \mid \mathbf{c})$; in this chapter it is assumed that both consequentiality and accountability hold. We can alternatively write Eq. (4.2) in somewhat generalized form as:

$$P\,[f]\,(e \mid \mathbf{c}) = \sum_{\mathbf{m}} f(\mathbf{m}) P(\mathbf{m} \mid \mathbf{c}) = \sum_{\mathbf{m}} f(\mathbf{m}) \prod_{i=1}^{n} P(\hat{m}_i \mid \hat{c}_i), \tag{4.3}$$

where we make use of the analogy between Boolean algebra and ordinary arithmetic by interpreting $\bot$ as 0 and $\top$ as 1, i.e., if $f(\mathbf{m}) = \bot$ this is interpreted as $f(\mathbf{m}) = 0$, and as $f(\mathbf{m}) = 1$ otherwise (Birkhoff and Mac Lane, 1997). We will sometimes employ functions f that are not Boolean; even then Eq. (4.3) still applies, where $P\,[f]\,(e \mid \mathbf{c})$ can be interpreted as the conditional expectation of f given $\mathbf{c}$. If f is a constant and there are no cause variables $\mathbf{C}$ then $P[f] = f$.

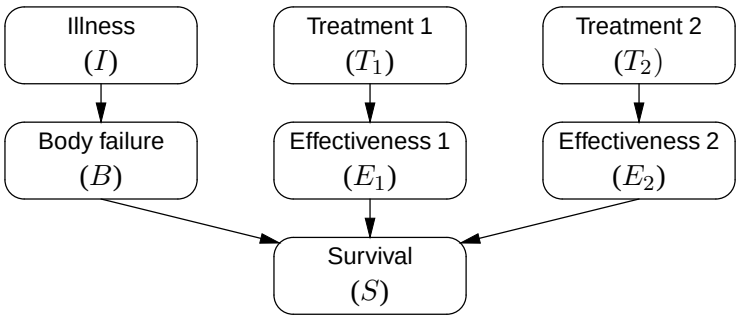


Figure 4.3: A prognostic model of survival in serious illness, modeling the interaction between two drugs, expressed as a causal independence model.

As an example of a realistic causal independence model that will be used to illustrate the theory developed in this chapter, consider the causal independence model shown in Figure 4.3 that represents a piece of medical knowledge with respect to the prognosis of a *serious illness* (I), such as malignant hypertension due to chronic kidney infection, infectious hypertension for short, which is handled by two alternative treatments T_1, an antihypertensive drug, and T_2, rifampin (an antibiotic).[1] The seriousness of the infectious hypertension is reflected by the fact that we are interested

[1]The choice of these drugs was inspired by the death of Slobodan Miloševic. It is hypothesized that his death was due to the combined effect of an antihypertensive drug, which was meant to reduce the height of his blood pressure, and the antibiotic rifampin, which counteracted the effect of the antihypertensive drug. Here, we abstract from the course of events.

in the *survival* (S) (e.g., within the next 5 years) of a patient with this illness. The resulting causal independence model is shown in Figure 4.3. The variable B stands for *body failure* due to the illness, E_1 stands for the *effectiveness* of treatment T_1 and E_2 for the effectiveness of treatment T_2. If body failure occurs and the disease cause is eradicated, it is assumed that the patient will survive. However, if both treatments T_1 and T_2 are effective then the patient will not survive due to the synergistic interaction between the two treatments (rifampin in conjunction with the antihypertensive drug). This can be expressed by means of a Boolean function f defined by the following Boolean expression:

$$\hat{s} = (\neg\hat{b} \wedge \neg\hat{e}_1 \wedge \neg\hat{e}_2) \vee (\hat{e}_1 \wedge \neg\hat{e}_2) \vee (\neg\hat{e}_1 \wedge \hat{e}_2) \tag{4.4}$$

(survival is equivalent to the absence of body failure or eradication of the disease due either treatment T_1 or T_2, but not both). In the sequel, we will use Boolean functions and Boolean expressions interchangeably. The qualitative behavior that arises from this choice should then be in accordance with the domain knowledge as stated above.

According to what has been said above, the Bayesian network model is an example of a CI model. It will be called the *prognostic model* in the following. Here, the variables I, T_1 and T_2 act as cause variables and B, E_1 and E_2 are the intermediate variables. For example, we have that $B = M_I$.

There are two main tasks in building a CI model. The first is to determine the underlying interaction function f, in the example a Boolean function that is assumed to model the interaction between the factors Body Failure (B), Effectiveness 1 (E_1) and Effectiveness 2 (E_2) with respect to Survival (S), where S is the effect variable. The second task is to estimate the parameters $P(B \mid I)$, $P(E_1 \mid T_1)$ and $P(E_2 \mid T_2)$. Notice that just three conditional probabilities need to be estimated, as $P(m_C \mid \bar{c})$ is assumed to be zero for each cause variable C. Examples of causal independence models that model other real-world problems and employ alternative interaction functions can be found in (Lucas, 2005).

4.1.2 Qualitative Probabilistic Networks

Recall that the aim of the research underlying this chapter was to develop a theory that is able to assist Bayesian network developers in quantifying Bayesian networks using qualitative knowledge from a problem domain. Qualitative probabilistic networks are at the core of this theory. We will, therefore, briefly summarize the theory of qualitative probabilistic networks.

Qualitative probabilistic networks (QPNs) were introduced by Wellman as a qualitative abstraction of ordinary Bayesian networks (Wellman, 1990). The relationships between variables are described by the concepts of *influences* and *synergies*. In the following, let (G, P) be a Bayesian network, let $A, B, C \in \mathbf{X}$ be binary random variables and let (A, C) and (B, C) be arcs in G.

A qualitative influence expresses how the value of one variable influences the probability of observing values for another variable.

Definition 4.2. *Let* $\mathbf{X}$ *denote* $\pi_G(C) \setminus \{A\}$. *We say that there is a* positive qualitative influence *of* A *on* C*, written as* $\delta_{A \to C} = +$*, if*

$$\delta_{A \to C}(\mathbf{x}) = P(c \mid a, \mathbf{x}) - P(c \mid \bar{a}, \mathbf{x}) \geq 0$$

regardless of the configuration $\mathbf{x}$*, with a strict inequality for at least one configuration* $\mathbf{x}$. Negative *(*$\delta_{A \to C} = -$*) and* zero *qualitative influences (*$\delta_{A \to C} = 0$*) are defined analogously, replacing* $\geq$ *by* $\leq$ *and* $=$ *respectively. If there are values* $\mathbf{x}$ *and* $\mathbf{x}'$*, such that*

$$P(c \mid a, \mathbf{x}) - P(c \mid \bar{a}, \mathbf{x}) > 0 \text{ and } P(c \mid a, \mathbf{x}') - P(c \mid \bar{a}, \mathbf{x}') < 0$$

then we say that the qualitative influence is non-monotonic, *denoted by* $\delta_{A \to C} = \sim$*. If none of these cases hold, i.e., when there is incomplete information about the probability distribution, then we say that the qualitative influence is* ambiguous, *written as* $\delta_{A \to C} = ?$.

Example 4.1. In order to illustrate the qualitative concepts we assume for the moment that the exact probabilities associated with the prognostic model are known. We assume $P(b \mid i) = 0.9$, $P(e_1 \mid t_1) = 0.3$ and $P(e_2 \mid t_2) = 0.6$. Hence, it is very likely that a serious illness gives rise to body failure, as it occurs in 90% of cases, treatment T_1 is effective in 30% of the patients and treatment T_2 is effective in 60% of the patients. What then, we might ask, is the qualitative influence of a serious illness on the survival? This is computed as follows, where the Boolean function f is defined by the Boolean expression (4.4):

$$\begin{aligned} \delta_{I \to S}(\{\hat{t}_1, \hat{t}_2\}) &= P[f](s \mid i, \hat{t}_1, \hat{t}_2) - P[f](s \mid \bar{\imath}, \hat{t}_1, \hat{t}_2) \\ &= P(\bar{e}_1 \mid \hat{t}_1) P(\bar{e}_2 \mid \hat{t}_2) \big(P(\bar{b} \mid i) - P(\bar{b} \mid \bar{\imath})\big). \end{aligned}$$

It follows that $\delta_{I \to S}(\{t_1, t_2\}) = -0.252$, $\delta_{I \to S}(\{\bar{t}_1, t_2\}) = -0.36$, $\delta_{I \to S}(\{t_1, \bar{t}_2\}) = -0.63$ and $\delta_{I \to S}(\{\bar{t}_1, \bar{t}_2\}) = -0.9$. In accordance with our expectations, serious illness appears to have a negative influence on survival.

An additive synergy expresses how the interaction between two variables influences the probability of observing values for a third variable.

Definition 4.3. *Let* $\mathbf{X}$ *denote* $\pi_G(C) \setminus \{A, B\}$. *We say that there is a* positive additive synergy *of* A *and* B *on* C*, written as* $\delta_{(A,B) \to C} = +$*, if*

$$\delta_{(A,B) \to C}(\mathbf{x}) = P(c \mid a, b, \mathbf{x}) + P(c \mid \bar{a}, \bar{b}, \mathbf{x}) - P(c \mid \bar{a}, b, \mathbf{x}) - P(c \mid a, \bar{b}, \mathbf{x}) \geq 0$$

regardless of the configuration $\mathbf{x}$*, with a strict inequality for at least one configuration* $\mathbf{x}$*. Negative, zero, non-monotonic and ambiguous additive synergies are defined analogous to qualitative influences.*

Example 4.2. With regard to the prognostic model, we might be interested in the additive synergy between serious illness and treatment T_1 with respect to survival. This is computed as follows, where again we employ Boolean expression (4.4):

$$\begin{aligned}\delta_{(I,T_1)\rightarrow S}(\{\hat{t}_2\}) &= P[f](s \mid i, t_1, \hat{t}_2) + P[f](s \mid \bar{\imath}, \bar{t}_1, \hat{t}_2) - \\ &\quad P[f](s \mid \bar{\imath}, t_1, \hat{t}_2) - P[f](s \mid i, \bar{t}_1, \hat{t}_2) \\ &= P(\bar{e}_2 \mid \hat{t}_2)\big(P(\bar{b} \mid i) - 1\big)\big(P(\bar{e}_1 \mid t_1) - 1\big).\end{aligned}$$

It follows that $\delta_{(I,T_1)\rightarrow S}(\{t_2\}) = 0.108$ and $\delta_{(I,T_1)\rightarrow S}(\{\bar{t}_2\}) = 0.27$ such that illness I and treatment T_1 have a positive additive synergy with respect to survival.

A product synergy expresses how upon observation of a common child of two vertices, observing the value of one parent vertex influences the probability of observing a value for the other parent vertex. The original definition of a product synergy is as follows (Henrion and Druzdzel, 1991).

Definition 4.4. *Let* $\mathbf{X}$ *denote* $\pi_G(C) \setminus \{A, B\}$. *We say that there is a* positive product synergy *of* A *and* B *with regard to the value* $\hat{c}$ *of variable* C*, written as* $\delta^{\hat{c}}_{(A,B)\rightarrow C} = +$*, if*

$$\delta^{\hat{c}}_{(A,B)\rightarrow C}(\mathbf{x}) = P(\hat{c} \mid a, b, \mathbf{x})P(\hat{c} \mid \bar{a}, \bar{b}, \mathbf{x}) - P(\hat{c} \mid \bar{a}, b, \mathbf{x})P(\hat{c} \mid a, \bar{b}, \mathbf{x}) \geq 0$$

regardless of the configuration $\mathbf{x}$*, with a strict inequality for at least one configuration* $\mathbf{x}$*. It is assumed that the value* $\hat{c}$ *of variable* C *is either true or false. Negative, zero, non-monotonic and ambiguous product synergies are again defined analogous to the corresponding types of qualitative influences.*

Example 4.3. With regard to the prognostic model, the product synergy between treatments T_1 and T_2 in the case of survival, is computed as follows.

$$\begin{aligned}\delta^{s}_{(T_1,T_2)\rightarrow S}(\{\hat{\imath}\}) &= P[f](s \mid \hat{\imath}, t_1, t_2) \cdot P[f](s \mid \hat{\imath}, \bar{t}, \bar{t}_2) - \\ &\quad P[f](s \mid \hat{\imath}, \bar{t}_1, t_2) \cdot P[f](s \mid \hat{\imath}, t_1, \bar{t}_2) \\ &= -P(e_1 \mid t_1)P(e_2 \mid t_2)\end{aligned}$$

It follows that $\delta^{s}_{(T_1,T_2)\rightarrow S}(\{\bar{\imath}\}) = \delta^{s}_{(T_1,T_2)\rightarrow S}(\{i\}) = 0.18$ such that treatments T_1 and T_2 have a positive product synergy with respect to survival. This positive product synergy arises due to the fact that in the case of survival of a patient, it is more likely that one of both treatments is given. The presence of both T_1 and T_2 and the absence of both T_1 and T_2 will lead to patient death.

The following lemma states that for binary random variables, the product synergy when $C = \perp$ is partially determined by the associated additive synergy.

Lemma 4.1. *For binary random variables, the product synergy when $C = \bot$ is determined by the product synergy when $C = \top$ and the additive synergy through the following equality:*

$$\delta^{\bar{c}}_{(A,B)\to C}(\mathbf{x}) = \delta^{c}_{(A,B)\to C}(\mathbf{x}) - \delta_{(A,B)\to C}(\mathbf{x}).$$

Proof.

$$\begin{aligned}
\delta^{\bar{c}}_{(A,B)\to C}(\mathbf{x}) &= P(\bar{c} \mid \bar{a}, \bar{b}, \mathbf{x})P(\bar{c} \mid a, b, \mathbf{x}) - P(\bar{c} \mid a, \bar{b}, \mathbf{x})P(\bar{c} \mid \bar{a}, b, \mathbf{x}) \\
&= (1 - P(c \mid \bar{a}, \bar{b}, \mathbf{x}))(1 - P(c \mid a, b, \mathbf{x})) - \\
&\quad (1 - P(c \mid a, \bar{b}, \mathbf{x}))(1 - P(c \mid \bar{a}, b, \mathbf{x})) \\
&= \big(P(c \mid \bar{a}, \bar{b}, \mathbf{x})P(c \mid a, b, \mathbf{x}) - P(c \mid a, \bar{b}, \mathbf{x})P(c \mid \bar{a}, b, \mathbf{x})\big) - \\
&\quad \big(P(c \mid \bar{a}, \bar{b}, \mathbf{x}) + P(c \mid a, b, \mathbf{x}) - P(c \mid a, \bar{b}, \mathbf{x}) - P(c \mid \bar{a}, b, \mathbf{x})\big) \\
&= \delta^{c}_{(A,B)\to C}(\mathbf{x}) - \delta_{(A,B)\to C}(\mathbf{x}),
\end{aligned}$$

which completes the proof. □

Modifications to the definition of a product synergy have been made after the observation that Def. 4.4 is incomplete when parent vertices in $\mathbf{X}$ are uninstantiated (Druzdzel and Henrion, 1993b,a). In other words,

$$\begin{aligned}
&\forall_{\mathbf{x}}[P(\hat{c} \mid a, b, \mathbf{x})P(\hat{c} \mid \bar{a}, \bar{b}, \mathbf{x}) - P(\hat{c} \mid a, \bar{b}, \mathbf{x})P(\hat{c} \mid \bar{a}, b, \mathbf{x}) \leq 0 \\
&\quad \nRightarrow P(\hat{c} \mid a, b)P(\hat{c} \mid \bar{a}, \bar{b}) - P(\hat{c} \mid a, \bar{b})P(\hat{c} \mid \bar{a}, b) \leq 0].
\end{aligned}$$

This so-called type II product synergy can be formalized in terms of the more intuitive notion of an *intercausal influence* (Renooij, 2001).

Definition 4.5. *Let $\mathbf{X}$ denote $\pi_G(B) \cup \pi_G(C) \setminus \{A\}$. Then a variable A exhibits a* positive intercausal influence *on B with regard to the value $\hat{c}$ if*

$$P(b \mid a, \hat{c}, \mathbf{x}) - P(b \mid \bar{a}, \hat{c}, \mathbf{x}) \geq 0,$$

regardless of the configuration $\mathbf{x}$. Negative, zero, non-monotonic and ambiguous intercausal influences are again defined analogous to the corresponding types of qualitative influences.

For causal independence models, intercausal influences describe the dependence between two causes C and C' when the value of the effect variable is observed. We, therefore, compute

$$P(c' \mid c, \hat{e}, \mathbf{c}_2) - P(c' \mid \bar{c}, \hat{e}, \mathbf{c}_2) \tag{4.5}$$

for all values $\mathbf{c}_2$ of the causes $\mathbf{C}_2 = \mathbf{C} \setminus \{C, C'\}$. Using Bayes' rule we obtain the equal expression

$$\frac{P(\hat{e} \mid c, c', \mathbf{c}_2)P(c' \mid c, \mathbf{c}_2)}{P(\hat{e} \mid c, \mathbf{c}_2)} - \frac{P(\hat{e} \mid \bar{c}, c', \mathbf{c}_2)P(c' \mid \bar{c}, \mathbf{c}_2)}{P(\hat{e} \mid \bar{c}, \mathbf{c}_2)}. \tag{4.6}$$

Note that $P(c' \mid c, \mathbf{c}_2) = P(c' \mid \bar{c}, \mathbf{c}_2) = P(c')$, as cause variables are independent. This leads to the following expression, whose sign equals that of Formula (4.6):

$$P(\hat{e} \mid \bar{c}, \mathbf{c}_2)P(\hat{e} \mid c, c', \mathbf{c}_2) - P(\hat{e} \mid c, \mathbf{c}_2)P(\hat{e} \mid \bar{c}, c', \mathbf{c}_2).$$

Rewriting $P(\hat{e} \mid \bar{c}, \mathbf{c}_2)$ as $P(\hat{e} \mid \bar{c}, c', \mathbf{c}_2)P(c') + P(\hat{e} \mid \bar{c}, \bar{c}', \mathbf{c}_2)P(\bar{c}')$ and $P(\hat{e} \mid c, \mathbf{c}_2)$ as $P(\hat{e} \mid c, c', \mathbf{c}_2)P(c') + P(\hat{e} \mid c, \bar{c}', \mathbf{c}_2)P(\bar{c}')$, we obtain:

$$P(\hat{e} \mid c, c', \mathbf{c}_2)P(\hat{e} \mid \bar{c}, \bar{c}', \mathbf{c}_2) - P(\hat{e} \mid \bar{c}, c', \mathbf{c}_2)P(\hat{e} \mid c, \bar{c}', \mathbf{c}_2)$$

which is the definition of the product synergy, specialized to causal independence models. Hence, for causal independence models over binary variables the product synergy and intercausal influences are equivalent.

So far, we have assumed that the parameters $P(m_C \mid c)$ are known when qualitative properties are computed. However, the goal of this chapter is to qualitatively characterize causal independence models with varying interaction functions. Therefore, we abstract away from the parameters and derive the qualitative properties solely by taking into account the properties of a causal independence model's interaction function. In the next section, we infer some general properties of causal independence models.

4.2 Properties of causal independence models

In this section, we will investigate general properties of the probability distribution $P[f]$, where it is assumed that f is a Boolean function. We will make use of the analogy between Boolean algebra and ordinary arithmetic by interpreting $\bot$ as 0 and $\top$ as 1 in an arithmetic context (Birkhoff and Mac Lane, 1997), in order to allow for a compact notation.

4.2.1 General properties

Lemma 4.2 states that $P[f]$ is bounded by $f = \bot$ and $f = \top$, which is a basic result due to the first axiom of probability theory.

Lemma 4.2. $0 = P[\bot] \leq P[f] \leq P[\top] = 1$.

The probability $P[\neg f]$ is determined through the following lemma.

Lemma 4.3. $P[\neg f] = 1 - P[f]$.

Proof.

$$P[\neg f](e \mid \mathbf{c}) = \sum_{\mathbf{m}} (1 - f(\mathbf{m}))P(\mathbf{m} \mid \mathbf{c}) = \sum_{\mathbf{m}} P(\mathbf{m} \mid \mathbf{c}) - \sum_{\mathbf{m}} f(\mathbf{m})P(\mathbf{m} \mid \mathbf{c})$$

which is equivalent to $1 - P[f](e \mid \mathbf{c})$. □

Sometimes, we will add two Boolean functions or compute the difference between two Boolean functions within a CI model. In that case, Lemma 4.2 does not hold and the expression is not a proper probability distribution anymore, but *can* be interpreted as a conditional expectation.

Lemma 4.4. *$P[af + bf'] = aP[f] + bP[f']$ for constants a and b.*

Proof. This follows from the linearity property of conditional expectation. □

$P[f]$ can be bounded from below and above through the following inequalities.

Corollary 1. *$P[f \wedge f'] \leq P[f] \leq P[f \vee f'] \leq P[f] + P[f']$.*

Proof.

$$\begin{array}{rcl} P[f \wedge f'](e \mid \mathbf{c}) & = & \sum_{\mathbf{m}} f(\mathbf{m}) f'(\mathbf{m}) P(\mathbf{m} \mid \mathbf{c}) \\ & \leq & \sum_{\mathbf{m}} f(\mathbf{m}) P(\mathbf{m} \mid \mathbf{c}) \\ & = & P[f](e \mid \mathbf{c}) \\ & \leq & \sum_{\mathbf{m}} (f(\mathbf{m}) + f'(\mathbf{m}) - f(\mathbf{m}) f'(\mathbf{m})) P(\mathbf{m} \mid \mathbf{c}) \\ & = & P[f \vee f'](e \mid \mathbf{c}) \\ & \leq & \sum_{\mathbf{m}} (f(\mathbf{m}) + f'(\mathbf{m})) P(\mathbf{m} \mid \mathbf{c}) \\ & = & P[f](e \mid \mathbf{c}) + P[f'](e \mid \mathbf{c}) \end{array}$$

which completes the proof. □

4.2.2 Analytical tools

Next, we introduce a number of analytical tools that will be used in the subsequent sections.

Definition 4.6. *Let $f : \mathbb{B}^n \rightarrow \mathbb{B}$ be a Boolean function. Then, the* curry *of f, denoted by $f_{X_j = \hat{x}_j}$, is defined as the function $f_{X_j = \hat{x}_j} : \mathbb{B}^{n-1} \rightarrow \mathbb{B}$, such that*

$$f_{X_j = \hat{x}_j}(\hat{x}_1, \ldots, \hat{x}_{j-1}, \hat{x}_{j+1}, \ldots, \hat{x}_n) = f(\hat{x}_1, \ldots, \hat{x}_{j-1}, \hat{x}_j, \hat{x}_{j+1}, \ldots, \hat{x}_n).$$

Central to the analysis is the notion of a partial order $\leq$ on configurations of $\mathbf{C}$ and $\mathbf{M}$.

Definition 4.7. *Let $\mathbf{m} = (\hat{m}_1, \ldots, \hat{m}_n), \mathbf{c} = (\hat{c}_1, \ldots, \hat{c}_n) \in \mathbb{B}^n$ be Boolean n-tuples. It holds that $\mathbf{m} \leq \mathbf{c}$ iff $\hat{m}_i \leq \hat{c}_i$ for all i, $1 \leq i \leq n$, where $\bot < \top$. The relation $\mathbf{m} < \mathbf{c}$ holds iff $\mathbf{m} \leq \mathbf{c}$ and $\mathbf{m} \neq \mathbf{c}$ and the relation $>$ is defined analogously.*

Note that for any two tuples $\mathbf{m}$ and $\mathbf{c}$ it holds that either $\mathbf{m} < \mathbf{c}$, $\mathbf{m} > \mathbf{c}$, $\mathbf{m} = \mathbf{c}$ or $\exists_C : \hat{m}_C < \hat{c} \wedge \exists_{C'} : \hat{m}_{C'} > \hat{c}'$. If the latter holds then we say that $\mathbf{m}$ and $\mathbf{c}$ are *incomparable*. By means of this ordering we are in a position to compare configurations $\mathbf{m}$ of the intermediate variables $\mathbf{M}$ with configurations $\mathbf{c}$ of the cause

variables $\mathbf{C}$. In other words, we can compare intermediate states with causal states. We have chosen for this partial order instead of for a lexicographic order, as the order of the cause and intermediate variables is not always important. By means of the partial order we can prove the following lemmas.

Lemma 4.5. $\mathbf{m} = \mathbf{c} \Rightarrow P(\mathbf{m} \mid \mathbf{c}) > 0$.

Proof. If $\mathbf{m} = \mathbf{c}$, then

$$P(\mathbf{m} \mid \mathbf{c}) \quad = \quad \prod_{C \in \mathbf{C}} P(m_C \mid c)^{\hat{c}} P(\bar{m}_C \mid \bar{c})^{1-\hat{c}} = \prod_{C \in \mathbf{C}} P(m_C \mid c)^{\hat{c}} > 0$$

due to the assumptions that $P(m_C \mid c) > 0$ and $P(m_C \mid \bar{c}) = 0$. □

Lemma 4.5 states that the probability that an intermediate state is equal to the causal state is always larger than zero. Hence, the causal state always conveys information about the actual state of the intermediate variables.

Lemma 4.6. $P(\mathbf{m} \mid \mathbf{c}) > 0 \Rightarrow \mathbf{c} \geq \mathbf{m}$.

Proof. If $\mathbf{c} \not\geq \mathbf{m}$ then there is some cause variable $C = \bot$ and intermediate variable $M_C = \top$. Since $P(m_C \mid \bar{c}) = 0$ it holds that $P(\mathbf{m}_C \mid \mathbf{c}) = 0$. □

Lemma 4.6 follows from the notion of accountability and states that the truth of an intermediate variable always implies the truth of its associated cause variable. It is an important lemma, as it essentially shows that we can ignore all configurations $\mathbf{m}$ that are not smaller than or equal, or incomparable, to a given configuration $\mathbf{c}$.

The following lemmas demonstrate how a choice of the parameters influences the value of $P(\mathbf{m} \mid \mathbf{c})$.

Lemma 4.7. $\forall_C P(m_C \mid c) = 1 \Rightarrow \forall_{\mathbf{m} \neq \mathbf{c}} P(\mathbf{m} \mid \mathbf{c}) = 0$ *for arbitrary* $\mathbf{c}$.

Proof. Choose $P(m_C \mid c) = 1$ for each $C \in \mathbf{C}$. If $\mathbf{m} = \mathbf{c}$, then $P(\mathbf{m} \mid \mathbf{c}) = 1$, and necessarily $P(\mathbf{m} \mid \mathbf{c}) = 0$ for $\mathbf{m} \neq \mathbf{c}$. □

Lemma 4.7 states that if the causal relationship between the causes C the intermediates M_C is deterministic, it is not allowed that the values of causes and intermediate variables differ, which is as expected.

Lemma 4.8. $\forall_C P(m_C \mid c) < 1 \Rightarrow \forall_{\mathbf{m} \leq \mathbf{c}} P(\mathbf{m} \mid \mathbf{c}) > 0$ *for arbitrary* $\mathbf{c}$.

Proof. Since $\mathbf{m} \leq \mathbf{c}$ we have that for each cause variable C such that $M_C = \top$ also $C = \top$ and for each C such that $M_C = \bot$ it is the case that either $C = \bot$ or $C = \top$. Therefore, we may write

$$P(\mathbf{m} \mid \mathbf{c}) = \prod_{C \in \mathbf{C}} P(m_C \mid c)^{\hat{m}_C} P(\bar{m}_C \mid c)^{(1-\hat{m}_C)\hat{c}_C}$$

since $P(\bar{m}_C \mid \bar{c}) = 0$ by assumption. Since it is also assumed that $0 < P(m_C \mid c) < 1$, we have $P(m_C \mid c) > 0$ and $P(\bar{m}_C \mid c) > 0$, which proves the proposition. □

Lemma 4.8 states that if there is an uncertain causal relationship between every cause C and its associated intermediate variable M_C, then it follows that each intermediate state whose true variables form a subset of the true cause variables, has a non-zero probability of occurring.

As the qualitative behavior of a CI model is completely determined by its interaction function, in the following we will frequently investigate how these functions behave. This analysis will frequently go beyond pure Boolean functions, as some of the interaction patterns are the result of adding and subtracting Boolean functions. Considerable insight into the interaction patterns is obtained by looking at the function values (positive, negative or zero) of the resulting function for configurations smaller than a given configuration. For this, introduction of a special notation will be convenient, as given by the following definition.

Definition 4.8. *Let $q : \mathbb{B}^m \to W$ be a function, where $W = \{-b, \ldots, 0, \ldots, b\} \subset \mathbb{Z}$, then q is said to have* initial non-negative function values*, denoted by V_q^+, if*

$$\exists_{\mathbf{m}} \left[[q(\mathbf{m}) \in \{1, \ldots, b\}] \wedge \forall_{\mathbf{m}' < \mathbf{m}} [q(\mathbf{m}') \in \{0, \ldots, b\}] \right].$$

Similarly, q is said to have initial non-positive function values*, denoted by V_q^-, if V_{-q}^+ holds.*

Thus, V_q^+ means that the function value of q is *positive* for some value $\mathbf{m}$, and takes non-negative values for any value $\mathbf{m}'$ lower in the ordering $<$. The meaning of V_q^- is analogous.

As an example, consider a function q that indicates quality of life, where the variables 'happiness' and 'beauty', abbreviated to H and B, are used as summary variables. It is defined as follows. With $q(h, b) = 1$ is indicated maximal quality of life; for all $(\hat{h}, \hat{b}) < (h, b)$, for example $(\bar{h}, b) < (h, b)$, unsatisfactory quality of life is quantified by $q(\hat{h}, \hat{b}) = 0$. Thus, for this quality of life function V_q^+ holds whereas V_q^- does not. The properties V_q^+ and V_q^- of a function q will be important tools for the qualitative analysis of CI models.

4.3 Qualitative properties of CI models

In this section, it is assumed that a Boolean interaction function underlying a causal independence model is given; we then identify the signs of qualitative influences (Section 4.3.1), additive synergies (Section 4.3.2) and product synergies (Section 4.3.3). These results can be used to identify Boolean functions that respect a particular qualitative characterization. Note that we can assume that the causes are direct parents of E as the intermediate variables are marginalized out of the final computation of $P[f](e \mid \mathbf{c})$ (cf. Eq. (4.2)). For our analysis, we assume some fixed CI model over a set $\mathbf{C}$ of n cause variables, in which we focus on the interaction between different cause variables C and C' and the effect variable E, where we abbreviate M_C

by M and $M_{C'}$ by M'. Throughout this chapter we will use $\mathbf{M}_1$ to denote $\mathbf{M} \setminus \{M\}$ and $\mathbf{M}_2$ to denote $\mathbf{M} \setminus \{M, M'\}$. Likewise, we will use $\mathbf{C}_1$ to denote $\mathbf{C} \setminus \{C\}$ and $\mathbf{C}_2$ to denote $\mathbf{C} \setminus \{C, C'\}$.

4.3.1 Qualitative Influences

Let $\delta_{C \to E}[f]$ denote $\delta_{C \to E}$ where f is the interaction function of the corresponding CI model. A qualitative influence $\delta_{C \to E}[f]$ between a cause C and effect E denotes how the observation of C influences the observation of the effect e. The sign of a qualitative influence for a CI model mediated by a function f is then determined by the sign of

$$\delta_{C \to E}[f](\mathbf{c}_1) = P[f](e \mid c, \mathbf{c}_1) - P[f](e \mid \bar{c}, \mathbf{c}_1). \tag{4.7}$$

The analysis of qualitative influences requires that we isolate the contribution of particular cause variables C with respect to the effect E. By writing

$$\begin{aligned} P[f](e \mid \hat{c}, \mathbf{c}_1) &= \sum_{\mathbf{m}} f(\mathbf{m}) P(\mathbf{m} \mid \mathbf{c}) \\ &= P(m \mid \hat{c}) P[f_m](e \mid \mathbf{c}_1) + (1 - P(m \mid \hat{c})) P[f_{\bar{m}}](e \mid \mathbf{c}_1) \\ &= P[f_{\bar{m}}](e \mid \mathbf{c}_1) + P(m \mid \hat{c}) P[g](e \mid \mathbf{c}_1) \end{aligned} \tag{4.8}$$

where g denotes the *difference function* $f_m - f_{\bar{m}}$, we obtain this isolation of C from the remainder of the cause variables. Sometimes, we wish to refer to the variable M over which we vary the interaction function f, and then the notation g_M is used. Note that it holds for the difference function that $g(\mathbf{m}_1) \in \{-1, 0, 1\}$. If we substitute Eq. (4.8) into (4.7) we obtain the following equation for the sign of a qualitative influence in CI models:

$$\delta_{C \to E}[f](\mathbf{c}_1) = (P(m \mid c) - P(m \mid \bar{c})) \cdot P[g](e \mid \mathbf{c}_1).$$

Under the assumption that $P(m \mid c) > P(m \mid \bar{c})$, which always holds under the assumption of accountability, i.e., $P(m \mid \bar{c}) = 0$ (cf. Section 4.1.1), we may write

$$\delta_{C \to E}[f](\mathbf{c}_1) \propto P[g](e \mid \mathbf{c}_1). \tag{4.9}$$

We use Def .4.7 and its associated lemmas to derive some properties of qualitative influences in causal independence models. We can write

$$P[g](e \mid \mathbf{c}_1) = \sum_{\mathbf{m}_1} g(\mathbf{m}_1) P(\mathbf{m}_1 \mid \mathbf{c}_1),$$

where the configuration $\mathbf{m}_1$ ranges over all elements of $\mathbb{B}^{n-1}$. Let these configurations $\mathbf{m}_1$ be represented by $\mathbf{m}_1^i$, for $i = 1, \ldots, 2^{n-1}$, and ordered such that if $\mathbf{m}_1^i < \mathbf{m}_1^j$ then $i < j$. The configurations $\mathbf{c}_1$ of $\mathbf{C}_1$ may also be any element of $\mathbb{B}^{n-1}$

and we assume that they are ordered likewise such that $\mathbf{c}_1^i = \mathbf{m}_1^i$ for $i = 1, \ldots, 2^{n-1}$. From Lemma 4.6 it follows that for each configuration $\mathbf{c}_1$:

$$P[g](e \mid \mathbf{c}_1) \quad = \quad \sum_{\mathbf{m}_1 \leq \mathbf{c}_1} g(\mathbf{m}_1) P(\mathbf{m}_1 \mid \mathbf{c}_1). \tag{4.10}$$

Therefore, we need only take into account intermediate states that precede a causal state in the ordering. Based on this ordering we derive the properties of qualitative influences in causal independence models. We will state these properties compactly in terms of the difference function g.

Proposition 4.1. $\delta_{C \to E}[f] = 0 \Leftrightarrow g = 0$.

Proof. Using Eq. (4.10), we prove by induction that if $P[g](e \mid \mathbf{c}_1^k) = 0$ then $g(\mathbf{m}_1^k) = 0$, for $k = 1, \ldots, 2^{n-1}$.

Basis. Let $k = 1$. Then $P[g](e \mid \mathbf{c}_1^k) = g(\mathbf{m}_1^k) \cdot P(\mathbf{m}_1^k \mid \mathbf{c}_1^k)$. Since $P(\mathbf{m}_1^1 \mid \mathbf{c}_1^1) > 0$ by Lemma 4.5, it must be the case that $g(\mathbf{m}_1^1) = 0$ if $P[g](e \mid \mathbf{c}_1^1) = 0$.

Inductive hypothesis. For $i = 1, \ldots, k$, it holds that from $P[g](e \mid \mathbf{c}_1^i) = 0$ it follows that $g(\mathbf{m}_1^i) = 0$, and vice versa.

Induction step. From the inductive hypothesis, it follows that:

$$P[g](e \mid \mathbf{c}_1^{k+1}) = \sum_{1 \leq i \leq k+1} g(\mathbf{m}_1^i) P(\mathbf{m}_1^i \mid \mathbf{c}_1^{k+1}) = g(\mathbf{m}_1^{k+1}) P(\mathbf{m}_1^{k+1} \mid \mathbf{c}_1^{k+1}).$$

As $P(\mathbf{m}_1^{k+1} \mid \mathbf{c}_1^{k+1}) > 0$ it follows that $g(\mathbf{m}_1^{k+1}) = 0$ if $P[g](e \mid \mathbf{c}_1^{k+1}) = 0$, and vice versa. But then $g(\mathbf{m}_1^i) = 0$, for $i = 1, \ldots, 2^{n-1}$. □

In order to distinguish the different signs of qualitative influences it is necessary to know when positive and negative contributions are possible in principle. We first state an elementary relationship between positive and negative contributions to the sign of a qualitative influence.

Lemma 4.9. $\delta_{C \to E}[f](\mathbf{c}_1) > 0 \Leftrightarrow \delta_{C \to E}[\neg f](\mathbf{c}_1) < 0$.

Proof. Using the result of Lemma 4.3, we derive

$$\begin{aligned}
\delta_{C \to E}[f](\mathbf{c}_1) > 0 \quad &\Leftrightarrow \quad P[f_m](e \mid \mathbf{c}_1) - P[f_{\bar{m}}](e \mid \mathbf{c}_1) > 0 \\
&\Leftrightarrow \quad (1 - P[f_m](e \mid \mathbf{c}_1)) - (1 - P[f_{\bar{m}}](e \mid \mathbf{c}_1)) < 0 \\
&\Leftrightarrow \quad P[\neg f_m](e \mid \mathbf{c}_1) - P[\neg f_{\bar{m}}](e \mid \mathbf{c}_1) < 0 \\
&\Leftrightarrow \quad \delta_{C \to E}[\neg f](\mathbf{c}_1) < 0
\end{aligned}$$

which completes the proof. □

Exploring the initial function values of the difference function g, as defined above in Def. 4.8, yields further insight into the properties of qualitative influences. Note that we use the definition here by taking $b = 1$.

Lemma 4.10 lists a sufficient condition for observing a positive value of $\delta_{C\rightarrow E}[f](\mathbf{c}_1)$.

Lemma 4.10. *For every CI model with interaction function f it holds that*

$$V_g^+ \Rightarrow \exists_{\mathbf{c}_1} : \delta_{C\rightarrow E}[f](\mathbf{c}_1) > 0$$

Proof. Recall from Def. 4.8 that it holds that

$$V_g^+ = \exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1 \wedge \forall_{\mathbf{m}_1' < \mathbf{m}_1} g(\mathbf{m}_1') \in \{0, 1\}.$$

Choosing $\mathbf{c}_1 = \mathbf{m}_1$ we obtain $P[g](e \mid \mathbf{c}_1) = \sum_{\mathbf{m}_1' \leq \mathbf{c}_1} g(\mathbf{m}_1')P(\mathbf{m}_1' \mid \mathbf{c}_1)$ according to Eq. (4.10). Since for each $\mathbf{m}_1' < \mathbf{c}_1$ it holds that $g(\mathbf{m}_1') \in \{0, 1\}$ and $g(\mathbf{m}_1) = 1$ with $P(\mathbf{m}_1 \mid \mathbf{c}_1) > 0$ we have proved the lemma. □

We present a similar result for negative values of $\delta_{C\rightarrow E}[f](\mathbf{c}_1)$.

Lemma 4.11. *For every CI model with interaction function f it holds that*

$$V_g^- \Rightarrow \exists_{\mathbf{c}_1} : \delta_{C\rightarrow E}[f](\mathbf{c}_1) < 0.$$

Proof. Recall that $V_g^- = \exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1 \wedge \forall_{\mathbf{m}_1' < \mathbf{m}_1} g(\mathbf{m}_1') \in \{-1, 0\}$. If we use $\neg f$ in Lemma 4.10 and the correspondence $\neg f_m(\mathbf{m}_1) - \neg f_{\bar{m}}(\mathbf{m}_1) = 1 \Leftrightarrow g(\mathbf{m}_1) = -1$ then we obtain

$$\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1 \wedge \forall_{\mathbf{m}_1' < \mathbf{m}_1} g(\mathbf{m}_1') \in \{-1, 0\} \Rightarrow \exists_{\mathbf{c}_1} : \delta_{C\rightarrow E}[\neg f](\mathbf{c}_1) > 0.$$

From Lemma 4.9 it follows that $\delta_{C\rightarrow E}[\neg f](\mathbf{c}_1) > 0 \Leftrightarrow \delta_{C\rightarrow E}[\neg\neg f](\mathbf{c}_1) < 0 = \delta_{C\rightarrow E}[f](\mathbf{c}_1) < 0$, which proves the proposition. □

The reason why we can find a positive (or negative) value of $\delta_{C\rightarrow E}[f](\mathbf{c}_1)$ follows from the fact that we may choose a configuration $\mathbf{c}_1$ that renders all configurations $\mathbf{m}_1$ that are larger than or incomparable with $\mathbf{c}_1$ irrelevant. This is visualized in Figure 4.4.

If we consider the functions f_m and $f_{\bar{m}}$ then one of four different situations may arise. First, if neither V_g^+ nor V_g^- hold then the inductive argument of Lemma 4.1 holds and $\delta_{C\rightarrow E}[f] = 0$. Second, if both V_g^+ and V_g^- hold, then we have two incomparable configurations $\mathbf{m}_1$ and $\mathbf{m}_1'$ that render $\delta_{C\rightarrow E}[f](\mathbf{c})$ positive and negative, respectively. This leads to the following proposition.

Proposition 4.2. $V_g^+ \wedge V_g^- \Rightarrow \delta_{C\rightarrow E}[f] = \sim$.

Proof. This follows from Lemma 4.10 and Lemma 4.11. □

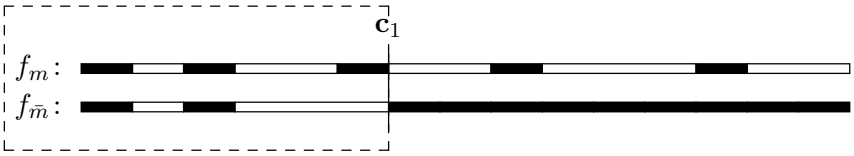


Figure 4.4: The horizontal bars represent the outcome for f_m and $f_{\bar{m}}$ respectively for those configurations from $\mathbf{m}_1^1$ to $\mathbf{m}_1^{2^{n-1}}$ of $\mathbf{M}_1$ that are comparable to $\mathbf{c}_1$. A black bar denotes that the output is true whereas a white bar denotes that the output is false. The vertical line represents a configuration $\mathbf{c}_1$ of $\mathbf{C}_1$. Due to a choice for $\mathbf{c}_1$ the only *reachable* configurations are contained within the dashed region, which must lead to a positive sign of $f_m - f_{\bar{m}}$.

Third, if V_g^+ holds and V_g^- does not hold then there is a positive value of $\delta_{C\to E}[f](\mathbf{c}_1)$ for some configuration $\mathbf{c}_1$ of $\mathbf{C}_1$ such that $\delta_{C\to E}[f]$ is either $+$ or $\sim$. Under a specific condition we can infer that the sign must be positive.

Proposition 4.3. *If* V_g^+ *and* $\neg\exists_{\mathbf{m}_1} : g(m_1) = -1$ *then* $\delta_{C\to E}[f] = +$.

Proof. The proposition follows from the observations that $\delta_{C\to E}[f](\mathbf{c}) > 0$ for some $\mathbf{c}$ and no negative contribution to the sign of the qualitative influence. □

Ref. (Lucas, 2005) includes tables for Boolean functions defined in terms of the 16 binary Boolean functions. We use these results in the following example.

Example 4.4. For both the AND and the OR operator, we have $\delta_{C\to E}[f] = +$ since for both operators it holds that the difference function $g = f_m - f_{\bar{m}}$ is non-negative and positive for at least one $\mathbf{m}_1$, which implies that the conditions of Proposition 4.3 hold.

If the conditions of Proposition 4.3 do not hold then we know for a fact that the sign is ambiguous, since it can be either non-monotonic or positive if the parameters are unknown.

Proposition 4.4. *If* V_g^+ *and* $\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1$ *then* $\delta_{C\to E}[f] = ?$.

In order to prove Proposition 4.4, we need to prove that if V_g^+ and $\exists_{\mathbf{m}_1} : f_m(\mathbf{m}_1) < f_{\bar{m}}(\mathbf{m}_1)$ then we can find parameters such that $\delta_{C\to E}[f] = \sim$ and other parameters such that $\delta_{C\to E}[f] = +$. The non-monotonic case is easily proven by the following lemma.

Lemma 4.12. *If* $\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1 \wedge \exists_{\mathbf{m}_1'} : g(\mathbf{m}_1') = -1$ *then we can choose parameters such* $\delta_{C\to E}[f] = \sim$.

Proof. From Lemma 4.7 it follows that we can choose parameters such that $\delta_{C\to E}[f](\mathbf{c}_1) = g(\mathbf{m}_1) = 1$ and $\delta_{C\to E}[f](\mathbf{c}_1') = g(\mathbf{m}_1') = -1$. □

It is more complex to prove that we can also find parameters such that $\delta_{C\to E}[f] = +$. The proof relies on the fact that we can always find parameters such that the negative contribution remains smaller than the positive contribution to the sign of the qualitative influence.

Lemma 4.13. *If V_g^+ and $\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1$ then we can find parameters such that $\delta_{C\to E}[f] = +$.*

Proof. It suffices to prove that $\forall_{\mathbf{c}}\delta_{C\to E}[f](\mathbf{c}) \geq 0$ for some choice of the parameters. We know that there must be some configuration $\mathbf{m}_1$ with $g(\mathbf{m}_1) = 1$ and for all configurations $\mathbf{m}_1'' < \mathbf{m}_1$ it holds that $g(\mathbf{m}_1'') \in \{0, 1\}$. We assume that $\forall_{\mathbf{m}_1''<\mathbf{m}_1} g(\mathbf{m}_1'') = 0$ and $\forall_{\mathbf{m}_1'>\mathbf{m}_1} g(\mathbf{m}_1') = -1$, which minimizes $P[g](e \mid \mathbf{c}_1)$. The incomparable configurations must be either zero or positive (otherwise a non-monotonic qualitative influence is implied) such that these cannot contribute negatively. We therefore obtain:

$$P[g](e \mid \mathbf{c}_1) \geq \prod_{C\in\mathbf{C}_1} P(m_C \mid \hat{c})^{\hat{m}_C} P(\bar{m}_C \mid \hat{c})^{1-\hat{m}_C} - \sum_{\mathbf{m}_1'>\mathbf{m}_1} P(\mathbf{m}_1' \mid \mathbf{c}_1)\,. \quad (4.11)$$

By choosing $P(m_C \mid c) = 1$ for each C such that $M_C = \top$, we obtain

$$\begin{aligned} P[g](e \mid \mathbf{c}_1) \quad \geq \quad & \prod_{C\in\mathbf{C}_1} P(\bar{m}_C \mid \hat{c})^{1-\hat{m}_C} - \\ & \sum_{\mathbf{m}_1'>\mathbf{m}_1} \prod_{C\in\mathbf{C}_1} P(m_C \mid \hat{c})^{(1-\hat{m}_C)\hat{m}_C'} P(\bar{m}_C \mid \hat{c})^{(1-\hat{m}_C)(1-\hat{m}_C')} \end{aligned}$$

due to the fact that if $M_C = \bot$ then $M_C' = \top$ or $M_C' = \bot$. Given that for each $\mathbf{m}_1'$ there must exist at least one cause $C_{u(\mathbf{m}_1')}$ with $u\colon \mathbb{B}^{n-1} \to \{1, \ldots, n\}$, such that $M'_{u(\mathbf{m}_1')} = \top$ and $M_{u(\mathbf{m}_1')} = \bot$, we obtain

$$P[g](e \mid \mathbf{c}_1) \geq \prod_{C\in\mathbf{C}_1} P(\bar{m}_C \mid \hat{c})^{1-\hat{m}_C} - \sum_{\mathbf{m}_1'>\mathbf{m}_1} P(m_{u(\mathbf{m}_1')} \mid \hat{c}_{u(\mathbf{m}')})\,.$$

By distinguishing present and absent causes, we may write

$$P[g](e \mid \mathbf{c}_1) \geq \prod_{C\in\mathbf{C}_1} P(\bar{m}_C \mid c)^{(1-\hat{m}_C)\hat{c}} - \sum_{\mathbf{m}_1'>\mathbf{m}_1} P(m_{u(\mathbf{m}_1')} \mid c_{u(\mathbf{m}_1')})^{\hat{c}_{u(\mathbf{m}_1')}} \cdot 0^{1-\hat{c}_{u(\mathbf{m}_1')}}\,.$$

A key step is to distinguish $\mathbf{C}_1$ into $\mathbf{C}_a = \left\{C \mid C \in \mathbf{C}_1, \forall_{\mathbf{m}_1'>\mathbf{m}_1} C \neq C_{u(\mathbf{m}_1')}\right\}$ and $\mathbf{C}_b = \left\{C \mid C \in \mathbf{C}_1, \exists_{\mathbf{m}_1'>\mathbf{m}_1} : C = C_{u(\mathbf{m}_1')}\right\}$, such that

$$\begin{aligned} P[g](e \mid \mathbf{c}_1) \quad \geq \quad & \prod_{C\in\mathbf{C}_a} P(\bar{m}_C \mid c)^{(1-\hat{m}_C)\hat{c}} \prod_{C'\in\mathbf{C}_b} P(\bar{m}_{C'} \mid c')^{(1-\hat{m}_{C'})\hat{c}'} - \\ & \sum_{\mathbf{m}_1'>\mathbf{m}_1} P(m_{u(\mathbf{m}_1')} \mid c_{u(\mathbf{m}_1')})^{\hat{c}_{u(\mathbf{m}_1')}} \cdot 0^{1-\hat{c}_{u(\mathbf{m}_1')}}\,. \end{aligned}$$

By choosing parameters $P(\bar{m}_C \mid c) = q$ for each $C \in \mathbf{C}_a$ such that $M_C = \bot$ and $P(m_{u(\mathbf{m}'_1)} \mid c_{u(\mathbf{m}'_1)}) = p$ for all $\mathbf{m}'_1 > \mathbf{m}_1$, we obtain

$$P[g](e \mid \mathbf{c}_1) \geq \prod_{C \in \mathbf{C}_a} q^{(1-\hat{m}_C)\hat{c}} \prod_{C' \in \mathbf{C}_b} (1-p)^{(1-\hat{m}_{C'})\hat{c}'} - \sum_{\mathbf{m}'_1 > \mathbf{m}_1} p^{\hat{c}_{u(\mathbf{m}'_1)}} \cdot 0^{1-\hat{c}_{u(\mathbf{m}'_1)}}.$$

Let w be the cardinality of $\{\mathbf{m}'_1 \mid \mathbf{m}'_1 \in \mathbb{B}^{n-1}, \mathbf{m}'_1 > \mathbf{m}_1\}$. As there are at most $n-1$ cause variables in $\mathbf{C}_1$, we obtain:

$$P[g](e \mid \mathbf{c}_1) \geq q^n(1-p)^n - wp$$

where w is the cardinality of $\{\mathbf{m}'_1 \mid \mathbf{m}'_1 \in \mathbb{B}^{n-1}, \mathbf{m}'_1 > \mathbf{m}_1\}$. It follows from Bernoulli's inequality that $P[g](e \mid \mathbf{c}_1) \geq q^n(1-np) - wp$, such that by choosing $p < \frac{q^n}{q^n n + w}$, we have ensured that $P[g](e \mid \mathbf{c}_1) \geq 0$. As there must be at least one configuration of $\mathbf{C}_1$ for which $P[g](e \mid \mathbf{c}_1) \neq 0$, we have proved the proposition. □

Finally, if V_g^- holds and V_g^+ does not hold then there is a negative value of $d_{C \to E}[f](\mathbf{c}_1)$ for some configuration $\mathbf{c}_1$ of $\mathbf{C}_1$ such that $\delta_{C \to E}[f]$ is either $-$ or $\sim$. Analogous to positive qualitative influences, under a specific condition we can infer that the sign must be negative.

Proposition 4.5. *If* V_g^- *and* $\neg\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1$ *then* $\delta_{C \to E}[f] = -$.

Proof. Analogous to the proof of Proposition 4.3. □

Symmetrically to positive qualitative influences, if this condition does not hold then we know for a fact that the sign is ambiguous since it can be either non-monotonic or negative if the parameters are unknown.

Proposition 4.6. *If* V_g^- *and* $\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1$ *then* $\delta_{C \to E}[f] = ?$.

The proof that parameters can always be found to generate negative or non-monotonic qualitative influences proceeds in the same way as that for the positive qualitative influences.

In the above, we have shown how properties of the interaction function f influence the qualitative properties of causal independence models. It is straightforward to recast properties of the difference function g in terms of properties of the interaction function f due to the identity $g = f_m - f_{\bar{m}}$ as is demonstrated by means of the prognostic model.

Example 4.5. We first consider the qualitative influence of I on S. In order to identify the qualitative behavior, we need to investigate the curries f_b and $f_{\bar{b}}$. If we restrict B to $\top$ (i.e., b), we have

$$\begin{aligned} f_b &\equiv (\neg b \wedge \neg E_1 \wedge \neg E_2) \vee (E_1 \wedge \neg E_2) \vee (\neg E_1 \wedge E_2) \\ &\equiv (E_1 \wedge \neg E_2) \vee (\neg E_1 \wedge E_2) \end{aligned}$$

In a similar vein we can reduce $f_{\bar{b}}$ to $\neg\big(E_1 \wedge E_2\big)$. It follows that for g we have[2]

$$g(e_1, e_2) = f_b(e_1, e_2) - f_{\bar{b}}(e_1, e_2) = (e_1 \wedge \neg e_2) \vee (\neg e_1 \wedge e_2) - \neg(e_1 \wedge e_2)) = 0$$
$$g(e_1, \bar{e}_2) = f_b(e_1, \bar{e}_2) - f_{\bar{b}}(e_1, \bar{e}_2) = (e_1 \wedge \neg \bar{e}_2) \vee (\neg e_1 \wedge \bar{e}_2) - \neg(e_1 \wedge \bar{e}_2)) = 0$$
$$g(\bar{e}_1, e_2) = f_b(\bar{e}_1, e_2) - f_{\bar{b}}(\bar{e}_1, e_2) = (\bar{e}_1 \wedge \neg e_2) \vee (\neg \bar{e}_1 \wedge e_2) - \neg(\bar{e}_1 \wedge e_2)) = 0$$
$$g(\bar{e}_1, \bar{e}_2) = f_b(\bar{e}_1, \bar{e}_2) - f_{\bar{b}}(\bar{e}_1, \bar{e}_2) = (\bar{e}_1 \wedge \neg \bar{e}_2) \vee (\neg \bar{e}_1 \wedge \bar{e}_2) - \neg(\bar{e}_1 \wedge \bar{e}_2)) = -1$$

It follows that Proposition 4.5 holds, such that $\delta_{I \to S}[f] = -$. This negative influence of the serious illness on prognosis is in accordance with the previously stated domain knowledge. We proceed in a similar way for the qualitative influences of T_1 on S and obtain the following results. For the qualitative influence of T_1 on S we have $f_{e_1} \equiv \neg E_2$ and $f_{\bar{e}_1} \equiv (\neg B \wedge \neg E_2) \vee E_2$. It follows that for g we have that $g(i, e_2) = 0$, $g(i, \bar{e}_2) = 1$, $g(\bar{\imath}, e_2) = -1$ and $g(i, e_2) = 0$. As $(i, \bar{e}_2)$ and $(\bar{\imath}, e_2)$ are incomparable and have opposing signs, it follows that $\delta_{T_1 \to S}[f] = \sim$ according to Proposition 4.2. We remark that $\delta_{T_2 \to S}[f] = \sim$ by symmetry. The qualitative influences are depicted in Figure 4.5.

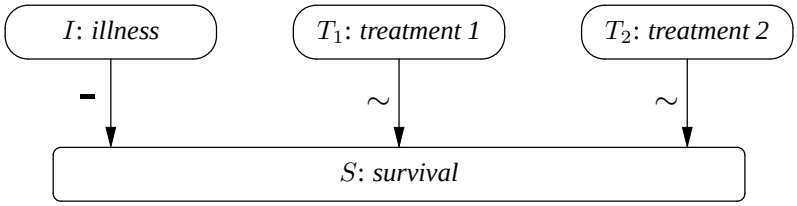


Figure 4.5: Qualitative influences with respect to patient survival.

Previously, we have shown how properties of the interaction function f influence the qualitative properties of causal independence models. Next, we show that, by means of the propositions and lemmas that have been derived, we can also immediately infer properties of interaction functions that should hold when a qualitative influence is known. First, observe that, based on the lemmas and propositions above

$$\begin{array}{llllll} \big(V_g^+ \wedge \neg\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1\big) & \vee & \big(V_g^+ \wedge \exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1\big) & \vee & \\ \big(V_g^- \wedge \neg\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1\big) & \vee & \big(V_g^- \wedge \exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1\big) & \vee & (g = 0) \end{array}$$

covers all possible cases. The first two conjunctions in this disjunction handle the positive qualitative influences (due to Proposition 4.3 and Lemma 4.13). The third and fourth conjunctions in this disjunction handle the negative qualitative influences (by symmetry), and the last conjunction is a necessary and sufficient condition for observing a zero qualitative influence (due to Proposition 4.1). The second and fourth conjunctions are conditions that may lead to non-monotonic qualitative influences, and whose disjunction is equivalent to $\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = -1 \wedge \exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1$. The properties of interaction functions given a qualitative influence are listed in Table 4.1.

[2]Recall that in an arithmetic context, we interpret $\top$ as 1 and $\bot$ as 0.

Table 4.1: Properties of interaction functions given a qualitative influence.

Qualitative Influence	Property of the Interaction Function
0	$g = 0$
$+$	V_g^+
$-$	V_g^-
$\sim$	$\exists_{\mathbf{m}_1} : g(\mathbf{m}_1) = 1 \wedge \exists_{\mathbf{m}'_1} : g(\mathbf{m}'_1) = -1$

Example 4.6. Suppose we knew the qualitative influences but not the underlying interaction function for the prognostic model of Section 4.1.1. According to Table 4.1 we have:

$$\begin{array}{lll} \delta_{I\to S}[f] = - & \Rightarrow & V_{g_B}^- \\ \delta_{T_1\to S}[f] = \sim & \Rightarrow & \exists_{\mathbf{m}_1} : g_{E_1}(\mathbf{m}_1) = 1 \wedge \exists_{\mathbf{m}'_1} : g_{E_1}(\mathbf{m}'_1) = -1 \\ \delta_{T_2\to S}[f] = \sim & \Rightarrow & \exists_{\mathbf{m}_1} : g_{E_2}(\mathbf{m}_1) = 1 \wedge \exists_{\mathbf{m}'_1} : g_{E_2}(\mathbf{m}'_1) = -1 \end{array}$$

where $g_B = f_b - f_{\bar{b}}$, $g_{E_1} = f_{e_1} - f_{\bar{e}_1}$, and $g_{E_2} = f_{e_2} - f_{\bar{e}_2}$. The results are indeed properties of the interaction function of the prognostic model, as represented by the Boolean expression (4.4). The first qualitative influence would, for example, preclude choosing the AND and OR interaction functions, as both do not satisfy property $V_{g_B}^-$.

4.3.2 Additive Synergies

Additive synergies express how two cause variables C and C' from the set of cause variables $\mathbf{C}$ jointly influence the probability of observing the effect E. Recall that the remaining cause variables are denoted by $\mathbf{C}_2 = \mathbf{C} \setminus \{C, C'\}$. Using the general definition of additive synergy from QPN theory, the additive synergy $\delta_{(C,C')\to E}[f]$ between C and C' given interaction function f is determined by

$$\begin{aligned} \delta_{(C,C')\to E}[f](\mathbf{c}_2) \;=\; & P[f](e \mid c, c', \mathbf{c}_2) + P[f](e \mid \bar{c}, \bar{c}', \mathbf{c}_2) - \\ & P[f](e \mid \bar{c}, c', \mathbf{c}_2) - P[f](e \mid c, \bar{c}', \mathbf{c}_2). \end{aligned} \tag{4.12}$$

The analysis requires an isolation of cause variables C and C'. We apply the decomposition (4.8) twice and obtain:

$$\begin{aligned} P[f](e \mid \hat{c}, \hat{c}', \mathbf{c}_2) \;=\; & P(m \mid \hat{c})P(m' \mid \hat{c}')P[h](e \mid \mathbf{c}_2) + P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) + \\ & P(m \mid \hat{c})P[f_{m,\bar{m}'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) + \\ & P(m' \mid \hat{c}')P[f_{\bar{m},m'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2), \end{aligned} \tag{4.13}$$

where the function $h : \mathbb{B}^{n-2} \to \{-2, -1, 0, 1, 2\}$ is defined as

$$h(\mathbf{m}_2) = f_{m,m'}(\mathbf{m}_2) + f_{\bar{m},\bar{m}'}(\mathbf{m}_2) - f_{\bar{m},m'}(\mathbf{m}_2) - f_{m,\bar{m}'}(\mathbf{m}_2). \tag{4.14}$$

The function h is also sometimes indicated by $h_{M,M'}$. By inserting Eq. (4.13) into (4.12) we obtain

$$\delta_{(C,C')\to E}[f](\mathbf{c}_2) = (P(m \mid c) - P(m \mid \bar{c}))\,(P(m' \mid c') - P(m' \mid \bar{c}'))P[h](e \mid \mathbf{c}_2).$$

Under the assumptions that $P(m \mid c) > P(m \mid \bar{c})$ and $P(m' \mid c') > P(m' \mid \bar{c}')$, which holds under the assumption of accountability, we may write

$$\delta_{(C,C')\to E}[f](\mathbf{c}_2) \propto P[h](e \mid \mathbf{c}_2).$$

We take a similar approach as for qualitative influences and use an ordering on configurations of $\mathbf{M}_2$ and $\mathbf{C}_2$ which now range from $\mathbf{m}_1$ to $\mathbf{m}_{2^{n-2}}$ and from $\mathbf{c}_1$ to $\mathbf{c}_{2^{n-2}}$ respectively.

The structure of the expression for qualitative influences and additive synergies is essentially the same, where the only difference is that we sum over 2^{n-2} instead of 2^{n-1} configurations and g is replaced by h. If we consider the proofs of Lemmas 4.9–4.13 and Propositions 4.1–4.6 in the previous section, then we find that none, with the exception of Lemma 4.13, are dependent upon these two differences. Due to the analogy between qualitative influences and additive synergies, we state the results in terms of the difference function h without proof.

A necessary and sufficient condition for observing a zero additive synergy is easily found.

Proposition 4.7. $\delta_{(C,C')\to E}[f] = 0 \Leftrightarrow h = 0$.

Again, interaction functions f and their negations $\neg f$ lead to opposite contributions to the qualitative sign.

Lemma 4.14. $\delta_{(C,C')\to E}(\mathbf{c}_2) > 0 \Leftrightarrow \delta_{(C,C')\to E}[\neg f](\mathbf{c}_2) < 0$.

We next investigate the implications of function values of the function h, as defined above in Eq. (4.14), using Def. 4.8, for the qualitative properties. Here we take $b = 2$. An analysis of positive and negative contributions to the sign of the additive synergy is given by Lemmas 4.15 and 4.16.

Lemma 4.15. *For every CI model with interaction function f it holds that*

$$V_h^+ \Rightarrow \exists_{\mathbf{c}_2} : \delta_{(C,C')\to E}[f](\mathbf{c}_2) > 0.$$

Lemma 4.16. *For every CI model with interaction function f it holds that*

$$V_h^- \Rightarrow \exists_{\mathbf{c}_2} : \delta_{C,C'}(\mathbf{c}_2)[f] < 0.$$

Non-monotonic additive synergies are identified by Proposition 4.8.

Proposition 4.8. $V_h^+ \wedge V_h^- \Rightarrow \delta_{(C,C')\to E}[f] = \sim$.

Positive additive synergies are identified by Proposition 4.9 and ambiguous additive synergies (either non-monotonic or positive signs) are identified by 4.10. We can always choose parameters such that this ambiguous additive synergy reduces to a non-monotonic or positive additive synergy. The proof is similar to the proof in case of qualitative influences and is omitted here.

Proposition 4.9. *If* V_h^+ *and* $\forall_{\mathbf{m}_2} h(\mathbf{m}_2) \in \{0,1,2\}$ *then* $\delta_{(C,C')\to E}[f] = +$.

Proposition 4.10. *If* V_h^+ *and* $\exists_{\mathbf{m}_2} : h(\mathbf{m}_2) \in \{-2,-1\}$ *then* $\delta_{(C,C')\to E}[f] = ?$.

Symmetric results are obtained for negative additive synergies in Proposition 4.11, where Proposition 4.12 identifies ambiguous additive synergies which can be either non-monotonic or negative, depending on the parameters.

Proposition 4.11. *If* V_h^- *and* $\forall_{\mathbf{m}_2} h(\mathbf{m}_2) \in \{-2,-1,0\}$ *then* $\delta_{(C,C')\to E}[f] = -$.

Proposition 4.12. *If* V_h^- *and* $\exists_{\mathbf{m}_2} : h(\mathbf{m}_2) \in \{1,2\}$ *then* $\delta_{(C,C')\to E}[f] = ?$.

We use the results of Ref. (Lucas, 2005) to verify some of our results.

Example 4.7. For the AND operator, we have $\delta_{(C,C')\to E}[f] = +$ since the difference function $h(\mathbf{m}_2) = f_{m,m'}(\mathbf{m}_2) + f_{\bar{m},\bar{m}'}(\mathbf{m}_2) - f_{\bar{m},m'}(\mathbf{m}_2) - f_{m,\bar{m}'}(\mathbf{m}_2)$ must be non-negative and positive for at least one configuration of $\mathbf{m}_2$. On the other hand, for the OR operator we have $\delta_{(C,C')\to E}[f] = -$ since h is non-positive and negative for at least one configuration of $\mathbf{m}_2$.

We can recast properties of the difference function h in terms of properties of the interaction function f as we have the identity $h = f_{m,m'} + f_{\bar{m},\bar{m}'} - f_{\bar{m},m'} - f_{m,\bar{m}'}$. We illustrate the results concerning additive synergies by means of the running example, shown in Figure 4.3.

Example 4.8. With regard to the additive synergy between the treatments T_1 and T_2, we have $f_{e_1,e_2} \equiv \bot$, $f_{\bar{e}_1,\bar{e}_2} \equiv \neg B$ and $f_{\bar{e}_1,e_2} \equiv f_{e_1,\bar{e}_2} \equiv \top$. We then have $h(b) = -2$ and $h(\bar{b}) = -1$ such that $\delta_{(T_1,T_2)\to S}[f] = -$ according to Proposition 4.11. This agrees with the observation that the administration of one of both treatments is optimal, whereas administration of both treatments yields a suboptimal result. With regard to the additive synergy between I and T_1, we have $f_{b,e_1} \equiv \neg E_2$, $f_{\bar{b},\bar{e}_1} \equiv \top$, $f_{\bar{b},e_1} \equiv \neg E_2$ and $f_{b,\bar{e}_1} \equiv E_2$. We then have that $h(e_2) = 0$ and $h(\bar{e}_2) = 1$ such that $\delta_{(I,T_1)\to S}[f] = +$ according to Proposition 4.9. We also have $\delta_{(I,T_2)\to S}[f] = +$ by symmetry. This is in agreement with the fact that when a treatment is administered to an ill person, or when no treatment is administered in the absence of the illness improves survival in comparison to when a non-ill person is treated or when treatment is not given to an ill person. The additive synergies are depicted in Figure 4.6.

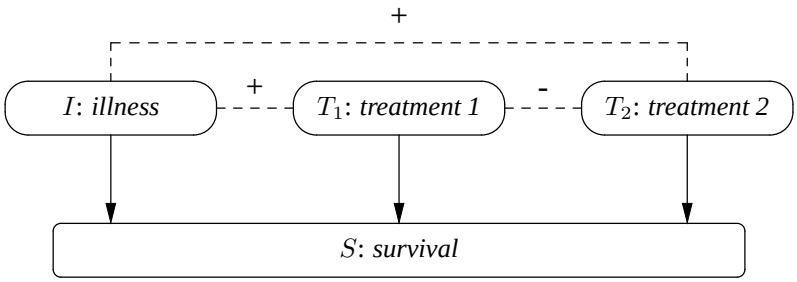


Figure 4.6: Additive synergies with respect to patient survival.

So far, we have only considered the qualitative behavior of a given interaction function. Again, we infer properties of interaction functions that should hold when an additive synergy is known. These properties are shown in Table 4.2 and have straightforward derivations due to the correspondence between qualitative influences and additive synergies. An example is again provided by considering the qualitative properties of the prognostic model.

Table 4.2: Properties of interaction functions given an additive synergy.

Additive Synergy	Property of the Interaction Function
0	$h = 0$
$+$	V_h^+
$-$	V_h^-
$\sim$	$\exists_{\mathbf{m}_2} : h(\mathbf{m}_2) \in \{1, 2\} \wedge \exists_{\mathbf{m}_2'} : h(\mathbf{m}_2') \in \{-2, -1\}$

Example 4.9. Suppose we knew the additive synergies but not the underlying interaction function for the prognostic model. According to Table 4.2 we have

$$\delta_{(T_1,T_2)\rightarrow S}[f] = - \quad \Rightarrow \quad V^-_{h_{E_1,E_2}}$$

where $h_{E_1,E_2} = f_{e_1,e_2} + f_{\bar{e}_1,\bar{e}_2} - f_{\bar{e}_1,e_2} - f_{e_1,\bar{e}_2}$. This is indeed a property of Boolean expression (4.4) that represents the prognostic model, as may be verified. This constraint would, for example, exclude the AND Boolean function, as it does not satisfy property $V^-_{h_{E_1,E_2}}$.

4.3.3 Product Synergies

Product synergies describe the created dependence between two causes when the value of the effect variable is observed. The sign $\delta^{\hat{e}}_{(C,C')\rightarrow E}[f]$ of a product synergy between C and C' with respect to $\hat{e}$ when f is the chosen interaction function, is

determined by

$$\begin{aligned}\delta^{\hat{e}}_{(C,C')\to E}[f](\mathbf{c}_2) &= P[f](\hat{e} \mid c, c', \mathbf{c}_2)P[f](\hat{e} \mid \bar{c}, \bar{c}', \mathbf{c}_2) - \\ &\quad P[f](\hat{e} \mid \bar{c}, c', \mathbf{c}_2)P[f](\hat{e} \mid c, \bar{c}', \mathbf{c}_2).\end{aligned}$$

This can be rewritten for $E = \top$ (presence of the effect has been observed) to:

$$\begin{aligned}\delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2) &= P(m \mid c)P(m' \mid c')\big(P[h](e \mid \mathbf{c}_2)P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) - \\ &\quad P[f_{m,\bar{m}'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2)P[f_{\bar{m},m'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2)\big),\end{aligned}$$

where again $h = f_{m,m'} + f_{\bar{m},\bar{m}'} - f_{\bar{m},m'} - f_{m,\bar{m}'}$. Under our standard assumption of accountability, this yields:

$$\begin{aligned}\delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2) &\propto P[h](e \mid \mathbf{c}_2)P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) - \\ &\quad P[f_{m,\bar{m}'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2)P[f_{\bar{m},m'} - f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2).\end{aligned}$$

This can be alternatively written as:

$$\begin{aligned}\delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2) &\propto P[f_{m,m'}](e \mid \mathbf{c}_2)P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) - \\ &\quad P[f_{\bar{m},m'}](e \mid \mathbf{c}_2)P[f_{m,\bar{m}'}](e \mid \mathbf{c}_2).\end{aligned}$$

Using the distributive law of arithmetic, we obtain:

$$\begin{aligned}\delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2) &\propto P[f_{m,m'}](e \mid \mathbf{c}_2)P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) - \\ &\quad P[f_{\bar{m},m'}](e \mid \mathbf{c}_2)P[f_{m,\bar{m}'}](e \mid \mathbf{c}_2) \\ &= \big(\sum_{\mathbf{m}_2} f_{m,m'}(\mathbf{m}_2)P(\mathbf{m}_2 \mid \mathbf{c}_2)\big)\big(\sum_{\mathbf{m}_2} f_{\bar{m},\bar{m}'}(\mathbf{m}_2)P(\mathbf{m}_2 \mid \mathbf{c}_2)\big) - \\ &\quad \big(\sum_{\mathbf{m}_2} f_{\bar{m},m'}(\mathbf{m}_2)P(\mathbf{m}_2 \mid \mathbf{c}_2)\big)\big(\sum_{\mathbf{m}_2} f_{m,\bar{m}'}(\mathbf{m}_2)P(\mathbf{m}_2 \mid \mathbf{c}_2)\big) \\ &= \sum_{\mathbf{m}_2,\mathbf{m}_2'} r(\mathbf{m}_2, \mathbf{m}_2')P(\mathbf{m}_2 \mid \mathbf{c}_2)P(\mathbf{m}_2' \mid \mathbf{c}_2)\end{aligned}$$

where the function $r : \mathbb{B}^{n-2} \times \mathbb{B}^{n-2} \to \{-1, 0, 1\}$ is defined as follows:

$$r(\mathbf{m}_2, \mathbf{m}_2') = f_{m,m'}(\mathbf{m}_2)f_{\bar{m},\bar{m}'}(\mathbf{m}_2') - f_{\bar{m},m'}(\mathbf{m}_2)f_{m,\bar{m}'}(\mathbf{m}_2'). \tag{4.15}$$

We will also sometimes use the notation $r_{M,M'}$. From the expression above, it follows that the behavior of the product synergy is determined by the function r.

It appears that it suffices to carry out the analysis for $E = \top$ (the effect has been observed to be present), as application of the following lemma renders the analysis of the qualitative behavior of the product synergy for $E = \bot$ (absence of the effect has been observed) a straightforward exercise.

Lemma 4.17. $\delta^{e}_{(C,C')\to E}[\neg f] = \delta^{\bar{e}}_{(C,C')\to E}[f]$.

Proof.

$$\begin{aligned}
\delta^{e}_{(C,C')\to E}[\neg f](\mathbf{c}_2) &\propto P[\neg f_{m,m'}](e \mid \mathbf{c}_2)P[\neg f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2) - \\
&\quad P[\neg f_{\bar{m},m'}](e \mid \mathbf{c}_2)P[f_{m,\bar{m}'}](e \mid \mathbf{c}_2) \\
&= (1 - P[f_{m,m'}](e \mid \mathbf{c}_2))(1 - P[f_{\bar{m},\bar{m}'}](e \mid \mathbf{c}_2)) - \\
&\quad (1 - P[f_{\bar{m},m'}](e \mid \mathbf{c}_2))(1 - P[f_{m,\bar{m}'}](e \mid \mathbf{c}_2)) \\
&= P[f_{m,m'}](\bar{e} \mid \mathbf{c}_2)P[f_{\bar{m},\bar{m}'}](\bar{e} \mid \mathbf{c}_2) - \\
&\quad P[f_{\bar{m},m'}](\bar{e} \mid \mathbf{c}_2)P[f_{m,\bar{m}'}](\bar{e} \mid \mathbf{c}_2) \\
&\propto \delta^{\bar{e}}_{(C,C')\to E}[f](\mathbf{c}_2)
\end{aligned}$$

which completes the proof. □

Hence, if $\delta^{e}_{(C,C')\to E}[\neg f](\mathbf{c}_2)$ has a particular sign for configuration $\mathbf{c}_2$ then $\delta^{\bar{e}}_{(C,C')\to E}[f](\mathbf{c}_2)$ will have the same sign. Therefore, the sign of the product synergy for $E = \top$ with interaction function $\neg f$ will be the same as that for $E = \bot$ with interaction function f. Due to this relationship between the signs of the product synergy for $E = \top$ and $E = \bot$, we will only consider the case where $E = \top$. Recall that by Lemma 4.1, we have the following interesting relationship between product synergies and additive synergies, which offers an alternative way to compute the product synergy $\delta^{\bar{e}}_{(C,C')\to E}[f](\mathbf{c}_2)$, based on the associated additive synergy $\delta_{(C,C')\to E}[f](\mathbf{c}_2)$ and the associated product synergy $\delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2)$:

$$\delta^{\bar{e}}_{(C,C')\to E}[f](\mathbf{c}_2) = \delta^{e}_{(C,C')\to E}[f](\mathbf{c}_2) - \delta_{(C,C')\to E}[f](\mathbf{c}_2).$$

Lemma 4.1 is useful for constructing tables of signs for particular Boolean functions, as it saves constructing one of these tables.

Example 4.10. Ref. (Lucas, 2005) includes tables for Boolean functions defined in terms of the 16 binary Boolean functions. Consider the AND operator, $\wedge$; its additive synergy is equal to $\delta_{(C,C')\to E}[\wedge] = +$, whereas its product synergy for $E = \top$ is equal to $\delta^{e}_{(C,C')\to E}[\wedge] = 0$. Lemma 4.1 tells us that the product synergy for $E = \bot$ is equal to $\delta^{\bar{e}}_{(C,C')\to E}[f] = -$, which is indeed the value for the product synergy for $E = \bot$ in Table 12 in (Lucas, 2005).

In the following, we derive sufficient conditions for observing particular qualitative behavior in terms of product synergies.

Proposition 4.13. $\delta^{e}_{(C,C')\to E}[f] = 0$ *if it holds that*

$$\forall_{\mathbf{m_2},\mathbf{m}'_2} \left[(f_{m,m'}(\mathbf{m}_2) \wedge f_{\bar{m},\bar{m}'}(\mathbf{m}'_2)) \Leftrightarrow (f_{\bar{m},m'}(\mathbf{m}_2) \wedge f_{m,\bar{m}'}(\mathbf{m}'_2)) \right].$$

Proof. Note that if the premise holds, then, according to Def. 4.15 of the function r, we have that $r(\mathbf{m}_2, \mathbf{m}'_2) = 0$, for each $\mathbf{m}_2, \mathbf{m}'_2$, and thus $\delta^e_{(C,C')\to E}[f] = 0$. □

A special case of this proposition, is the following condition:

$$(f_{m,m'} \equiv \bot \vee f_{\bar{m},\bar{m}'} \equiv \bot) \wedge (f_{m,\bar{m}'} \equiv \bot \vee f_{\bar{m},m'} \equiv \bot),$$

i.e., if at least one Boolean function at both sides of the negation of Def. 4.15 is equal to falsum, then a zero product synergy results.

We again determine conditions under which $\delta^e_{(C,C')\to E}[f](\mathbf{c}_2)$ is positive or negative. Similar to previous sections, we use the notations V_r^+ and V_r^-, this time in terms of the function r defined above; for example V_r^+ means that

$$\exists_{\mathbf{m}_2,\mathbf{m}'_2} \left[\left[r(\mathbf{m}_2, \mathbf{m}'_2) = 1\right] \wedge \forall \mathbf{m}''_2 < \mathbf{m}_2, \mathbf{m}'''_2 < \mathbf{m}'_2 \left[r(\mathbf{m}''_2, \mathbf{m}'''_2) \in \{0, 1\}\right]\right]$$

Lemma 4.18. *For every CI model with interaction function f we have*

$$V_r^+ \Rightarrow \exists_{\mathbf{c}_2} : \delta^e_{(C,C')\to E}[f](\mathbf{c}_2) > 0.$$

Proof. Simply note that if r is initially non-negative, we have a positive $\delta^e_{(C,C')\to E}[f](\mathbf{c}_2)$ for at least one value $\mathbf{c}_2$ by definition. □

An example of a positive value of $\delta^e_{(C,C')\to E}[f](\mathbf{c}_2)$ is demonstrated in Figure 4.7.

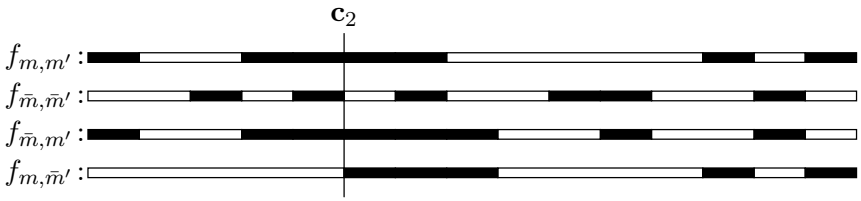


Figure 4.7: Similar to Figure 4.4, the horizontal bars represent the outcome for $f_{m,m}$, $f_{\bar{m},\bar{m}}$, $f_{\bar{m},m}$ and $f_{m,\bar{m}}$ for configurations $\mathbf{m}_2^1$ to $\mathbf{m}_2^{2^{n-2}}$ of $\mathbf{M}_2$. The vertical line represents a configuration $\mathbf{c}_2$ of $\mathbf{C}_2$. Due to a choice for $\mathbf{c}_2$ the only *reachable* configurations are contained within the dashed region, which must lead to a positive sign of $\delta^e_{(C,C')\to E}[f](\mathbf{c}_2)$.

A similar result holds for negative values of $\delta^e_{(C,C')\to}[f](\mathbf{c}_2)$.

Lemma 4.19. *For every CI model with interaction function f we have*

$$V_r^- \Rightarrow \exists_{\mathbf{c}_2} : \delta^e_{(C,C')\to E}[f](\mathbf{c}_2) < 0.$$

Proof. Analogous to the proof of Lemma 4.18. □

Proposition 4.14. *If both V_r^+ and V_r^- hold then $\delta^e_{(C,C')\to E}[f] = \sim$.*

Proof. This follows from the definition of a non-monotonic product synergy. □

It also follows directly from Lemmas 4.18 and 4.19 that if V_r^+ holds and V_r^- does not hold, then the sign of the product synergy is either positive or non-monotonic. Conversely, if V_r^- holds and V_r^+ does not hold, then it follows that the sign of the product synergy is either negative or non-monotonic. The following two propositions identify under which conditions the sign of a product synergy is known to be positive or negative, respectively.

Proposition 4.15. *If* $\exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = 1$ *and* $\forall_{\mathbf{m}_2,\mathbf{m}_2'}\,[r(\mathbf{m}_2,\mathbf{m}_2') \geq 0]$ *then it holds that* $\delta^e_{(C,C')\rightarrow E}[f] = +$.

Proof. This is just the general case of Lemma 4.18, where we ensure that the conditions listed for configurations $\mathbf{m}_2'' < \mathbf{m}_2, \mathbf{m}_2''' < \mathbf{m}_2'$ such that $r(\mathbf{m}_2'',\mathbf{m}_2''') \geq 0$ not only hold for configurations smaller than $\mathbf{m}_2, \mathbf{m}_2'$, but for all configurations $\mathbf{m}_2'' \neq \mathbf{m}_2, \mathbf{m}_2''' \neq \mathbf{m}_2'$. □

Proposition 4.16. *If* $\exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = -1$ *and* $\forall_{\mathbf{m}_2,\mathbf{m}_2'}\,[r(\mathbf{m}_2,\mathbf{m}_2') \leq 0]$ *then it holds that* $\delta^e_{(C,C')\rightarrow E}[f] = -$.

Proof. This is the generalized case of Lemma 4.19. □

The cases that are not covered by the above propositions will be categorized as ambiguous.

Proposition 4.17. *If none of Propositions 4.13–4.16 hold then* $\delta^e_{(C,C')\rightarrow E}[f] = ?$.

Proposition 4.17 collects those cases for which no sufficient conditions for observing a particular sign of a product synergy have been derived. In such cases, the sign can still be positive, negative or non-monotonic, depending on the parameters and depending on the structure of the interaction function. It is important to realize that due to Lemma 4.17, the above results equally hold for the case where $E = \bot$ whenever we replace each occurrence of f by $\neg f$.

We illustrate the results concerning product synergies again by means of the prognostic model, depicted in Figure 4.3.

Example 4.11. We first focus on the case where we hypothesize that the patient will survive, i.e., $S = \top$. With regard to the product synergy between treatments T_1 and T_2, we have that $f_{e_1,e_2} \equiv \bot$, $f_{\bar{e}_1,\bar{e}_2} \equiv \neg B'$ and $f_{\bar{e}_1,e_2} \equiv f_{e_1,\bar{e}_2} = \top$. Condition 3 of Proposition 4.16 is satisfied since $r(B,B') = -1$ for each value of B, B', and thus $\delta^s_{(T_1,T_2)\rightarrow S}[f] = -$. This agrees with the observation that we expect that one of both treatments was administered given that we observe patient survival. With regard to the product synergy between I and T_1, we have that $f_{b,e_1} \equiv \neg E_2$, $f_{\bar{b},\bar{e}_1} \equiv \top$, $f_{\bar{b},e_1} \equiv \neg E_2$ and $f_{b,\bar{e}_1} \equiv E_2'$. Condition 1 of Proposition 4.15 is satisfied since $r(\bar{e}_2,\bar{e}_2') = 1$, whereas $r(E_2,E_2') = 0$ for any value of E_2, E_2', with the exception

of $E_2 = \bot$ and $E_2' = \bot$; thus $\delta^s_{(I,T_1)\to S}[f] = +$. Hence, it is likely that treatment T_1 is administered given disease progression and patient survival and that treatment T_1 is not administered given no progression and patient survival. It is less likely that treatment T_1 is administered given no progression and patient survival and that treatment T_1 is not administered given disease progression and patient survival. The same holds for the product synergy between I and T_2 by symmetry. The results are summarized by Figure 4.8.

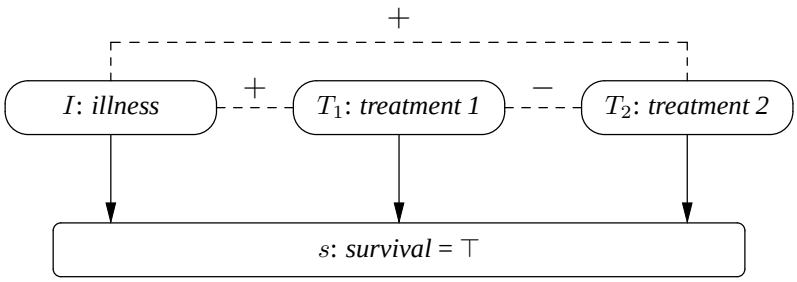


Figure 4.8: Product synergies with respect to patient survival.

As has been proved in Lemma 4.17, we can use also the derived propositions for $E = \bot$ by replacing f with $\neg f$. With regard to the product synergy between T_1 and T_2, we have that $\neg f_{e_1,e_2} \equiv \top$, $\neg f_{\bar{e}_1,\bar{e}_2} \equiv B$ and $\neg f_{\bar{e}_1,e_2} \equiv \neg f_{e_1,\bar{e}_2} = \bot$. Condition 3 of Proposition 4.15 is satisfied, since $r(B, B') = B$, thus $\delta^{\bar{s}}_{(T_1,T_2)\to S}[f] = +$. With regard to the product synergy between I and T_1, we have that $\neg f_{b,e_1} \equiv E_2$, $\neg f_{\bar{b},\bar{e}_1} \equiv \bot$, $\neg f_{\bar{b},e_1} \equiv E_2$ and $\neg f_{b,\bar{e}_1} = \neg E_2'$, thus $r(E_2, E_2') = -(E_2 \wedge \neg E_2')$. We classify the product synergy as $\delta^{\bar{s}}_{(I,T_1)\to S}[f] = -$. The same holds for the product synergy between I and T_2 by symmetry. The results are summarized by Figure 4.9.

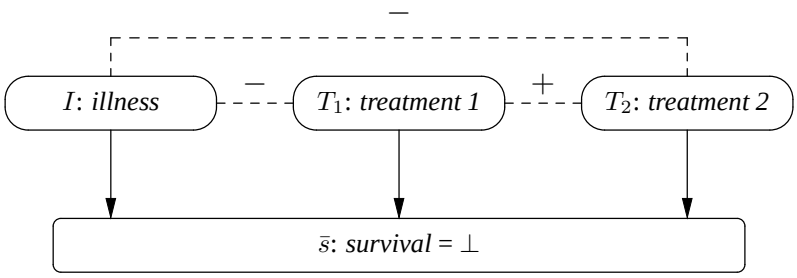


Figure 4.9: Product synergies with respect to patient death.

Again, we look at the converse analysis from qualitative specification to constraints on interaction functions using the propositions and lemmas that have been derived. Properties of product synergies with the effect observed to be present ($E = \top$) are shown in Table 4.3 and are derived by negating the properties for opposite signs when $E = \top$. For example, since V_r^+ with $E = \top$ implies that there is a con-

figuration $\mathbf{c}_2$ of cause variables such that $\delta^e_{(C,C')\rightarrow E}[f](\mathbf{c}_2) > 0$ (Lemma 4.18), we know that $\neg V_r^+$ must hold for negative product synergies with $E = \top$. Likewise, $\neg V_r^-$ must hold for positive product synergies with $E = \top$. For the same reason, $\neg V_r^+ \wedge \neg V_r^-$ must hold for zero product synergies with $E = \top$. For non-monotonic product synergies it holds that Propositions 4.15 and 4.16 must both be false. Since, according to Proposition 4.13, it cannot be the case that $\forall_{\mathbf{m}_2,\mathbf{m}_2'}\,[r(\mathbf{m}_2,\mathbf{m}_2') = 0]$, it must hold that both $\exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = 1$ and $\exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = -1$. Properties of product synergies with $E = \bot$ are obtained using Lemma 4.17 by replacing the function r with the function

$$\bar{r}(\mathbf{m}_2,\mathbf{m}_2') = \neg f_{m,m'}(\mathbf{m}_2)\neg f_{\bar{m},\bar{m}'}(\mathbf{m}_2') - \neg f_{\bar{m},m'}(\mathbf{m}_2)\neg f_{m,\bar{m}'}(\mathbf{m}_2').$$

Table 4.3: Properties of interaction functions given a product synergy for $E = \top$.

Product Synergy	Property of the Interaction Function
0	$\neg V_r^+ \wedge \neg V_r^-$
$+$	$\neg V_r^-$
$-$	$\neg V_r^+$
$\sim$	$\exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = 1 \wedge \exists_{\mathbf{m}_2,\mathbf{m}_2'} : r(\mathbf{m}_2,\mathbf{m}_2') = -1$

In order to demonstrate this converse analysis, we look at the product synergy between treatments T_1 and T_2 of the prognostic model.

Example 4.12. Suppose we knew the product synergies but not the underlying interaction function for the prognostic model. For the product synergy between treatment T_1 and T_2 with the effect assumed to be present ($E = \top$), we have

$$\delta^e_{(T_1,T_2)\rightarrow S}[f] = - \quad \Rightarrow \quad \neg V^+_{r_{E_1,E_2}}$$

whereas its product synergy for the effect assumed to be absent ($E = \bot$) is given by

$$\delta^{\bar{e}}_{(T_1,T_2)\rightarrow S}[f] = + \quad \Rightarrow \quad \neg V^-_{\bar{r}_{E_1,E_2}}$$

Note that here we use the complementary function $\bar{r}_{E_1,E_2}$. Again, it may be verified that these are properties of the Boolean expression (4.4) that underlies the prognostic model. For example, these properties are not satisfied by the AND function, which, therefore, cannot be selected as a basis for a prognostic model that satisfies the given qualitative constraints.

4.4 Summary

In this chapter, causal independence models that employ Boolean interaction functions have been analyzed. In contrast to previous work, (Lucas, 2005), the chapter

offers a characterization of causal independence models based on Boolean functions in general, and it can, thus, be used as a foundation for the analysis of any of such causal independence models. It was shown that QPN theory can be applied to these models in order to characterize model behavior in terms of influences and synergies. By making use of difference functions and an order on Boolean tuples we were able to derive both the conditions under which positive, negative, zero, non-monotonic and ambiguous signs for qualitative influences, additive synergies and product synergies are observed and the constraints these signs impose on the underlying interaction functions.

In conclusion, we believe that the theory can aid in Bayesian network construction, where the prognostic model served as an example to illustrate the usefulness of the theory in practice. If the causal independence assumptions hold then the appropriateness of an interaction function can be determined without the need to specify the parameters in advance and properties of the interaction function can be derived from a qualitative specification.

Chapter 5

Dynamic Decision Making with DLIMIDs

According to the norms dictated by utility theory, rational clinical decision making implies the maximization of patient benefit, while simultaneously minimizing the cost of treatment (Von Neumann and Morgenstern, 1947). For instance, in our research, we have focused on finding treatment strategies for high-grade carcinoid tumors; an aggressive type of neuroendocrine tumor (Zuetenhorst and Taal, 2005). For these tumors, it is of the utmost importance that chemotherapy is administered at the right moments in time. Treating a patient too early, or too long, may cause an unnecessary deterioration in general health status, whereas treating a patient too late, or too short, may fail to stop or reverse tumor progression. Solving such *dynamic decision problems* (Magni and Bellazzi, 1997) is a difficult task, since it requires the physician to take appropriate action at each point in time, by taking into account the patient's history, in a world that is characterized by uncertainty.

The selection of strategies that lead to optimal patient treatment has received considerable attention from both the Operations Research and Artificial Intelligence communities, where it is known as *stochastic control* and *planning* respectively. In recent years, emphasis has been placed on the similarities and differences between stochastic control and *decision-theoretic* planning, where probability theory and utility theory are used to represent decision-making under uncertainty (Dean and Wellmann, 1991; Boutilier et al., 1996a). In this work, we introduce *dynamic limited-memory influence diagrams* (DLIMIDs) which inherit characteristics from both approaches to decision making under uncertainty. They can be represented compactly as a *temporal limited-memory influence diagram* (TLIMID) and allow the modeling of dynamic decision problems that are only partially observable and may go on for an unbounded amount of time. We also introduce a number of algorithms that approximate the optimal strategy for dynamic decision problems that are modeled as a DLIMID. This is demonstrated by a DLIMID that models high-grade carcinoid tumor

This chapter is based on (van Gerven et al., 2006a; van Gerven and Díez, 2006; van Gerven et al., 2006b).

pathophysiology and has been constructed in collaboration with an expert physician at the Netherlands Cancer Institute (NKI).

This chapter proceeds as follows. In Section 5.1, we describe the perspectives on dynamic decision making that are offered by stochastic control and decision-theoretic planning, in order to make clear the differences and similarities between the two approaches. DLIMIDs and algorithms that approximate optimal strategies are defined in Sections 5.2 and 5.3 respectively. Section 5.4 describes the construction of the oncological model, and Section 5.5 describes experimental results concerning the strategies found for the model, using the described algorithms. We end with some concluding remarks in Section 5.6.

5.1 Perspectives on dynamic decision making

Stochastic control and decision-theoretic planning both offer a different approach to dynamic decision making. Stochastic control is often realized by means of *Markov decision processes*, whereas decision-theoretic planning is often realized by means of *(dynamic) influence diagrams*. In this section we describe both approaches, their solution strategies, and their respective strengths and weaknesses.

5.1.1 Markov decision processes

One way to model dynamic decision problems is by means of the theory of *Markov decision processes* (MDPs) (Howard, 1960). MDPs are extensions of *Markov chains*, defined as follows (Grimmett and Stirzaker, 1992).

Definition 5.1. *Let S be a* discrete-time random process*, that is, a family of random variables $\{S(t) : t \in T\}$ that take values from Ω_S and are indexed by some set $T = \{0, \ldots, N\}$, where N denotes the* horizon*. A* Markov chain *is a discrete-time process that satisfies the* Markov condition*:*

$$P(S(n) = s_n \mid S(1) = s_1, \ldots, S(n-1) = s_{n-1}) = P(S(n) = s_n \mid S(n-1) = s_{n-1})$$

for $n \geq 1$ and $s_1, \ldots, s_n \in \Omega_S$.

The Markov condition ensures that the future state is independent of the past state given the current state of a random process. A Markov decision process extends a Markov chain by allowing *actions* and *rewards* to incorporate both choice and motivation (Fig. 5.1).

Definition 5.2. *A* Markov decision process *(MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the probability that the system ends up in state s' at time $t+1$, given that action a was performed in state s at time t, and $R(s, a) \in \mathbb{R}$ is the reward for taking an action a in state s.*

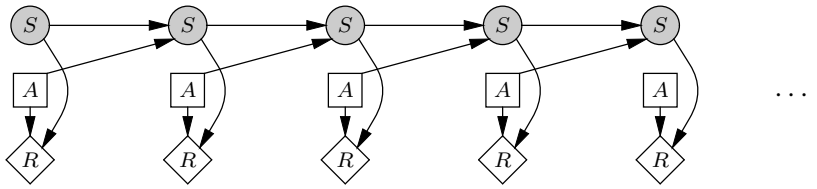


Figure 5.1: A Markov decision process, where shading indicates observability of the state.

Markov decision processes have proven very useful for cost-effectiveness analysis in medicine (Sonnenberg and Beck, 1993; Kuntz and Weinstein, 2001). The goal of a rational decision maker is to maximize expected reward

$$\mathrm{E}\left(\sum_{t\in T}\gamma^t R(s_t,a_t)\right)$$

where $\gamma \in [0,1]$ is a *discount factor*. Usually $\gamma < 1$, which implies that delayed rewards are less valuable to the decision maker. The expected reward is maximized by choosing an optimal sequence of actions for all $t \in T$, as represented by a *policy* $\pi_t \colon S \to A$, which maps states to actions at each decision moment $t \in T$. This mapping can be either *stochastic*, allowing for randomness in the actions, or *deterministic*, defining a fixed mapping between states and actions. If the index set T is finite then we speak of a *finite-horizon* MDP and if it is infinite then we speak of an *infinite-horizon* MDP.

An important result is that for infinite-horizon MDPs, the optimal policy is *stationary* (independent of t) and deterministic, whereas for finite-horizon MDPs the optimal policy is typically non-stationary (Howard, 1960). Let $V_{\pi,t}(s)$ denote the expected value of starting in state s, when there are still t steps to go, while executing policy π. In the finite-horizon case, the expected value that is gained by using the optimal policy π^*, is given by the Bellman equations:

$$\begin{aligned} V_{\pi^*,1}(s) &= \max_{a\in\mathcal{A}}\{R(s,a)\} \\ V_{\pi^*,t}(s) &= \max_{a\in\mathcal{A}}\left\{R(s,a)+\gamma\sum_{s'\in\mathcal{S}}P(s'\mid a,s)V_{\pi^*,t-1}(s')\right\}, \quad t>1 \end{aligned}$$

with $\pi^* = \{\pi_t^* \colon t \in T\}$. In the infinite-horizon case, we simply have

$$V_{\pi^*}(s) = \max_{a\in\mathcal{A}}\left\{R(s,a)+\gamma\sum_{s'\in\mathcal{S}}P(s'\mid a,s)V_{\pi^*}(s')\right\}$$

since both policy and expected reward are independent of t. Finding (approximations to) optimal policies for MDPs is relatively straightforward. In the finite-horizon case,

the standard method is to perform a backward recursion on the Bellman equations, whereas in the infinite-horizon case, we may use techniques such as *value iteration* (Bellman, 1957) or *policy iteration* (Howard, 1960).

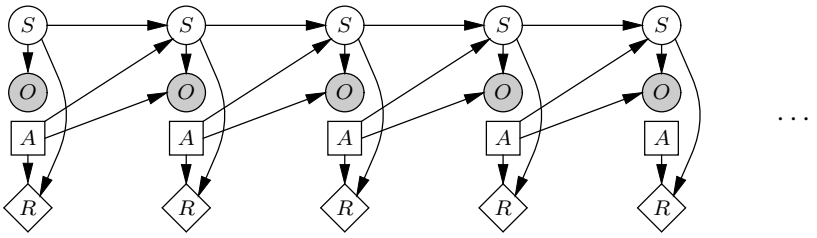


Figure 5.2: A partially-observable Markov decision process, where shading indicates observability.

MDPs assume that the state of the process is completely observable. In practice, however, we often have incomplete state information. For instance, in medicine, progression of a disease can often only be determined by observable symptoms or laboratory findings. This brings us into the realm of *partially observable Markov decision processes* (Aström, 1965; Monahan, 1982), as shown in Fig. 5.2.

Definition 5.3. *A* partially observable Markov decision process *(POMDP) is a tuple* $(\mathcal{S}, \mathcal{A}, \mathcal{O}, P, R, Q)$*, such that* $(\mathcal{S}, \mathcal{A}, P, R)$ *defines a Markov decision process,* $\mathcal{O}$ *is a finite set of* observations*, and* $Q(o \mid a, s)$ *is the probability of observing* o *given that we landed in state* s *at time* $t+1$*, while performing action* a *at time* t*.*

In order to make optimal decisions, POMDPs take into account all past observations by maintaining a *belief state* with respect to the (hidden) state of the process (Smallwood and Sondik, 1973). Let $b(s)$ denote the current belief of the decision-maker that the process is in state s. Given $b(s)$, an observation o and executed action a, we estimate the next belief state from Bayes' rule as:

$$b'(s') = \alpha \cdot Q(o \mid a, s') \sum_{s \in S} P(s' \mid a, s) b(s) \tag{5.1}$$

where α is a normalizing constant. We define the *state estimator*: $P(b' \mid a, b, o)$, which assigns a probability of one to the belief state that is compatible with Eq. (5.1) and zero to all other belief states. The corresponding Bellman equation for infinite-horizon POMDPs is then given by:

$$V_{\pi^*}(b) = \max_{a \in \mathcal{A}} \left\{ R(b, a) + \gamma \sum_{b'} P(b' \mid a, b) V_{\pi^*}(b') \right\}$$

where

$$R(b, a) = \sum_{s \in \mathcal{S}} b(s) R(s, a)$$

and

$$P(b' \mid a, b) = \sum_{o \in \mathcal{O}} P(b' \mid a, b, o) P(o \mid a, b)\,,$$

with state estimator $P(b' \mid a, b, o)$, and $P(o \mid a, b) = \sum_{s \in \mathcal{S}} Q(o \mid a, s) b(s)$.

By reformulating the POMDP in terms of an MDP in belief space (Aström, 1965; Smallwood and Sondik, 1978), the POMDP can be solved by applying dynamic programming techniques to the corresponding MDP. The difficulty is, however, that a belief is a point in the n-dimensional simplex, where n is the number of states. This implies an infinite number of belief states, and requires the construction of a policy that maps this infinite number of states to actions. It has been shown that, in the finite-horizon case, the optimal value function is piecewise-linear and convex (Smallwood and Sondik, 1973), thus requiring only a finite mapping of situations to actions. In the infinite-horizon case, however, the optimal value function no longer consists of a finite number of linear elements, although it can be approximated arbitrarily closely by a finite horizon-value function (Smallwood and Sondik, 1978). This being said, even approximating the optimal strategy to a sufficient degree is computationally very costly (Papadimitriou and Tsitsiklis, 1987; Lusena et al., 2001), and finding optimal strategies is feasible only for small decision problems (Boutilier et al., 1996a). Another problem associated with the use of (partially-observable) Markov decision processes for modeling dynamic decision problems, is the fact that the state space $\mathcal{S}$ quickly becomes unmanageably large for realistically sized decision problems. This leads to problems, both during specification of the decision process (Magni and Bellazzi, 1997), as well as at computation time (Boutilier et al., 1996a).

5.1.2 Dynamic influence diagrams

An alternative point of departure for modeling the types of decision problems described above is by means of dynamic influence diagrams (Tatman and Shachter, 1990). They extend standard influence diagrams (Howard and Matheson, 1984b) in order to represent finite-horizon decision processes, by decomposing the global utility function into a set of local utility functions. A *dynamic influence diagram* (DID) is a tuple $(\mathbf{C}, \mathbf{D}, \mathbf{U}, \mathbf{A}, P)$, where $\mathbf{N} = \mathbf{C} \cup \mathbf{D} \cup \mathbf{U}$ is a set of nodes that is partitioned into *chance variables* $\mathbf{C}$, *decision variables* $\mathbf{D}$, and *utility functions* $\mathbf{U}$, $\mathbf{A}$ is a set of arcs such that $G = (\mathbf{N}, \mathbf{A})$ forms an acyclic directed graph (ADG), and P is a family of probability distributions. When a DID is used to model a dynamic decision problem, chance variables, decision variables and utility functions are indexed by times $t \in T$.

Chance variables (graphically depicted by circles), are random variables that represent the stochastic component of the model. Decision variables (graphically depicted by squares), are ordinary variables that represent the actions that may be performed by a decision maker. Utility functions (graphically depicted by diamonds), represent the utility of being in a certain state, as defined by configurations of chance

and decision variables. The graph $G = (\mathbf{N}, \mathbf{A})$ represents the qualitative structure of the decision problem. The meaning of an arc $(X, Y) \in \mathbf{A}$ is determined by the type of Y. If $Y \in \mathbf{C}$ then the conditional probability distribution associated with Y is conditioned by X. If $Y \in \mathbf{D}$ then X represents information that is available to the decision maker prior to deciding upon Y; we call the parents $\pi(Y) = \{X\colon (X, Y) \in \mathbf{A}\}$ of decision Y its *informational predecessors*. We also require that there exists a directed path between all decisions $D \in \mathbf{D}$ in G, which represents the order in which decisions are made. Decisions that are made later in time must always inherit the informational predecessors of decision that are made earlier in time, which is known as the *no-forgetting* principle. If $Y \in \mathbf{U}$ then X takes part in the specification of the utility function Y such that $Y\colon \Omega_{\pi(Y)} \rightarrow \mathbb{R}$. Utility functions must either have a subset of the chance and decision variables, or other utility functions, as their parents. In the latter case, we call the utility function U a *super-value node*, where we require that the global utility function is decomposed into a set of local utility functions which combine additively:

$$U(x_1, \ldots, x_n) = \sum_{i=1}^{n} u_i(x_i)\,.$$

The family of probability distributions P is a set $\{P(C \mid \pi(C))\colon C \in \mathbf{C}\}$, such that we have for each configuration $\mathbf{d} \in \Omega_{\mathbf{D}}$ a distribution:

$$P(\mathbf{C}\colon \mathbf{d}) = \prod_{C \in \mathbf{C}} P(C \mid \pi(C)) \tag{5.2}$$

that represents the distribution over $\mathbf{C}$ when the decision maker has set $\mathbf{D}$ equal to $\mathbf{d}$ (Cowell et al., 1999). Hence, $\mathbf{C}$ is not conditioned on $\mathbf{D}$, but rather parameterized by $\mathbf{D}$.[1] Figure 5.3 shows an example of a DID for three consecutive time-slices. A *stochastic policy* for decisions $D \in \mathbf{D}$ is a distribution $P(D \mid \pi(D))$ that maps configurations of $\pi(D)$ to a distribution over alternatives for D. If $P(D \mid \pi(D))$ is degenerate (i.e. consisting of ones and zeros only) then we say that the policy is deterministic. Let $\mathbf{V}$ denote $\mathbf{C} \cup \mathbf{D}$. A *strategy* is a set $\Delta = \{P(D \mid \pi(D))\colon D \in \mathbf{D}\}$ of policies that induces the following joint distribution over the variables in $\mathbf{V}$:

$$P_{\Delta}(\mathbf{V}) = P(\mathbf{C}\colon \mathbf{D}) \prod_{D \in \mathbf{D}} P(D \mid \pi(D))\,. \tag{5.3}$$

We may then compute the expected utility of a strategy Δ as:

$$\mathrm{EU}(\Delta) = \sum_{\mathbf{v}} P_{\Delta}(\mathbf{v}) U(\mathbf{v})\,. \tag{5.4}$$

The aim of any rational decision maker is then to maximize the expected utility by finding an optimal strategy $\Delta^* \equiv \arg\max_{\Delta} \mathrm{EU}(\Delta)$.

[1] This is equivalent to Pearl's *do* operator (Pearl, 2000).

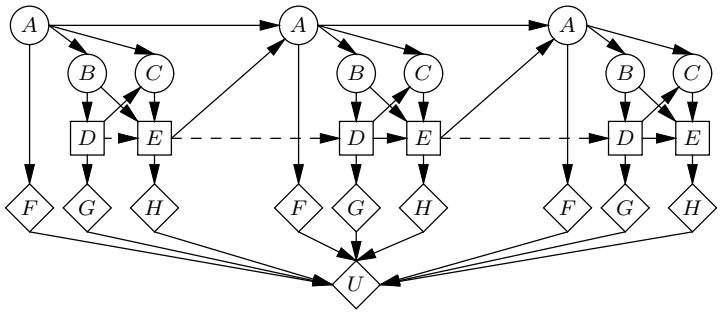


Figure 5.3: A dynamic influence diagram. The dashed arcs emphasize the directed path between decision nodes, which stands for the decision sequence. If a node is an informational predecessors of a decision node, then its use as an informational predecessor for decision nodes occurring later in the sequence, is left implicit in the diagram. The super-value node U combines the local utility functions.

In order to solve a DID, we can resort to a graph reduction algorithm that corresponds to the dynamic programming technique used to solve finite-horizon Markov decision processes (Tatman and Shachter, 1990). There are, however, some salient differences between DIDs and Markov decision processes. The main advantage of (dynamic) influence diagrams over Markov decision processes is the fact that they make use of a factorization of the state-space defined by the variables in the domain. This often allows for more efficient probabilistic inference, the estimation of fewer parameters, and a more meaningful specification in terms of cause-effect relations (Druzdzel, 1997; Owens et al., 1997). A second difference is the way in which partial observability is handled in DIDs. As described, optimal policies for POMDPs are found by solving an MDP in belief space. DIDs follow an alternative strategy, where each decision variable is conditioned by all past observations. Since a belief state follows uniquely from an initial belief state together with a sequence of observations, the approaches give equivalent results. However, DIDs replace the problem of making optimal decisions for an infinite number of belief states, by making optimal decisions for each possible configuration of past observations. Since this becomes infeasible for long decision processes, DIDs are limited to short finite-horizon decision processes.

One way to manage a factorized representation of infinite-horizon Markov decision processes is to specify the state transition matrix, going from time t to time $t+1$, in terms of an influence diagram-like structure. For example, *influence views*, as introduced by Leong as part of her DynaMol framework for dynamic decision analysis (Leong, 1994; Cao et al., 1998), provide, for each possible action, a factorized representation of the transition matrix. Since the influence view distinguishes between *state variables*, which explicitly denote the informational predecessors of a decision node, and *event variables*, which play a supporting role in the representation of the transition matrix, the computational burden of solving a dynamic decision problem

can be reduced (Magni and Bellazzi, 1997). Magni et al. have demonstrated that influence views are suitable for the modeling of realistic dynamic decision problems in medicine (Magni, 1998; Magni et al., 2000). The solution of a dynamic decision problem by means of influence views proceeds by transforming the factorized representation into a normal MDP and applying value iteration. Consequently, the technique is restricted to completely observable MDPs. A similar approach was advocated by Boutilier et al. (Boutilier et al., 1996a), who factorized stationary and completely observable MDPs in terms of a so-called *two-stage temporal Bayes net* (2TBN) (Dean and Kanazawa, 1989). This leads not only to gains in representational efficiency, but also allows for the efficient computation of transition probabilities by means of probabilistic inference over the factorized representation. Boutilier and Poole have used this same factorized representation in order to solve POMDPs in terms of a factorized MDP in belief space (Boutilier and Poole, 1996), and approximate solution techniques have been developed for these factorized representations (Guestrin et al., 2001). The use of POMDPs as factorized MDPs in belief space for clinical decision making has been discussed in (Peek, 1999), and has been applied to the treatment of ischemic heart disease in (Hauskrecht and Fraser, 2000).

5.1.3 LIMIDs

In this chapter, instead of representing a POMDP as a factorized MDP in belief space, we take influence diagrams as our point of departure. We describe an alternative representation that crucially depends on the limited-memory assumption that strategies based on a limited amount of memory for each decision will be able to approximate the optimal strategy. *Limited-memory influence diagrams* (LIMIDs) (Lauritzen and Nilsson, 2001) incorporate this assumption, and are otherwise defined analogous to DIDs.[2] The limited-memory assumption allows us to drop the requirement that a complete order is defined over decisions, thereby increasing the variety of decision problems that can be handled (Fig. 5.4).

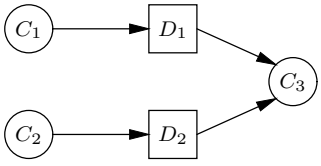


Figure 5.4: A LIMID allows the decisions D_1 and D_2 to be made in parallel and with different informational predecessors, thereby increasing the variety of decision problems that can be handled.

The algorithm that approximates the optimal strategy in LIMIDs, as described in (Lauritzen and Nilsson, 2001), is much more efficient than the algorithms that find

[2] A dynamic influence diagram is just the special case of a LIMID that takes all past observations into account.

the optimal strategy in standard influence diagrams. The latter keep track of all past observations, which becomes infeasible when many (subsequent) decisions need to be made. However, in case of infinite-horizon decision processes, finding approximately optimal strategies in LIMIDs becomes infeasible as well, since decisions are represented explicitly at each point in time.

5.2 Dynamic limited-memory influence diagrams

In order to enable the representation of infinite-horizon POMDPs in terms of LIMIDS, we define dynamic LIMIDs, that can be represented compactly by means of temporal LIMIDS. In Section 5.3, we introduce a number of algorithms that approximate the optimal strategy for a dynamic LIMID.

5.2.1 Constructing DLIMIDs

A *dynamic LIMID* (DLIMID) is defined as a LIMID $(\mathbf{C}, \mathbf{D}, \mathbf{U}, \mathbf{A}, P)$, that models a dynamic decision problem, such that chance variables, decision variables, or utility functions at time t can only depend on other chance variables or decision variables at times $\mathbf{K}_t = \{t-K, \ldots, t\}$. Hence, a DLIMID is a factorized representation of a K-th order POMDP. If a DLIMID is explicitly defined at times $\mathbf{K}_0 = \{0, \ldots, K-1\}$ and has fixed structure and parameters for all $t \in \{K, \ldots, N\}$, where N is the horizon, then a DLIMID can be represented more compactly as a *temporal* LIMID. In the following, we omit time indices when clear from context.

Definition 5.4. *A* temporal LIMID *(TLIMID) is a pair of LIMIDs $(\mathcal{L}_0, \mathcal{L}_t)$ that respects the following conditions:*

- *The* prior model $\mathcal{L}_0 = (\mathbf{C}_0, \mathbf{D}_0, \mathbf{U}_0, \mathbf{A}_0, P_0)$ *is defined for times $\mathbf{K}_0$ where for all arcs $(X(u), Y(v)) \in \mathbf{A}_0$ it holds that $u \leq v$, and*

$$P_0 = \{P(X \mid \pi(X))\colon X \in \mathbf{N}_0\}$$

  with $\mathbf{N}_0 = \mathbf{C}_0 \cup \mathbf{D}_0 \cup \mathbf{U}_0$.

- *The* transition model $\mathcal{L}_t = (\mathbf{C}_t, \mathbf{D}_t, \mathbf{U}_t, \mathbf{A}_t, P_t)$ *is defined for times $\mathbf{K}_t$ where for all arcs $(X(u), Y(v)) \in \mathbf{A}_t$ it holds that $u \leq v$ and $v = t$, and*

$$P_t = \{P(X \mid \pi(X))\colon X \in \mathbf{N}_t\}.$$

  with $\mathbf{N}_t = \mathbf{C}_t \cup \mathbf{D}_t \cup \mathbf{U}_t$.

A TLIMID allows for the representation of (infinite-horizon) POMDPs. The prior model is used to represent the initial distribution $P(\mathbf{C}_0\colon \mathbf{D}_0)$ and utility functions $U \in \mathbf{U}_0$ at the first K time slices. The transition model is not yet bound to any

specific t, but if bound to some $t \in \{K, \ldots, N\}$, then it is used to represent the conditional distribution $P(\mathbf{C}_t \colon \mathbf{D}_{t-K}, \ldots, \mathbf{D}_t)$ and utility functions $U \in \mathbf{U}_t$ for some $t \geq N$. The graph $G = (\mathbf{N}_t, \mathbf{A}_t)$ does not depend on t, and normally it is assumed that P_t does not depend on t either. This is not a strict requirement however, which allows the representation of non-stationary POMDPs where probability distributions are a function of t. Figure 5.5 shows an example of a TLIMID as a factorized representation of a K-th order POMDP.

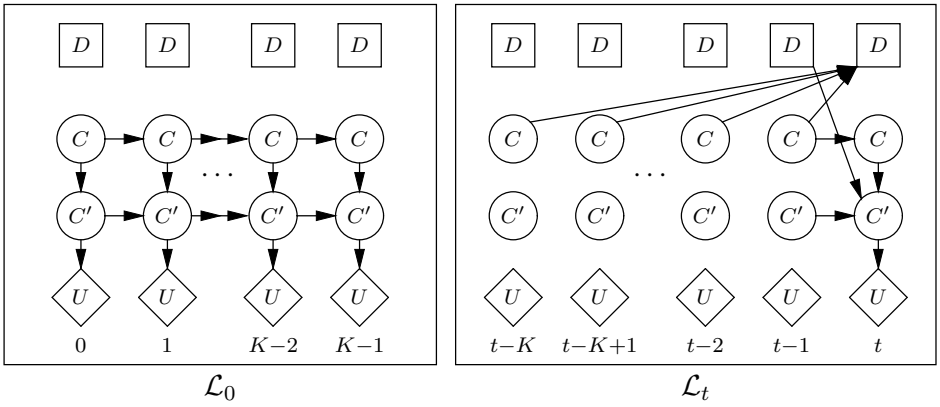


Figure 5.5: Representation of a K-th order POMDP by a TLIMID, where chance nodes are shown as circles, decision nodes as squares and utility nodes as diamonds. The prior model $\mathcal{L}_0$ depicts the situation for the initial K time points, whereas the transition model $\mathcal{L}_t$ depicts how the situation at a time t depends on the previous K time points. In this particular case, we have a model where the decision D has no effect in the prior model, whereas in the transition model it influences C'_t through D_{t-1}, and has $C_{t-K}, \ldots, C_{t-1}$ as its informational predecessors.

Given a horizon N, we may *unroll* a TLIMID for $N - K$ *time-slices* in order to obtain a DLIMID with the following joint distribution:

$$P(\mathbf{C} \colon \mathbf{D}) = P(\mathbf{C}_0 \colon \mathbf{D}_0) \prod_{t=K}^{N} P(\mathbf{C}_t \colon \mathbf{D}_{t-K}, \ldots, \mathbf{D}_t)\,. \tag{5.5}$$

Let $\mathbf{V}$ again denote $\mathbf{C} \cup \mathbf{D}$, and let $\Delta_t = \{P(D \mid \pi(D)) \mid D \in \mathbf{D}_t\}$ denote the strategy for time t. Given a strategy $\Delta_0 = \bigcup_{t \in \mathbf{K}_0} \Delta_t$, $\mathcal{L}_0$ defines the following distribution over the variables in $\mathbf{V}_0$:

$$P_{\Delta_0}(\mathbf{V}_0) = P(\mathbf{C}_0 \colon \mathbf{D}_0) \prod_{D \in \mathbf{D}_0} P(D \mid \pi(D))\,, \tag{5.6}$$

and given a strategy $\Delta_t = \bigcup_{t \in \mathbf{K}_t} \Delta_t$, with $t \geq K$, $\mathcal{L}_t$ defines the following conditional distribution over the variables in $\mathbf{V}_t$:

$$P_{\Delta_t}(\mathbf{V}_t \mid \mathbf{I}_t) = P(\mathbf{C}_t \colon \mathbf{D}_{t-K}, \ldots, \mathbf{D}_t) \prod_{D \in \mathbf{D}_t} P(D \mid \pi(D)) \tag{5.7}$$

where $\mathbf{I}_t = \{X(t') \colon t' < t, (X(t'), Y(t)) \in \mathbf{A}_t\}$ is the *interface* of the transition model, representing the variables that have a direct influence on variables in $\mathbf{V}_t$.

Combining Eqs. (5.6) and (5.7), given a horizon N and strategy $\Delta = \bigcup_{t \in T} \Delta_t$, a TLIMID induces the following distribution over variables in $\mathbf{V}$:

$$P_\Delta(\mathbf{V}) = P_{\Delta_0}(\mathbf{V}_0) \prod_{t=K}^{N} P_{\Delta_t}(\mathbf{V}_t \mid \mathbf{I}_t)\,. \tag{5.8}$$

Let $\mathcal{U}_t = \sum_{U \in \mathbf{U}_t} U$ denote the joint utility for a time-slice t. We define the joint utility function for a dynamic LIMID as

$$\mathcal{U} = \sum_{t=0}^{N} \gamma^t \mathcal{U}_t \tag{5.9}$$

with discount factor $\gamma \in [0, 1]$, such that the expected utility of a strategy Δ is given by $\mathrm{EU}(\Delta) = \sum_{\mathbf{v}} P_\Delta(\mathbf{v}) \mathcal{U}(\mathbf{v})$.

5.2.2 Representing observed history

As remarked before, if decisions are allowed to depend on all past observations, then a DLIMID becomes computationally intractable for all but small finite-horizon decision processes. Therefore, we can only hope to find (approximations to) the optimal strategy, where each policy is based on a limited number of past observations.[3] It is clear from Fig. 5.5 that, if we use a TLIMID, policies take into account *at most* all chance and decision variables in K subsequent time-slices since $\pi(D_t) \subseteq \mathbf{V}_{t-K} \cup \cdots \cup \mathbf{V}_t$ (cf. Eq. (5.5)). Observations made earlier in time will not be taken into account and as a result, states that are qualitatively different can appear the same to the decision maker, leading to suboptimal policies. In reinforcement learning, this phenomenon is known as *perceptual aliasing* (Whitehead and Ballard, 1991), indicating that active perception of the world can have as a consequence that the agent's internal representation confounds external world states. In order to alleviate the problem of perceptual aliasing, there are a number of ways to relax the strong limited-memory assumption implied by TLIMIDs. One way to resolve this problem is by using a large value for K. This still allows us to represent large decision processes, and as K approaches N, we will find better approximations to the optimal strategy Δ^* in general.

An alternative way to deal with perceptual aliasing, as used in this chapter, is to assume that the first-order Markov assumption that the future is independent of the past given the present holds ($K = 1$), and to represent part of the observational history by means of *memory variables* $\mathbf{M} \subseteq \mathbf{C}$. As shown in Fig. 5.6, one way to

[3] In the context of POMDPs, methods that rely on the use of a finite history are common, and can be dated back to (Brown, 1972).

maintain memory concerning chance and decision variables, is to associate a unique memory variable $M \in \mathbf{M}$ with each informational predecessor $V \in \pi(D)$ for all $D \in \mathbf{D}$. The TLIMID is then redefined by using memory variables M as the informational predecessors of D, and by requiring that $(V(0), M(0))$ is an arc in $\mathbf{A}_0$, and both $(V(t), M(t))$ and $(M(t\text{-}1), M(t))$ are arcs in $\mathbf{A}_t$.

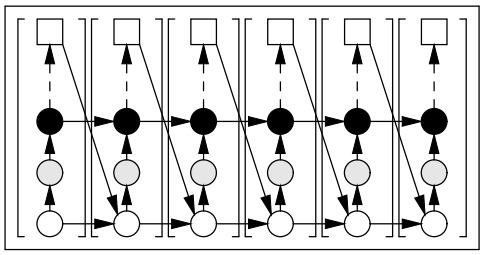


Figure 5.6: Dealing with perceptual aliasing by introducing memory variables (black circles). Memory variables are used instead of associated observed variables (shaded circles), as the informational predecessor for a decision variable (squares).

Memory about the past is maintained by means of distributions $P(M(0) \mid V(0))$ and $P(M(t) \mid M(t\text{-}1), V(t))$. For example, suppose we would like to maintain a memory about the past two time-slices. Then it suffices to define

$$\Omega_{M(t)} \;=\; \Omega_{V(t)} \cup \Omega_{V(t)} \times \Omega_{V(t-1)} \cup \Omega_{V(t)} \times \Omega_{V(t-1)} \times \Omega_{V(t-2)}\,,$$

which represents all possible observational histories of length three, and to use the distributions to maintain changes in the observational history. Note that such an explicit enumeration of all observational histories leads to a huge state space for M. Therefore, we normally represent the observational history of V more compactly by partitioning all possible observational histories into a small set of states. In this way, we use *aggregation* (Boutilier et al., 1996a) to group states that are indistinguishable from the point of view of the decision maker. The choice of the states of M is problem dependent, and we will not further address this issue in this chapter. Instead, it is assumed that their definition is based on available domain knowledge.

In Section 5.4 we define a TLIMID for a dynamic decision problem in medicine that uses two memory variables treathist and bmdhist. Here, treathist is a short-term memory variable that represents the three latest observations, while bmdhist is a long-term memory variable that indicates whether the patient has ever had bone-marrow depression.

5.3 Improving strategies in infinite-horizon DLIMIDs

In the previous section we have shown how a DLIMID, constructed from a TLIMID, can represent an infinite-horizon Markov decision process. We proceed by exploring techniques for approximating the optimal strategy.

5.3.1 Computing expected utility

In order to compute the expected utility for a TLIMID, we resort to an indirect approach, where we make use of the fact that given Δ, an influence diagram $(\mathbf{N}, \mathbf{A}, P)$ may be converted into a Bayesian network, which can subsequently be used as a computational architecture for decision making under uncertainty (Cooper, 1988; Shachter and Peot, 1992). Since a strategy Δ induces a distribution over variables in $\mathbf{V}$ (cf. Eq. (5.8)), we can use Δ to convert decision variables $D \in \mathbf{D}$ into random variables $X \in \mathbf{X}$, with parents $\pi(D)$ such that:

$$P(X \mid \pi(X)) = P(D \mid \pi(D)) .$$

Additionally, utility functions $U \in \mathbf{U}$ may be converted into random variables $X \in \mathbf{X}$, with parents $\pi(U)$. We define the distribution $P(X \mid \pi(X))$ with $\Omega_X = \{0, 1\}$ by means of a transformation:

$$P(X{=}1 \mid \mathbf{x}') = \frac{U(\mathbf{x}') - \min_{\mathbf{x}} U(\mathbf{x})}{\max_{\mathbf{x}} U(\mathbf{x}) - \min_{\mathbf{x}} U(\mathbf{x})}$$

with $\mathbf{x}, \mathbf{x}' \in \Omega_{\pi(U)}$, as defined in (Cooper, 1988). This allows us to compute the expected utility $\mathrm{EU}(\Delta)$ given a strategy Δ directly, by using the Bayesian network to compute the posterior probability of X, and performing the reverse transformation on the probability of X. We use $B(\mathcal{L}, \Delta)$ to denote the conversion of a LIMID $\mathcal{L}$, given a strategy Δ, into a Bayesian network $\mathcal{B}$.

Given Δ, we may convert a TLIMID $(\mathcal{L}_0, \mathcal{L}_t)$ into the pair $(\mathcal{B}_0, \mathcal{B}_t)$ with $\mathcal{B}_0 = B(\mathcal{L}_0, \Delta_0)$ and $\mathcal{B}_t = B(\mathcal{L}_t, \Delta_t)$, the latter of which is also known as a two-stage temporal Bayes net. The pair $(\mathcal{B}_0, \mathcal{B}_t)$ is often used to construct a *dynamic Bayesian network (DBN)* (Dean and Kanazawa, 1989; Boutilier et al., 1996a). The transformation of a TLIMID into $(\mathcal{B}_0, \mathcal{B}_t)$ and of a DLIMID into an unrolled DBN are depicted in Fig. 5.7.

As the figure suggests, one way to do probabilistic inference is to unroll $(\mathcal{B}_0, \mathcal{B}_t)$ into one big static network and to use a standard inference algorithm, such as the junction tree algorithm (Cowell et al., 1999). However, although the complexity of inference is determined by the size of the largest clique that is obtained after triangularization of the graph underlying the static network (Dechter and Rish, 1994), space complexity also grows linearly in the horizon N, and therefore this approach is unsuitable for large horizons. For online inference, more efficient inference algorithms exist that operate directly on $(\mathcal{B}_0, \mathcal{B}_t)$. We have used the *interface algorithm*, which uses a triangulation method such that the space and time taken to compute $P(\mathbf{X}(t) \mid \mathbf{X}(t\text{-}1))$ does not depend on the number of time-slices (Murphy, 2002). As the size of the model grows, exact inference may become infeasible, and then we may resort to deterministic or stochastic approximate inference schemes like loopy belief propagation (Murphy et al., 1999) or particle filtering (Doucet et al., 2001).

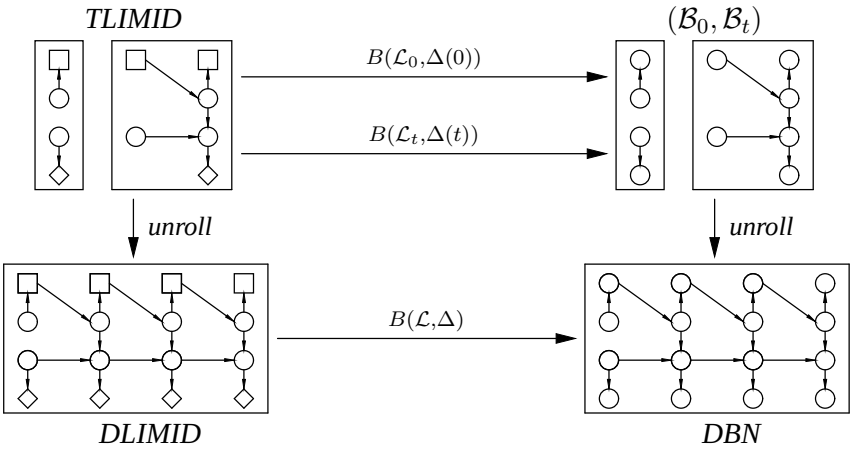


Figure 5.7: Converting between different representations for the special case that the TLIMID represents a first-order POMDP ($K = 1$), as assumed throughout the remainder of this chapter.

In order to compute an approximation to the expected utility given Δ, we assume that the TLIMID $(\mathcal{L}_0, \mathcal{L}_t)$ represents a first order POMDP ($K = 1$), and Δ can be expressed as a pair (Δ_0, Δ_t), where Δ_0 is the strategy at $t = 0$ and Δ_t is a *stationary* strategy that does not depend on t for $t > 0$. Recall that the optimal strategy is deterministic and stationary for infinite-horizon Markov decision processes (Ross, 1983). However, in the partially observable case, we can only expect to find approximations to the optimal strategy by using memory variables that represent part of the observational history (Meuleau et al., 1999). The approximation $\mathrm{EU}^{\kappa}(\Delta)$ to the expected utility is made by computing the discounted expected utility ($\gamma < 1$) using $(B(\mathcal{L}_0, \Delta_0), B(\mathcal{L}_t, \Delta_t))$ for a finite number of time-slices κ. Here, κ may be chosen based on the problem characteristics, or based on some error criterion ϵ. For instance, by choosing

$$\kappa = \log_{\gamma}(\epsilon(1-\gamma)/2u_{\max}),$$

where $u_{\max}$ stands for the maximum utility obtainable during one time-slice, we ensure that at most $\epsilon/2$ error is introduced into the approximation (Ng and Jordan, 2000).

5.3.2 Single policy updating

One way to improve strategies in standard LIMIDs is to use an iterative procedure called *single policy updating* (SPU) (Lauritzen and Nilsson, 2001). Let

$$\Delta^0 = \{p_1, \ldots, p_n\}$$

be an ordered set representing the initial strategy, where p_j with $1 \leq j \leq n$ stands for a (randomly initialized) policy P_{D_j}. We say p_j is the *local maximum policy* for

a strategy Δ at decision D_j if EU(Δ) cannot be improved by changing p_j. In SPU, each cycle iterates over all decision variables to find local maximum policies, and reiterates until no further improvement in expected utility can be achieved. SPU converges in a finite number of cycles to a *local maximum strategy* Δ where each $p_j \in \Delta$ is a local maximum policy. Note that this local maximum strategy is not necessarily the global maximum strategy Δ^*. Let

$$\Delta^0 = \Delta_0 \cup \Delta_t$$

be the initial strategy, with $\Delta_0 = \{p_1, \ldots, p_m\}$ and $\Delta_t = \{p_{m+1}, \ldots, p_n\}$, where m is the number of decision variables in $\mathcal{L}_0$ and $n - m$ is the number of decision variables in $\mathcal{L}_t$. Following (Lauritzen and Nilsson, 2001), we define $p'_j * \Delta$ as the strategy obtained by replacing p_j with p'_j in Δ. SPU based on a TLIMID $\mathcal{T}$ with initial strategy Δ^0 is then defined by Algorithm 5.1.

Algorithm 5.1 Single policy updating for TLIMIDs.

input: TLIMID $\mathcal{T}$, initial random strategy Δ^0, stopping criterion κ
$\Delta = \Delta^0$, *euMax* $= \text{EU}^\kappa(\Delta^0)$.
repeat
  euMaxOld $=$ *euMax*
  for $j = 1$ to n **do**
    for all policies p'_j for Δ at D_j **do**
      $\Delta' = p'_j * \Delta$
      if $\text{EU}^\kappa(\Delta') >$ *euMax* **then**
        $\Delta = \Delta'$ **and** *euMax* $= \text{EU}^\kappa(\Delta')$
      end if
    end for
  end for
until *euMax* $=$ *euMaxOld*
return Δ

In case of a (non-temporal) LIMID, a locally optimal policy can be found by optimizing each single rule independently of the others, such that we need to evaluate km^r different policies at each decision variable D, where k denotes the cardinality of Ω_D, and r is the number of informational predecessors of D, assuming that the cardinality of Ω_{V_j} equals m for all $V_j \in \pi(D)$. However, in case of *dynamic* LIMIDs with stationary policies, the optimal rule for a certain scenario at time t depends on the policies applied at future times, which leads to a coupling of the rules. The number of policies that need to be evaluated at each decision variable D therefore grows as $k^{(m^r)}$, such that it becomes impossible in practice to iterate over all possible policies for D.

5.3.3 Single rule updating

For reasons exposed in the previous section, we use a hill-climbing search for DLIMIDs, called *single rule updating* (SRU), that is equivalent to single policy updating for LIMIDs. A deterministic policy can be viewed as a mapping $p_j : \Omega_{\pi(D_j^t)} \rightarrow \Omega_{D_j^t}$, describing for each configuration of the informational predecessors of a decision variable D_j^t an action $a \in \Omega_{D_j^t}$. We call $(\mathbf{x}, a) \in p_j$ a *decision rule*. Instead of exhaustively searching over all possible policies for each decision variable, we try to increase the expected utility by local changes to the decision rules within the policy. I.e., at each step we change one decision-rule within the policy, accepting the change when the expected utility increases. We use $(\mathbf{x}, a') * p_j$ to denote the replacement of $(\mathbf{x}, a)$ by $(\mathbf{x}, a')$ in p_j. Similarly to SPU, we keep iterating until there is no further increase in the expected utility. Using single rule updating, we decrease the number of policies that need to be evaluated in each *local* cycle for a decision node to only km^r, where notation is as before, albeit at the expense of replacing the exhaustive search by a hill-climbing strategy, which increases the risk of ending up in a local maximum, and having to run local cycles until convergence. SRU based on a TLIMID $\mathcal{T}$ with initial strategy Δ^0 is then defined by Algorithm 5.2.

Algorithm 5.2 Single rule updating for TLIMIDs.

input: TLIMID $\mathcal{T}$, initial random strategy Δ^0, stopping criterion κ
$\Delta = \Delta^0$, *euMax* $= \text{EU}^{\kappa}(\Delta^0)$
repeat
  euMaxOld $=$ *euMax*
  for $j = 1$ to n **do**
    repeat
      euMaxLocal $=$ *euMax*
      for all configurations $\mathbf{x}$ of $\pi(D_j)$ **do**
        for all actions $a' \in \Omega_{D_j}$ **do**
          $p'_j = (\mathbf{x}, a') * p_j$
          $\Delta' = p'_j * \Delta$
          if $\text{EU}^{\kappa}(\Delta') >$ *euMax* **then**
            $\Delta = \Delta'$ **and** *euMax* $= \text{EU}^{\kappa}(\Delta')$
          end if
        end for
      end for
    until *euMax* $=$ *euMaxLocal*
  end for
until *euMax* $=$ *euMaxOld*
return Δ

The local maximum strategies returned by SRU (and occasionally also SPU) may differ from the global maximum strategy, Δ^*, as can be seen in the following example.

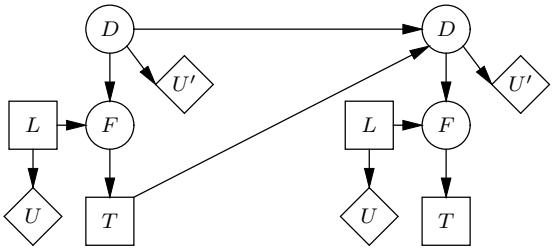


Figure 5.8: A DLIMID for treatment of patients that may or may not have a *disease* D. The disease can be identified by a *finding* F, which is the result of a *laboratory test* L, having an associated cost that is captured by the utility function U. Based on the finding, we decide whether or not to perform *treatment* T. If the patient does not have the disease then this has an associated utility U'.

Example 5.1. Suppose the best strategy for the DLIMID shown in Fig. 5.8 is to always test, to treat when the outcome is positive, and not to treat when the outcome is negative. Suppose the initial strategy Δ^0 is to never test and always treat. Trying to improve the policy for the laboratory test L we find that performing the test will only decrease the expected utility since the test has no informational value (we always treat) but does have an associated cost. Conversely, trying to improve the policy for treatment we find that, as the test has not been performed, it is safer to always treat. Hence, SPU and SRU will stop after one cycle, returning the proposed strategy as the local optimal strategy.

5.3.4 Simulated annealing

In order to improve upon the strategies found by SRU, we resort to *simulated annealing* (SA), which is a heuristic search method that tries to avoid getting trapped into local maximum solutions found by hill-climbing techniques such as SRU (Kirkpatrick et al., 1983). SA chooses candidate solutions by looking at neighbors of the current solution as defined by a *neighborhood function*. Local maxima are avoided by sometimes accepting worse solutions according to an *acceptance function*. In this chapter, we have chosen the acceptance function

$$P(a(\Delta') = \text{yes} \mid eu, eu', t) = \begin{cases} 1 \text{ if } eu' > eu \\ \mathrm{e}^{\frac{eu'-eu}{T(t)}} \text{ otherwise} \end{cases}$$

where $a(\Delta')$ stands for the acceptance of the proposed strategy Δ', $eu' = \mathrm{EU}^{\kappa}(\Delta')$, $eu = \mathrm{EU}^{\kappa}(\Delta)$ for the current strategy Δ, and T represents the temperature in an *annealing schedule* defined as

$$T(t+1) = \alpha \cdot T(t)$$

where $T(0) = \beta$ with $\alpha < 1$ and $\beta > 0$. The annealing schedule ensures that initially a random search through the space of strategies is performed, which gradually

changes into a hill-climbing search. We refer to (Eglese, 1990) for a discussion about choices that can be made for SA parameters α and β. With respect to strategy finding in dynamic LIMIDs, we propose an initial simulated annealing scheme and a subsequent application of SRU in order to greedily find a local maximum solution. Let θ denote a random variable that is repeatedly chosen uniformly at random between 0 and 1, and let $T_{\min}$ stand for the minimum temperature for which we perform the annealing. SA based on a TLIMID $\mathcal{T}$ with initial strategy Δ^0 is then defined by Algorithm 5.3.

Algorithm 5.3 Simulated annealing for TLIMIDs.

input: TLIMID $\mathcal{T}$, initial random strategy Δ_0, stopping criterion κ,
annealing schedule T, minimum temperature $T_{\min}$
$\Delta = \Delta^0$, $t = 0$, $eu = \text{EU}^{\kappa}(\Delta)$
repeat
  select a random decision variable D_j
  select a random decision rule $(\mathbf{x}, a) \in p_j$
  select a random action $a' \in \Omega_{D_j}$, $a' \neq a$
  $p'_j = (\mathbf{x}, a') * p_j$
  $\Delta' = p'_j * \Delta$
  $eu' = \text{EU}^{\kappa}(\Delta')$
  if $\theta \leq P(a(\Delta') = \text{yes} \mid eu, eu', t)$ **then**
    $\Delta = \Delta'$
    $eu = eu'$
  end if
  $t = t + 1$
until $T(t) < T_{\min}$
return SRU$(\mathcal{T}, \Delta, \kappa)$

In Section 5.5, we illustrate the application of the simulated annealing algorithm to a real-world problem in oncology that is described in the following section.

5.4 A dynamic decision problem in medicine

We have applied DLIMIDs to the problem of treatment selection for high-grade carcinoid tumor patients.[4] A carcinoid tumor is a type of neuroendocrine tumor that is predominantly found in the midgut and is normally characterized by the production of excessive amounts of biochemically active substances, such as serotonin (Modlin et al., 2005). These neuroendocrine tumors are often differentiated according to the histological findings (Capella et al., 1995) and in a small minority of cases tumors are of high-grade histology, which, although biochemically much less active than low-

[4] Although a patient's life-span is bounded, it is useful to describe a treatment selection problem as an infinite-horizon POMDP, where the process has an exponentially decreasing but non-zero probability of continuing at each time-slice.

grade carcinoids, show much more rapid tumor progression. Therefore, carcinoid treatment that concentrates on reducing biochemical activity is not considered applicable, and more aggressive chemotherapy in the form of an etoposide and cisplatin-containing scheme is the only remaining treatment option (Moertel et al., 1991). The dynamic decision problem then becomes whether or not to administer chemotherapy at each decision moment. Our aim is to validate if the treatment strategy that is used in practice will also be found by a TLIMID as a formal domain model, thereby confirming the quality of the employed strategy.

In order to solve this problem, we have constructed a TLIMID as a model of high-grade carcinoid tumor pathophysiology in collaboration with an expert physician.[5] Figure 5.9 depicts the structure of the model, where shaded variables are observable. Since patients return to the clinic for follow-up every three months, we assume that each time-slice represents the patient status at three-month intervals, at which time treatment can be adjusted.

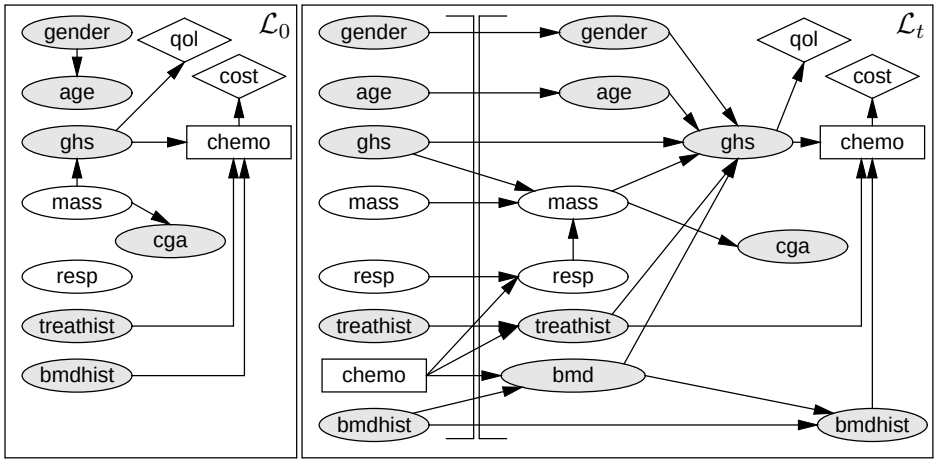


Figure 5.9: A TLIMID for high-grade carcinoid tumor pathophysiology. Chemotherapy has not yet been given at the initial time, which renders the tumor response to chemotherapy (RESP) independent of all other variables at $\mathcal{L}_0$.

In the model, the patient's *general health status* (ghs) is of central importance. In oncology, one way to estimate the general health status is by means of the *performance status* (Oken et al., 1982), which is distinguished into *normal* (0), *mild complaints* (1), *ambulatory* (2), *nursing care* (3), *intensive care* (4), and *death* (5). Modeling the evolution of ghs is a non-trivial task; it depends on the current general health status, and on patient properties such as age, and gender, since these are risk factors that may lead to patient death due to causes other than the disease. Furthermore, ghs is influenced by the tumor mass (mass) and the treatment strategy. Tumor mass has a negative influence on the general health status and is the first cause

[5]The model was developed with Hugin Developer™: `http://www.hugin.com`.

of death for patients with high-grade carcinoid tumors. Hepatic metastases normally account for the majority of the tumor mass, and the primary localization does not normally contribute significantly to the tumor mass. Tumor mass is expressed in terms of standard units, ranging from a patient with just a primary localization (mass = 0) to a patient that shows the maximal amount of metastases (mass = 16), as depicted in Fig. 5.10.

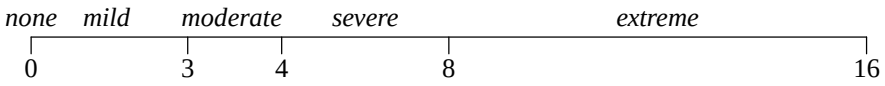


Figure 5.10: Tumor mass.

Most patients with high-grade tumors have extensive metastatic disease when admitted to the hospital, and if there is no tumor response due to treatment, then the physician estimates an exponential growth in tumor mass:

$$x(t) = x_0 \cdot \mathrm{e}^{1.41t}.$$

If there is a tumor response due to treatment then we will see a reduction in tumor mass according to Table 5.1. If no chemotherapy is given, then we use *nt* (no treatment) to denote the absence of tumor response. Finally, if dghs = *dead* then there is no change in tumor mass.

Table 5.1: The WHO criteria for tumor response.

Tumor Response	**Criteria**
Complete remission (*cr*)	Disappearance of all lesions.
Partial remission (*pr*)	More than 50% decrease in tumor mass.
Progressive disease (*pd*)	More than 25% increase in lesions, or a new lesion.
Stable disease (*sd*)	Neither pr nor pd.

Chemotherapy (chemo), with $\Omega_{\mathsf{chemo}} = \{\mathit{none}, \mathit{reduced}, \mathit{standard}\}$, is the only available treatment to reduce tumor growth, where a reduced dose is at 75% of the standard dose. We use treathist, with $\Omega_{\mathsf{treathist}} = \{0, 1, 2, 3\}$, as a memory variable to represent the patient's relevant treatment history, such that treathist $= i$ represents continued chemotherapy over the past i trimesters. Reductions in tumor mass due to chemotherapy are often described by means of the WHO criteria for tumor response (resp), as defined in Table 5.1. In (Moertel et al., 1991), given chemotherapy, 17% of patients showed complete regression, 50% showed partial regression and the remaining 33% of patients showed stable disease. Hence, a patient did not experience progressive disease if he had not been treated previously. For reduced chemotherapy we estimate that 5% of patients show complete regression, 45% show partial regression and the remaining 50% of patients show stable disease. If a patient has been

treated previously, then the effectiveness of treatment changes. In case resp$(t\text{-}1)$ is either *pr* or *cr*, then it is assumed that continued chemotherapy will lead to stable disease (*sd*). If, on the other hand, resp$(t\text{-}1) = sd$ then continued chemotherapy will become less effective. Even when chemotherapy is discontinued, we expect some residual effect of chemotherapy due to the knock-out effect on tumor-cells. It is estimated that after three months, the effect of chemotherapy is at 70% of its normal effectiveness.

Note that chemotherapy may have both positive and negative effects on general health status. Positive due to reductions in tumor mass, and negative due to severe bone-marrow depression (bmd) and damage associated with prolonged chemotherapy. Severe bone-marrow depression may cause patient death due to associated neutropenic sepsis and/or internal bleeding and it has been reported that 5 out of 45 patients experienced grade 4 leucopenia due to chemotherapy (Moertel et al., 1991). We therefore estimate that 11% of patients will experience life-threatening forms of bone-marrow depression when given standard chemotherapy. When reduced dose chemotherapy is administered, we estimate that in the order of 3% of patients will be affected. We use bmdhist, with states *no-bmd* and *bmd*, as a memory variable to represent whether or not the patient has experienced bmd in the past. No decision variable has been defined that determines whether or not to assess bmd status, since this status is assumed to be given by routine laboratory tests.

The global utility is defined as a discounted additive combination of the *quality of life* (qol) and the cost of chemotherapy (cost):

$$\mathcal{U} = \sum_{t=0}^{n} \gamma^t \left(\mathsf{qol}(t)(\mathsf{ghs}(t))) - \mathsf{cost}(t)(\mathsf{chemo}(t))))\right) .$$

Our measure of quality of life is based on *quality-adjusted life-years*, or QALYs (Weinstein and Stason, 1977), which simultaneously captures gains in quantity and quality of life (Drummond et al., 2005). QALYs are computed by multiplying a *quality-adjustment weight* for each health state by the discounted time spent in this state. We associate quality-adjustment weights with the states of ghs based on the *quality of well-being* scale (Kaplan and Anderson, 1988), taking into account that each time-slice stands for a three-month period (Table 5.2). We have associated a small economical cost with chemotherapy, that is regarded insignificant compared with the benefit gained in terms of quality of life.

Table 5.2: Quality-adjustment weights for ghs.

ghs	0	1	2	3	4	5
weight	0.214	0.184	0.168	0.121	0.109	0.000

In our model, we used a discounting factor of 0.95 as suggested in (Haddix et al., 1996), such that the three-month discount factor is $\gamma \approx 0.987$. The expected utility

then becomes:

$$\begin{aligned} \mathrm{EU}(\Delta) &= \mathrm{E}_{\Delta}\left(\sum_{t=0}^{n}\gamma^{t}\mathsf{qol}(t)(\mathsf{ghs}(t))\right) - \\ &\quad \mathrm{E}_{\Delta}\left(\sum_{t=0}^{n}\gamma^{t}\mathsf{cost}(t)(\mathsf{chemo}(t))\right). \end{aligned} \tag{5.10}$$

The first term in Eq. (5.10) is the discounted quality-adjusted life expectancy (QALE) and the second term is the discounted expected cost of treatment. The goal of our model then is to find a policy for chemotherapy that maximizes this expression.

The physician has indicated that the informational predecessors of chemo are given by ghs, treathist and bmdhist, where both treathist and bmdhist are used as memory variables within the model. Changes in treatment history are specified as follows. Given that chemo equals *standard* or *reduced*, treathist increases from x to $x+1$ until the maximum of 3 is reached, and given that chemo $=$ *none*, treathist decreases from x to $x-1$ until the minimum of 0 is reached. In order to represent whether or not a patient has ever experienced bone-marrow depression, we assume that bmdhist$(t) =$ *no-bmd* if bmd$(t) =$ *no* and bmdhist$(t$-$1) =$ *no-bmd*. Otherwise, it is assumed that bmdhist$(t) =$ *bmd*. Note that in this case, we represent memory of infinite length by restricting ourselves to the event whether or not severe bone-marrow depression has occurred. Contrary to what may be expected, cga, as a correlate of tumor mass, is not regarded to be an informational predecessor by the physician, since a patient who is known to have a high-grade carcinoid tumor is treated as often as possible, irrespective of the current state of the tumor.

5.5 Experimental results

Our aim is to find a treatment strategy for high-grade carcinoid tumors using the developed model and the described algorithms. We have applied the simulated annealing scheme, followed by SRU, as suggested in Section 5.3.[6] Since the informational predecessors are equal for chemo in $\mathcal{L}_0$ and $\mathcal{L}_t$, we assume that $\Delta_0 = \Delta_t$. We use Δ to denote this strategy, containing a stationary policy for chemo. The number of possible policies for chemo is then given by:

$$\Omega_{\mathsf{chemo}}^{\Omega_{\mathsf{ghs}}\cdot\Omega_{\mathsf{treathist}}\cdot\Omega_{\mathsf{bmdhist}}} = 3^{5\cdot 4\cdot 2} \approx 1.22\cdot 10^{19}.$$

Note that single policy updating would require an exhaustive search through this space of possible policies, which is clearly computationally intractable. For our model, we have used $\kappa = 40$ as the stopping criterion for the approximation to

[6] The proposed algorithms have been implemented using Intel's Probabilistic Networks Library (PNL): `http://www.intel.com/technology/computing/pnl`.

the expected utility, based on the observation that ten-year survival is rarely attained for this aggressive form of cancer. After some initial experiments, we have chosen $\alpha = 0.995$, $\beta = 0.5$ and $T_{\min} = 1.225 \cdot 10^{-3}$ for the simulated annealing parameters.

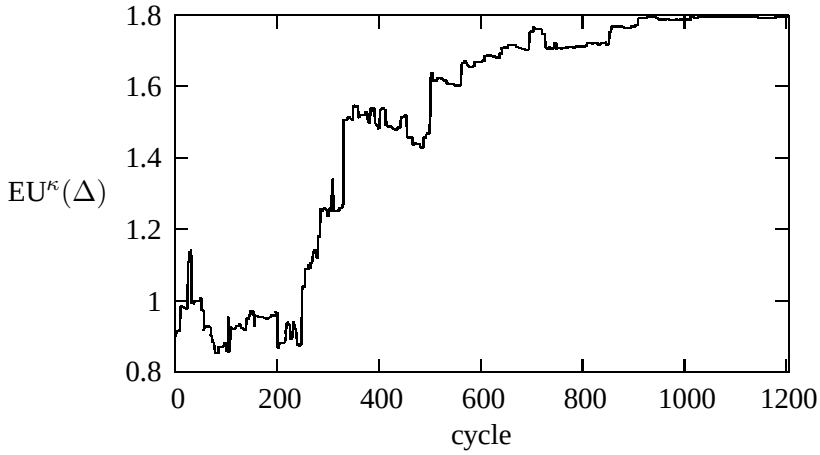


Figure 5.11: Change in EU$^{\kappa}(\Delta)$ for the treatment strategies, selected during simulated annealing, followed by single rule updating at the end.

The SA algorithm was repeated twenty times, starting from random initial strategies. It consistently found the same treatment strategy Δ, with an expected utility of 1.795. Figure 5.11 shows the subsequent values of EU$^{\kappa}(\Delta)$ of the strategies found during one of these experiments. The figure depicts how the initial explorative behavior of the simulated annealing scheme gradually changes into a hill-climbing strategy. The application of single rule updating after the simulated annealing phase caused a small increase in expected utility from 1.795 to 1.798. For this particular example, the solution found by simulated annealing (followed by SRU) was the same as the solution found by SRU alone, although this does not hold in general (cf. Example 5.1).

One way to depict the found strategy is by means of a *policy graph*, which is a finite state machine that represents state-transitions based on observations and actions associated with the nodes (Smallwood and Sondik, 1973). The policy graph for the found treatment strategy is shown in Fig. 5.12 and can be interpreted as an abstract representation of a treatment protocol. Arrows on the left-hand side of the figure depict the starting state, which depends on the initial observations. Each state has an associated action and the next state is chosen based on the next observation. The protocol states that we treat once if the health status is good enough ($\mathsf{ghs} \leq 3$), where patients with severe bmd receive reduced chemotherapy. Then we wait to let the patient recover and treat again, depending on whether or not the general health status is good enough. According to the expert physician, the found strategy was in

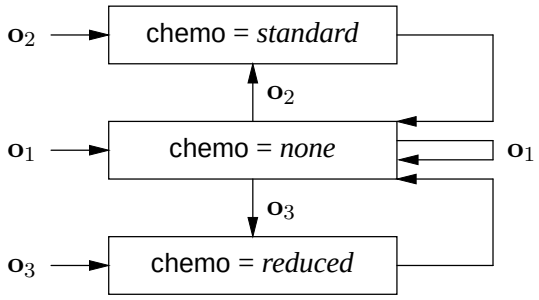


Figure 5.12: Policy graph for the best strategy that was found by simulated annealing, where $\mathbf{o}_1$ = ghs > 3, $\mathbf{o}_2$ = ghs $\leq 3 \wedge$ bmdhist $=$ *no-bmd* and $\mathbf{o}_3$ = ghs $\leq 3 \wedge$ bmdhist $=$ *bmd*.

agreement with the treatment protocol that is used in clinical practice, even though in exceptional cases, patients are given chemotherapy for more than three consecutive months. Hence, our formal domain model validates the treatment protocol that physicians use to treat carcinoid patients.

5.6 Summary

We have defined DLIMIDs, represented as TLIMIDs, as a framework for dynamic decision-making under uncertainty and used them as the basis for a pathophysiological model for high-grade carcinoid patients. Although the repetitive structure of a TLIMID has been used implicitly in (Lauritzen and Nilsson, 2001), the explicit use of a TLIMID and its transformation into a pair of Bayesian network fragments allows for the representation of infinite-horizon POMDPs. This benefit comes at the expense of using strategies that may suffer from perceptual aliasing, which we resolve by means of memory variables that represent part of the observed history.

We have demonstrated that reasonable strategies can be found for infinite-horizon DLIMIDs, where both SRU and SA do not suffer from the intractability of SPU when the number of informational predecessors increases. The approach does require that good strategies can be found using a limited amount of memory, since otherwise, found strategies will fail to approximate the optimal strategy. This requirement should hold especially between time-slices, since the state-space of memory variables can become prohibitively large when a large part of the observed history is required for optimal decision-making. Although this restricts the types of decision problems that can be managed, DLIMIDs, as constructed from a TLIMID, allow the representation of large or even infinite-horizon decision problems that cannot be managed by standard influence diagrams.

Our approach is particularly useful in the case of problems that cannot be properly approximated by a short number of time-slices, which was shown for a toy problem

in (van Gerven and Díez, 2006). Application of the theory to the selection of a treatment strategy for high-grade carcinoid tumors has demonstrated the usefulness of our approach for real-world medical problems. The theory provides a formal basis for the validation of existing treatment strategies and may actually be used to modify existing treatment strategies when more optimal solutions are found.

Chapter 6

A Probabilistic Model for Carcinoid Prognosis

An important task in clinical patient management is to determine a prognosis for a patient that suffers from a disease, where prognosis is defined as: *the prediction of the future course of a disease process conditional on patient history and a projected treatment strategy*. This prediction is non-trivial since the physician often has incomplete information and treatment itself can have a multitude of uncertain effects. As a result, predictions made by the physician can be poor (Lee et al., 1986; Knaus et al., 1991b; Christakis and Lamont, 2000) or miscalibrated (Glare et al., 2003). Therefore, patient management can benefit greatly from the development of prognostic models that aid the physician in this task. Next to its use in clinical decision making, prognostic models can also be of value to the patient (notification, quality-of-life decisions), as well as to the policy-maker (comparative audit, patient selection for clinical trials, development of treatment protocols) (Wyatt and Altman, 1995; Abu-Hanna and Lucas, 2001).

Various approaches to develop a prognostic model exist. Traditionally, a prognostic model consists of simple decision rules that are based on a prognostic score and classify patients into different risk categories (Mazumdar and Glassman, 2000). Such scores are often based on clinical variables, and have been constructed for the general patient population (Knaus et al., 1991a; Le Gall et al., 1993) as well as for specific patient subgroups (Schuchter et al., 1996; Groeger et al., 1998). Survival analysis takes a different approach, and models survival rate by taking into account patient-specific covariates, such as by means of the proportional hazards model (Cox, 1972; Cox and Oakes, 1984; Collett, 2003). In decision analysis, stochastic processes which evolve over time, known as Markov decision processes, are used as the basis for prognostic models (Beck and Pauker, 1983; Sonnenberg and Beck, 1993). More recently, techniques such as decision-trees, neural networks, support vector machines, and Bayesian networks, as developed by the artificial intelligence community, have become popular as prognostic models (Cruz and Wishart, 2006; Delen et al., 2005;

This chapter is based on (van Gerven and Taal, 2006; van Gerven et al., 2007b).

Ohno-Machado, 1997; Abu-Hanna and Lucas, 2001).

The above techniques have all proven their worth as prognostic models in medicine, but they are not always applicable. Although sophisticated techniques, such as neural networks and support vector machines, generally improve upon the performance of simple decision rules, they also require the availability of large amounts of high-quality data. Unfortunately, this data is not always available, which renders the methods inapplicable. Another perceived deficiency is the fact that most of the described techniques do not provide insight into *how* a certain prognostic conclusion is reached; they are so-called *black-box* models, which is an undesirable property of clinical decision support systems (Hart and Wyatt, 1990). For instance, even though the proportional hazards model has an interpretation in terms of the patient-specific covariates that modulate patient hazard, the model cannot give a causal explanation of how the covariates interact and influence patient survival.

Bayesian networks (Pearl, 1988) do allow for an interpretation in terms of causes and effects, and have the additional benefit that they can be constructed from available expert knowledge. If a Bayesian network incorporates time, then it is known as a *dynamic* Bayesian network (DBN), and if it includes decision making, as is often needed for accurate prognostication (Hilden and Habbema, 1987), then it can be regarded as a factorized representation of a partially-observable Markov decision process. The usefulness of such a representation for clinical patient management has already been discussed in (Peek, 1999), but, to date, there are only few systems for clinical patient management that were built using this approach (Hauskrecht and Fraser, 2000; Charitos et al., 2005). As will be shown, representation of a prognostic model in terms of a DBN is beneficial, since they allow for a causal explanation, can be constructed from data and/or expert knowledge, and allow for flexible query answering. However, DBNs constructed from expert knowledge are difficult to develop, which is thought to be one of the main reasons for their limited use at present.

In this chapter, our aim is to describe the construction and validation of a DBN for prognosis of patients that present with low-grade carcinoid tumors; a neuroendocrine tumor that displays a complex symptomatology. This is a difficult task, since the domain requires the incorporation of decision-making and the representation of temporal interactions. We proceed by describing the clinical problem, carcinoid pathophysiology, and carcinoid treatment in Section 6.1. Section 6.2 describes the prognostic model, which we call, henceforth, the carcinoid model. The carcinoid model is validated in Section 6.3 by means of a database that has been collected at the Netherlands Cancer Institute (NKI). In order to obtain insight into the quality of the model, we use a number of techniques, where we focus not only on prognostic accuracy, but also on the intelligibility of the prognostic conclusions. We end with a discussion of the results in Section 6.4.

6.1 Prognosis of carcinoid tumors

6.1.1 Problem description

Low-grade carcinoid tumors are a type of neuroendocrine tumor that can produce high levels of serotonin, kinins, prostaglandins, and other vasoactive peptides. They are most commonly found in the midgut (Taal and Smits, 2005) and typically behave less aggressively than conventional adenocarcinomas (van Eeden et al., 2002). During the early stages, carcinoid tumors often remain undiagnosed, where vague abdominal pain is commonly ascribed to irritable bowel or spastic colon (Bast-Jr et al., 2000). Progressive carcinoid disease is often accompanied by the *carcinoid syndrome*. This syndrome is mainly characterized by diarrhea caused by increased bowel motility due to serotonin overproduction (Öberg et al., 1987), periodical flushing attacks due to the synergistic interaction between histamine, kinins, and prostaglandin released by the tumor into the general circulation, and less frequently wheezing (Zuetenhorst et al., 1999). Extreme cases of the carcinoid syndrome are known as a *carcinoid crisis*, which may lead to cardiovascular collapse and ultimately death. Often, only if symptoms of the carcinoid syndrome are present, a carcinoid tumor is suspected and the patient is sent to the hospital. Since the clinical department of the *Netherlands Cancer Institute* (NKI) acts as a referral centre, most patients that are admitted are already diagnosed to have carcinoid disease, most often of the midgut type. Hence, for physicians at the NKI, diagnosis of carcinoids is not of primary concern. However, due to the complex nature of carcinoid disease, and recent advances in carcinoid treatment, the need for appropriate prognostication has increased.

6.1.2 Pathophysiology of carcinoid tumors

The midgut is the region in which carcinoids are predominantly found, and neuroendocrine tumors that derive from other sites often show markedly different behavior and hence need alternative models for prognostication (Zuetenhorst and Taal, 2005). Carcinoid tumor histology is determined by mitotic activity and tissue necrosis, and distinguished into well differentiated, or *low-grade* malignancies, and poorly differentiated, or *high-grade* malignancies (Capella et al., 1995). A minority of patients presents with high-grade tumors, which grow faster but are biochemically less active, and therefore require a different prognostic model. We restrict ourselves to carcinoids of the midgut with a low-grade histology.

As mentioned, the most prominent clinical sign of carcinoid disease is the carcinoid syndrome, which is caused by high levels of circulating bioactive substances (Zuetenhorst et al., 1999). Although many of these substances are thought to play a role in the disease, the exact interactions are as yet unclear, and in practice, diagnosis relies on the assessment of serotonin overproduction by measuring urinary 5-hydroxyindole-3-acetic acid (5-HIAA) levels, which we distinguish into *normal*,

elevated, and *extreme*. Serotonin overproduction is caused by the carcinoid tumor in the presence of particular metastases. Hormones released by carcinoid tumors are often destroyed by the liver before they reach the general circulation to cause symptoms, and therefore, only liver metastases or metastases that release hormones directly into the general circulation such as gonad (ovary or testes) or lung metastases, can produce the carcinoid syndrome. Most of the hormone-producing tumor-mass is accounted for by the liver, and consequently carcinoids are often accompanied by widespread hepatic metastases. Plasma *chromogranin A* (CgA) levels can be used as a marker of tumor load, in terms of neuroendocrine activity (Nobels et al., 1998) and tumor mass (D'Herbomez and Gouze, 2002). We distinguish *normal*, *elevated*, and *extreme* CgA levels, and patients with extreme CgA levels have a significantly poorer 5-year survival than patients with elevated CgA levels (Janson and Öberg, 1996). The production of CgA and serotonin is determined by tumor activity and tumor extensiveness.

Sometimes, excessive release of bioactive substances leads to a carcinoid crisis, which is characterized by severe flushing, severe diarrhea, and vomiting. A crisis may lead to dehydration, acute hypotension and may ultimately cause cardiovascular collapse, which is a life-threatening situation. It is thought to arise from an excessive release of vasoactive substances into the general circulation (Sutton et al., 2003). Serotonin is known to cause diarrhea and is used as a correlate of the vasoactive substances that cause flushing. Which substances are exactly responsible for flushing remains unclear.

A major complication of carcinoid tumors is carcinoid heart disease (CHD), which is a consequence of enlargement and distortion of the endocardium and subendocardium of the tricuspid valve, leading to tricuspid insufficience and decompensatio cordis. CHD may lead to right heart failure which is the cause of death in approximately half of carcinoid patients (Taal et al., 1999); as the pump function of the heart deteriorates the patient's health deteriorates rapidly. A trend can be seen between the degree of right atrium dilatation, and the level of the *brain natriuretic peptide* (BNP); especially its biologically inactive N-terminal fragment NT-pro-BNP (Zuetenhorst et al., 2004). This level is distinguished into *normal* and *elevated*. CHD-related mortality is dependent on the progression of CHD in the patient, which is defined as tricuspid valve thickening with additional severe or extreme regurgitation. Mesenterial fibrosis is another major complication of carcinoid tumors, where small-bowel tumors cause shrinkage and fibrosis of the mesentery, leading to bowel obstruction and/or ischaemia, with finally necrosis and perforation of the bowel wall, which is frequently accompanied by acute abdominal pain (Modlin et al., 2004).

6.1.3 Treatment of carcinoid tumors

Treatment is distinguished into *interventions* and *systemic treatments*. We disregard symptomatic treatment since this does not influence disease progression.

Interventions

Treatment of a carcinoid tumor often amounts to surgical intervention, and can be either curative or palliative. Although curative surgical removal of the primary tumor is the treatment of choice for small localized tumors, it is almost impossible in the presence of intra-abdominal or hepatic metastases. In the context of the prognostic model, it is assumed that the patient has already received appropriate primary tumor surgery. The remaining applicable interventions are shown in Table 6.1. These interventions also present a risk to the patient since they cause patient death in a minority of cases.

Table 6.1: Interventions for carcinoid tumors.

Intervention	Usage
bowel resection	Performed in case of severe mesenterial fibrosis.
cardiac surgery	Performed in case of carcinoid heart disease.
partial liver resection	Treatment of mild liver metastases.
radiofrequency ablation	Treatment of moderate liver metastases.
embolization	Treatment of severe liver metastases.

Only when the primary tumor leads to mesenterial fibrosis, treatment in the form of bowel resection becomes necessary (Sutton et al., 2003). Bowel resection is a preventive palliative treatment that is performed whenever the patient experiences a curable form of mesenterial fibrosis given an acceptable health status.

If a patient suffers from carcinoid heart disease then, given that the patient has an acceptable health status, cardiac surgery is performed. This normally amounts to tricuspid valve replacement, reducing tricuspid valve thickening and regurgitation. Unfortunately, cardiac surgery has a relatively high associated mortality rate.

In case of hepatic metastases one may opt for one of the hepatic treatments: *partial liver resection* (PLR), *radiofrequency ablation* (RFA), or *hepatic artery embolization* (Meij et al., 2005). Hepatic metastases are operable only if there are no more than three localized metastatic regions, and, according to the physician, PLR can be administered at most two times, given an acceptable health status. If PLR fails, then hepatic metastases may be treated by RFA when there are no more than six metastatic localizations, each region being less than 4 cm in diameter. RFA heats tumors and thereby kills the cancer cells. The procedure has a low complication rate, can be performed without major open surgery, only involves overnight hospitalization, and can be administered at most three times, given an acceptable health status. Embolization is a method to treat diffuse carcinoid localizations in the liver. Selective embolization leads to occlusion of the liver artery, cutting off blood supply to the tumor, depriving it of oxygen and nutrients. Embolization of the liver arteries leads to the post-embolization syndrome, which is characterized by temporary fever and pain, and may cause life-threatening complications (Eriksson et al., 1998; Meij et al., 2005). According to the physician, embolization can be administered at most two

times, and is performed only in case of diffuse hepatic metastases, when systemic treatment has failed and given an acceptable health status. The effect of hepatic treatment is a reduction in hepatic tumor mass, and can be interpreted in terms of the tumor response, as depicted in Table 6.2.

Table 6.2: The criteria for tumor response.

Tumor Response	**Criteria**
Complete remission (*cr*)	Disappearance of all lesions.
Partial remission (*pr*)	$> 50\%$ decrease in tumor mass.
Progressive disease (*pd*)	$> 25\%$ increase in lesions, or appearance of a new lesion.
Stable disease (*sd*)	Neither pr nor pd.

Systemic treatment

Systemic treatment focuses on reducing overall tumor activity and tumor growth, and can be distinguished into the treatments shown in Table 6.3. Systemic treatment is administered in case of biochemically active metastases in conjunction with extreme 5-HIAA levels and/or both severe diarrhea and flushing. We call these conditions the *systemic conditions*.

Table 6.3: Systemic treatment of carcinoid tumors.

Systemic treatment	**Description**
farmacological somatostatin	Synthetic forms of native somatostatin.
interferon	Synthetic form of an immune system stimulant.
radiolabeled somatostatin	Radioactive somatostatin used for autoradiation therapy.
radiolabeled MIBG	Radioactive MIBG used for autoradiation therapy.
farmacological MIBG	Inhibitor of mitochondrial respiration.

Reductions in tumor growth are captured by the tumor response of Table 6.2, whereas reductions in tumor activity are captured by the biochemical response, as quantified by means of the criteria in Table 6.4. In general, systemic treatment is characterized by positive effects (such as tumor reduction and reduction of biological activity), and possible side-effects, such as bone-marrow depression.

Table 6.4: The criteria for biochemical response.

Biochemical response	**Criteria**
Complete remission (*cr*)	Normal biochemical activity.
Partial remission (*pr*)	$> 50\%$ decrease in biochemical activity.
Progressive disease (*pd*)	No treatment response.
Stable disease (*sd*)	$< 50\%$ decrease or $< 25\%$ increase in biochemical activity.

Somatostatin is a peptide that has widespread inhibitory effects and leads to a reduction in the release and production of serotonin and vasoactive substances by the tumor. It binds to the somatostatin receptors which are expressed on more than 80% of the carcinoid tumors. Native somatostatin has limited use by its short half life, but a number of longer acting somatostatin analogues have been developed. Octreotide (Lamberts et al., 1996) is a somatostatin analogue, that often induces symptomatic improvement, although this is not always accompanied by a reduction in 5-HIAA excretion. Somatostatin analogues have been reported to inhibit tumor growth, but a reduction in tumor volume is seldom observed (Zuetenhorst and Taal, 2005). We refer to this form of medication as *farmacological somatostatin* (f-soma). Farmacological somatostatin may induce increased bowel motility, and over time, farmacological somatostatin efficacy decreases due to somatostatin receptor down-regulation. A tracer dose of radiolabeled octreotide is used to detect somatostatin receptors by means of a so-called *octreoscan*, and in order to treat with f-soma, the octreoscan must be positive. Once started, we increase the dosage when the disease becomes progressive despite treatment, until the highest dosage is reached and maintained.

Interferon-α (ifn) is a synthetic copy of a substance that is produced naturally by monocyte/ macrophages and is considered after failure of f-soma treatment. Due to binding of ifn to interferon receptors a complex series of signal transduction events takes place, resulting in the production of a multitude of proteins with different actions. ifn works directly on cancer cells by interfering with cells growth and multiplication, and stimulates the immune system, by encouraging killer T cells and other cells that attack cancer cells (Öberg and Eriksson, 1991). Side-effects amount to flu-like symptoms, diarrhea, general sickness, tiredness, loss of appetite, and a temporary drop in bone marrow functioning. Due to these side-effects, ifn is administered for at most a year. Note that the health status should be acceptable, and bone-marrow depression must be absent, in order to give treatment.

Once interferon treatment has failed, we may use either ^{177}Lu-labeled Octreotide (Lutetium), or ^{131}I-labeled MIBG, to invoke autoradiation. Meta-Iodobenzylguanidin (MIBG) resembles noradrenalin and serotonin, and it is taken up in the carcinoid tumor cells and stored in the neurosecretory granules. We refer to the respective treatments as *radiolabeled somatostatin* (r-soma) and *radiolabeled MIBG* (r-mibg) treatment. R-soma has a strong tumor reducing effect, and is only administered once in a series of four treatments with two month intervals. Observed toxicities of r-soma autoradiation therapy are nausea and vomiting, haematological toxicity and renal function impairment (Zuetenhorst and Taal, 2005), and therefore, a good renal function is required, and bone-marrow may not be severely depressed. Renal failure may arise due to various causes such as medication, vascular obstruction, or hypertension.

Radiolabeled MIBG is also used for scanning purposes (*mibgscan*), and a positive mibgscan is a prerequisite for r-mibg treatment. Predosing with f-mibg leads to improved tumor targeting of r-mibg since f-mibg has the capacity to render a negative

mibgscan positive. Radiolabeled MIBG treatment consists of two 200 mCi dosages within a six to eight week interval, and can be administered at most twice due to radiation damage, leading to severe bone-marrow depression in a minority of cases.

Farmacological MIBG (f-mibg) is administered when other treatments have failed. The cytotoxic effect of f-mibg is related to inhibition of mitochondrial respiration, resulting in enhanced glucose consumption, increased lactic acid production, inhibition of oxygen consumption and decreased adenosine triphosphate levels (Zuetenhorst et al., 1999). F-mibg treatment requires a normal blood pressure since it induces changes in blood pressure (Zuetenhorst et al., 1999). The treatment strategy for f-mibg is to treat for three months, to stop for six months and then to repeat treatment if previous results were positive.

6.2 Structure of the carcinoid model

We proceed with a description of the architecture of the carcinoid model, which is specified in terms of a dynamic Bayesian network.

6.2.1 Dynamic Bayesian networks

A *Bayesian network* $\mathcal{B} = (G, P)$ is a pair where G is an *acyclic directed graph*, with nodes corresponding to a set of random variables $\mathbf{X}$, and P is a joint probability distribution (JPD) of variables in $\mathbf{X}$, which factorizes as:

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X \mid \pi(X))$$

where $\pi(X)$ denotes the parents of X in G. The representation of a JPD by a Bayesian network generally reduces the number of parameters that need to be estimated and allows for efficient probabilistic inference. In case we are dealing with problems of a temporal nature, we explicitly include time within a Bayesian network, by reasoning over random processes $X = \{X(t) \colon t \in T\}$ instead of random variables. The resulting model is known as a *dynamic Bayesian network*, and if it is assumed that the Markov property holds, which states that the future is independent of the past, given the present, we obtain the following factorization:

$$P(\mathbf{X}) = \prod_{t \in T} \prod_{X(t) \in \mathbf{X}(t)} P(X(t) \mid \pi(X(t)))$$

with $\mathbf{X}(t) = \{X(t) \colon X \in \mathbf{X}\}$.

In this work, we will focus on discrete-time and discrete-space random processes, which implies that $T \subseteq \mathbb{N}$ and $P(\cdot \mid \cdot)$ can be specified by a finite look-up table. If the structure of the dynamic Bayesian network is invariant for all times $t \in \{1, 2, \ldots\}$ then it can be specified in terms of:

- a *prior model* $P(\mathbf{X}(0))$, specifying the initial distribution of the joint process, and
- a *transition model* $P(\mathbf{X}(t) \mid \pi(\mathbf{X}(t)))$, specifying how the process evolves as we go from time t to time $t+1$ for $t \in \{1, 2, \ldots\}$.

In the following, we describe the structure of the carcinoid model, focusing first on pathophysiology (Section 6.2.2), and second on treatment (Section 6.2.3). The prior model and transition model together consist of 218 variables and 74 342 CPT entries. In order to compute distributions of interest, we use the exact junction tree algorithm (Lauritzen and Spiegelhalter, 1988) and approximate particle filtering (Doucet et al., 2001) where appropriate. For a more complete description of the model and its required parameter estimates, we refer to (van Gerven and Taal, 2006).

6.2.2 Architecture of the pathophysiological component

The shaded nodes in Fig. A.2 in Appendix A are an abstract representation of carcinoid tumor pathophysiology as it is embodied in the carcinoid model. It is depicted how health is influenced by carcinoid disease, through the tumor, its biochemistry, and its major complications of carcinoid crisis, carcinoid heart disease, and mesenterial fibrosis in the bowel. Observable symptoms arise due to the biochemistry, bowel problems, and tumor progression. Furthermore, it is shown that health is influenced by patient specific risk factors.

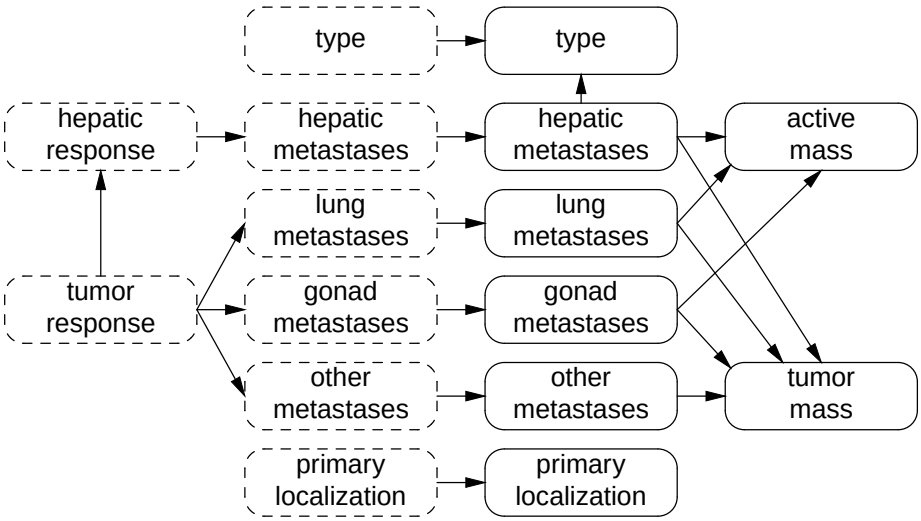


Figure 6.1: Representation of tumor progression.

Figure 6.1 depicts the progression of the tumor in detail. As described, the tumor may lead to various metastases, some of which may be biochemically active.

Furthermore, hepatic metastases are distinguished into different types, since hepatic treatment depends on this. The variable hepatic response captures the effect of hepatic treatment, whereas the variable tumor response captures the tumor effect of systemic treatment. We only represent primary tumor localization and not its size, since disease progression is mainly determined by metastatic disease. Active metastases and total metastatic tumor mass are represented by variables active mass and tumor mass. Parameter estimates are based on clinical expertise and Ref. (Skinazi et al., 1996).

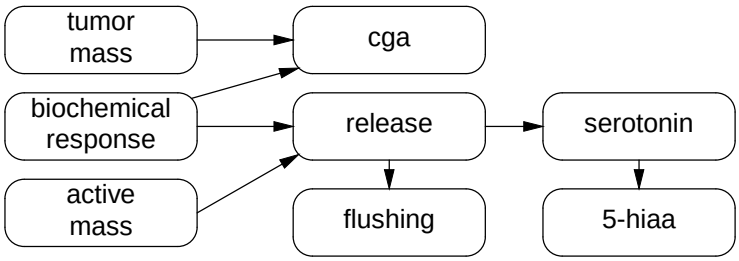


Figure 6.2: Representation of tumor biochemistry.

Carcinoid tumor biochemistry is captured in Fig. 6.2. Here, it is shown that all metastases determine CgA production, whereas biochemically active metastases determine the release of various biochemical compounds. One of these compounds is serotonin, whose product 5-HIAA can be measured in a urine sample. Note that the release of CgA and other biochemical compounds is influenced by the biochemical response of systemic treatment. Parameter estimates are based on clinical expertise and Ref. (Nehar et al., 2004). The release of biochemical compounds may in severe cases lead to a carcinoid crisis through a cascade of events. Since our interest is not in modeling this cascade, we simply capture this by assuming a dependence between release at time t and crisis at time $t + 1$, where parameters were estimated by an expert physician.

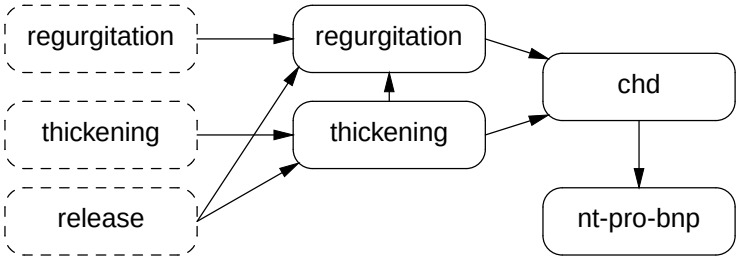


Figure 6.3: Representation of carcinoid heart disease.

The release of biochemical compounds is also responsible for the development of carcinoid heart disease, as shown in Fig. 6.3. Note that thickening is a prerequisite

for regurgitation, and CHD is defined in terms of both as follows:

$$P(\mathsf{chd} \mid \mathsf{thickening}, \mathsf{regurgitation}) = 1_{\mathsf{thickening}=\mathit{yes} \wedge \mathsf{regurgitation}>\mathit{moderate}}$$

where 1_X is the indicator function, which is equal to one if X evaluates to *true*, and equal to zero if X evaluates to *false*. Even though CHD is fully determined by thickening and regurgitation, it is still useful to represent the variable CHD in the model, as it facilitates subsequent parameter estimation. For example, NT-pro-BNP concentrations are normally expressed conditional on the absence or presence of CHD (e.g., (Zuetenhorst et al., 2004)).

Bowel-related problems are another complication of carcinoid tumors (Fig. 6.4). Mesenterial fibrosis is induced by biochemically active small-bowel primary tumors, and may lead to ischaemia and/or obstruction. Abdominal pain is a symptom of these complications, but may also be caused by other metastases or increased bowel motility, for instance due to serotonin overproduction. Increased bowel motility leads to diarrhea, and cessation of diarrhea may be experienced in case of bowel obstruction. Parameter estimates are based on clinical expertise and Ref. (Taal and Visser, 2004).

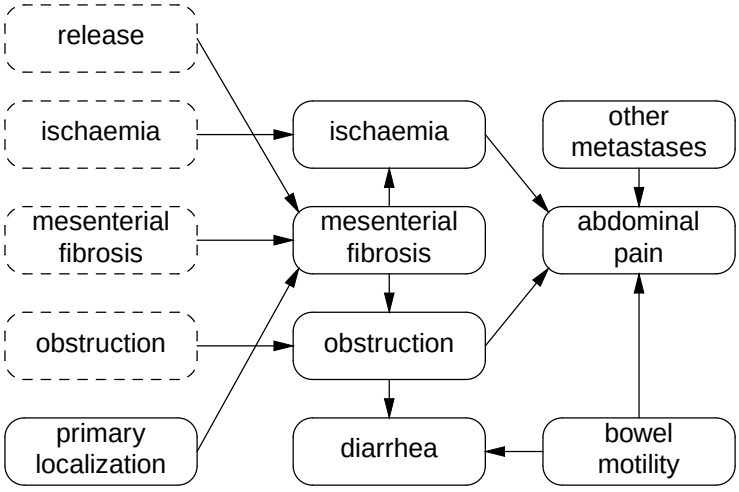


Figure 6.4: Representation of bowel-related problems.

The core part of the carcinoid model is formed by modeling how a patient's health status is influenced by the disease, by risk factors independent of the disease, and by possible treatment complications. In oncology, one way to represent the patient's health status is in terms of the *performance status* (Oken et al., 1982), which is distinguished into *normal* (0), *mild complaints* (1), *ambulatory* (2), *nursing care* (3), *intensive care* (4), and *death* (5), where we say that the health status is acceptable if $\mathsf{health} < 3$. Figure 6.5 depicts the influences on patient health, where the variables age and gender are major risk factors that determine patient death independent of the disease. Their influence has been estimated from demographic data collected by

the *Central Bureau of Statistics* for the period 2000–2004 (Centraal Bureau voor de Statistiek, 2005).

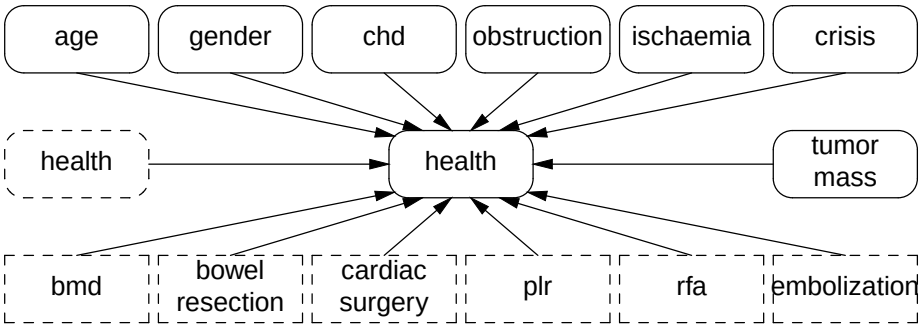


Figure 6.5: Representation of patient health.

The large number of conditioning variables that (partially) determine patient health, makes estimation of conditional probabilities for this variable very difficult. However, a large subset of these variables are risk factors that influence health due to the fact that they may cause immediate patient death. If we let the variable endurance with $\Omega_{\text{endurance}} = \{yes, no\}$ stand for survival of such risk factors, then we obtain a much simpler model, as given in Fig. 6.6.

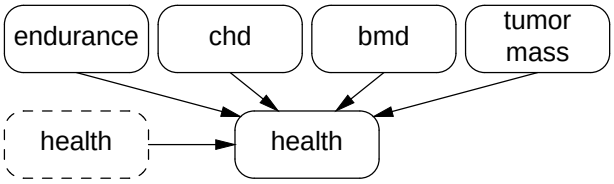


Figure 6.6: A simplified representation of patient health.

The structure which we associate with the variable endurance can be interpreted as a causal interaction model (Meek and Heckerman, 1997). Here, endurance is indirectly influenced by risk factors $C_i \in \mathbf{C}$ through intermediate variables $X_i \in \mathbf{X}$, as modulated by patient health. The influences are then combined by a logical OR, and we obtain

$$P(\text{endurance}=\textit{true} \mid \mathbf{C}, \text{health}(t\text{-}1)) \;=\; 1 - \prod_{C_i \in \mathbf{C}} P(X_i = \textit{false} \mid C_i, \text{health}(t\text{-}1)) \tag{6.1}$$

with causes $\mathbf{C} = \{\{\text{age, gender}\}, \text{obstruction, ischaemia, crisis, bowel resection, cardiac surgery, plr, rfa, embolization }\}$, where age and gender condition the same intermediate variable, as they together quantify death risk in the general population.

As can be seen in Eq. (6.1), the variable health also plays an important role if our interest is in computing posteriors for other variables. We have found it useful to use

the notion of a *default influence*, in order to facilitate the estimation of how health influences various risk factors. Consider for instance the influence of patient health on the risk of dying from a carcinoid crisis: $P(\text{crisis-death}(t) = \textit{yes} \mid \text{crisis}(t\text{-}1) = \textit{yes}, \text{health}(t\text{-}1) = g)$. Since $\Omega_{\text{health}} = \{0, \ldots, 5\}$, we need to estimate six different probability values. It is assumed that health has a default influence on the various risk factors, which is accomplished by assuming that the influence of health on a risk factor x can be written as:

$$\frac{P(\textit{yes} \mid \textit{yes}, g)}{P(\textit{no} \mid \textit{yes}, g)} = \frac{P(\textit{yes} \mid \textit{yes}, 1)}{P(\textit{no} \mid \textit{yes}, 1)} \cdot \theta_{\text{health}}(g) \tag{6.2}$$

where

$$\theta_{\text{health}}(g) = \frac{P(\textit{yes} \mid \textit{yes}, g)}{P(\textit{no} \mid \textit{yes}, 1)}$$

represents the change in the odds for x $\text{death}(t) = \textit{yes}$ given a change in health. This change is estimated by the physician as

$$\theta_{\text{health}} = \{(0, 0.99), (1, 1), (2, 1.75), (3, 10), (4, 100), (5, 0)\},$$

where the choice of 0 for $\text{health} = 5$ represents the fact that a risk factor has no influence whenever the patient is already dead. This use of default influences of health leads to a six-fold decrease in the number of probabilities that need to be specified for variables that are conditioned on health, since we can use Eq. (6.2) to compute probabilities for $\text{health} \neq 1$. In the following, conditioning of risk factors by patient health is left implicit.

6.2.3 Architecture of the treatment component

A prognostic model also requires the representation of decisions and their outcomes. For each treatment, we need to specify its negative and positive effects, and the treatment protocol; i.e., under which conditions the various treatments are applied.

Treatment effects

Negative treatment effects have already been shown in Figs. 6.5 and 6.6, where bone-marrow depression (bmd) may be caused by ifn, r-soma, or r-mibg treatment. Positive treatment effects are modeled as follows. The intervention cardiac surgery$(t\text{-}1)$ simply conditions tricuspid valve thickening(t), where it is assumed that $\text{thickening}(t) = \textit{absent}$ given that $\text{cardiac surgery}(t\text{-}1) = \textit{yes}$. The intervention bowel resection$(t\text{-}1)$ conditions mesenterial fibrosis(t), where it is assumed that $\text{mesenterial fibrosis}(t) = \textit{absent}$ given that $\text{bowel resection}(t\text{-}1) = \textit{yes}$.

Hepatic treatments influence the hepatic metastases through the hepatic response(t), which represents the combined effect of all hepatic treatments. An example is given in Fig. 6.7, which models the positive effect of an arbitrary hepatic

treatment. Note that the effect of hepatic treatment is modulated by the metastatic type, which can be *localized, multiple,* or *diffuse*. The total hepatic response can be modeled by means of the following causal interaction model:

$$P(\text{hepatic response} = e \mid \mathbf{C}, \text{type}(t\text{-}1)) = \sum_{\mathbf{x}:\ \max(\mathbf{x})=e} \prod_{C_i \in \mathbf{C}} P(x_i \mid C_i, \text{type}(t\text{-}1)) \tag{6.3}$$

with $\mathbf{C} = \{\text{plr}(t\text{-}1), \text{rfa}(t\text{-}1), \text{embolization}(t\text{-}1)\}$. States of hepatic response are ordered: *progressive disease* $\prec$ *stable disease* $\prec$ *partial response* $\prec$ *complete response*.

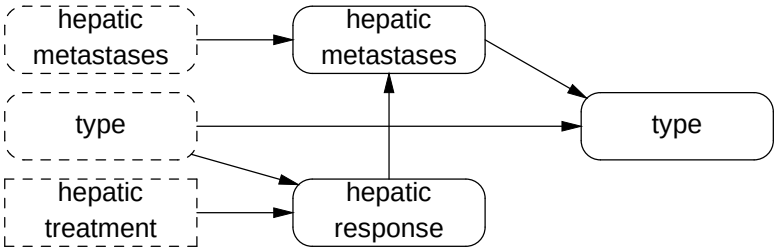


Figure 6.7: The positive effect of hepatic treatment on hepatic metastases.

Figure 6.8 depicts the tumor and biochemical response of the various systemic treatments. The effect of some of these treatments is modulated by other variables (not shown).

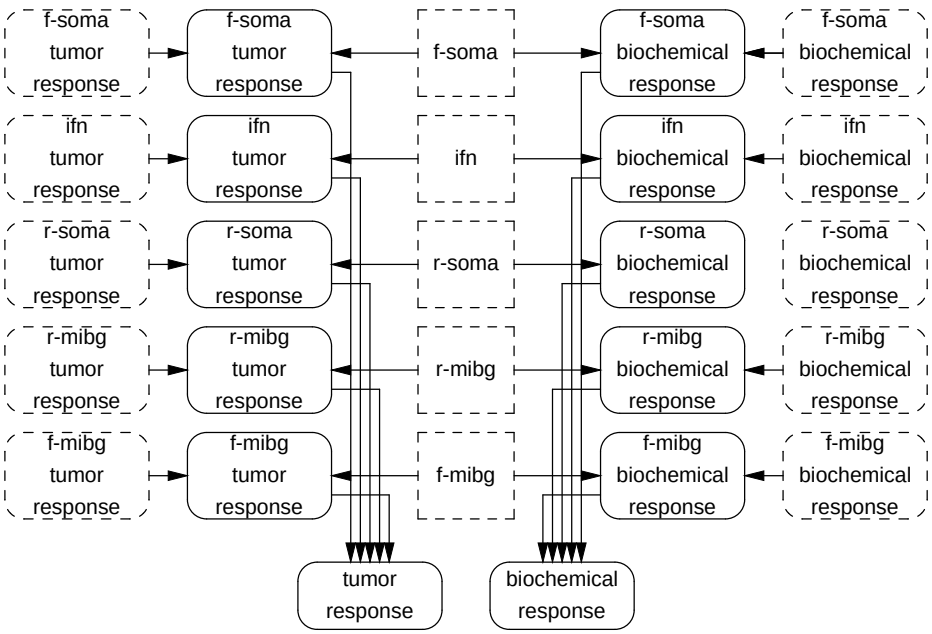


Figure 6.8: The tumor and biochemical response of systemic treatment.

For f-soma, we condition tumor and biochemical response on a variable increase, which captures if the f-soma dosage has recently been increased, since an increase in dosage induces a stronger response. The tumor and biochemical response of both f-soma and r-soma is modulated by the octreoscan (since these treatments have no effect in case the octreoscan is negative). Similarly, the tumor and biochemical response of r-mibg is conditioned by the mibgscan. The combined effects can again be modeled by means of a causal interaction model, similar to that of Eq. (6.3), where the current responses are also modulated by the previous responses. The positive effect of tumor response(t) and biochemical response(t) has already been shown in Figs. 6.1 and 6.2, with states as given by Tables 6.2 and 6.4.

Treatment protocol

The protocol for the various treatments was mentioned in passing in Section 6.1.3. Bowel resection is applied in case of curable mesenterial fibrosis and/or obstruction due to other causes, together with an acceptable health status (Fig. 6.9).

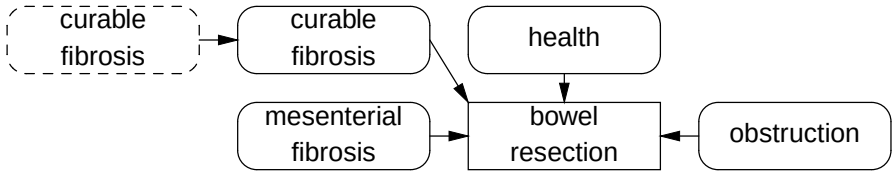


Figure 6.9: The treatment strategy for bowel resection.

A similar situation holds for cardiac surgery, where we treat in case of carcinoid heart disease given an acceptable health status (Fig. 6.10).

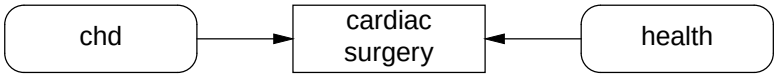


Figure 6.10: The treatment strategy for cardiac surgery.

For the hepatic treatments, the strategy is determined by the extensiveness of hepatic metastases, the type of hepatic metastases, health status, and the history of treatment. Additionally, for embolization, we require that autoradiation treatments have failed (Fig. 6.11). For the systemic treatments, the strategies are more complex. Consider for instance the treatment strategy for f-soma (Fig. 6.12). The figure depicts that the systemic conditions must be present and the octreoscan must be positive in order to administer f-soma. It is also shown that if biochemical and tumor responses are absent despite f-soma treatment, then there is tumor progression despite treatment (f-soma progression). This progression determines whether f-soma treatment dosage is increased, or whether f-soma treatment fails. Finally, if the patient comes in with severe or extreme amounts of tumor mass, then the patient receives f-soma, possibly together with other systemic treatment.

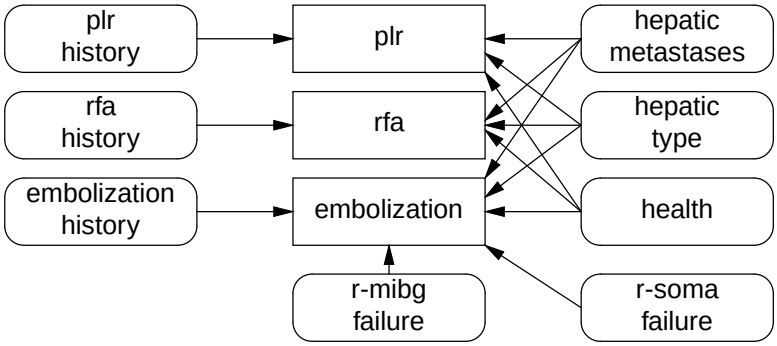


Figure 6.11: The treatment strategies for plr, rfa, and embolization.

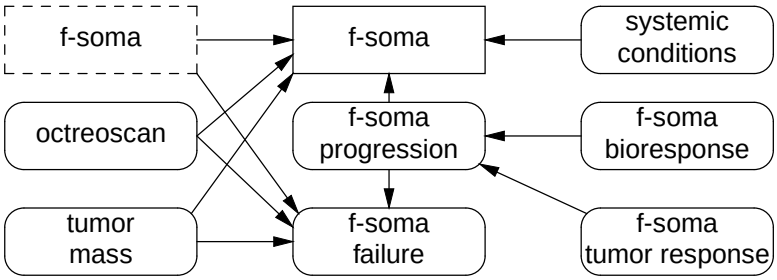


Figure 6.12: The treatment strategy for f-soma.

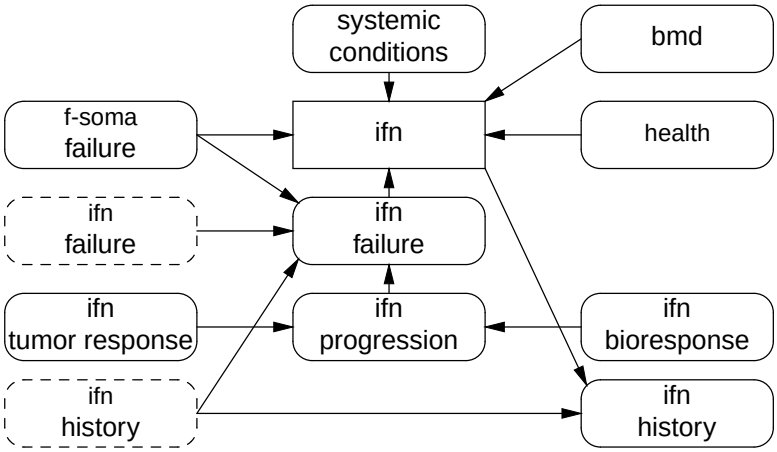


Figure 6.13: The treatment strategy for ifn.

For interferon treatment (Fig. 6.13), we additionally need to take into account whether or not f-soma treatment has failed, since ifn treatment is only given after f-soma failure when the systemic conditions hold, health is acceptable, and there is no bone-marrow depression. We also need to take into account the treatment history,

since interferon may only be administered for a year.

For r-soma and r-mibg treatment, we use a similar structure, where r-soma treatment also takes into account that the patient may not suffer from renal failure. Additionally, we need to take into account that we make a random choice between the two treatments, as represented by Fig. 6.14.

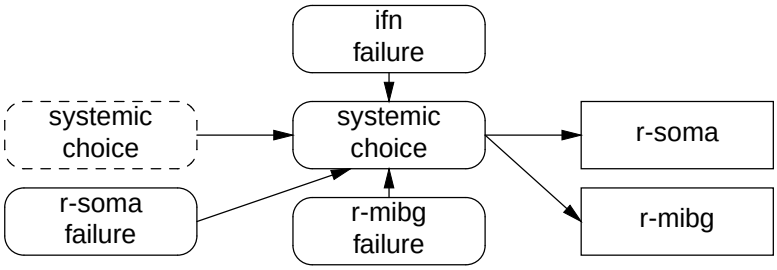


Figure 6.14: Representing the choice between two treatments.

Farmacological MIBG is administered in case of a good blood pressure when f-mibg has failed and other treatments are not applicable for instance due to a poor condition. The treatment history is used to represent the notion that we treat for three months and then stop for six months (Fig. 6.15).

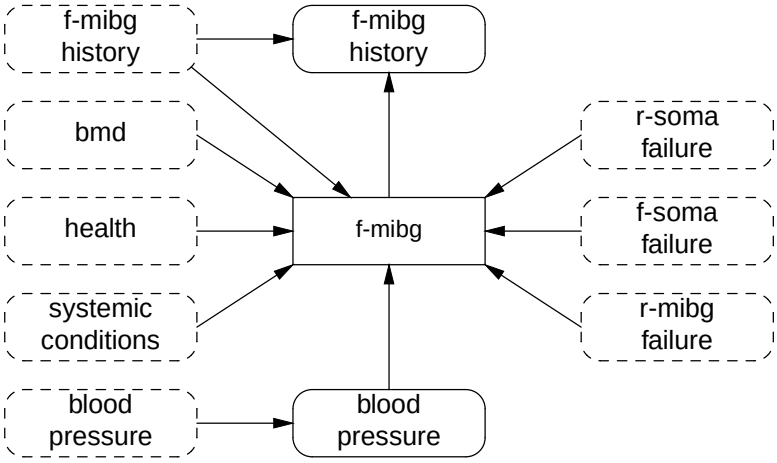


Figure 6.15: The treatment strategy for f-mibg.

Once the transition model for the pathophysiological and treatment components has been specified, we need to define a prior model. This prior model can be generated in part from the independencies that are already represented in the transition model, although we need to take into account possible associations between random variables. For example, patient age is conditioned by both patient gender and the primary localization of the tumor since these variables are correlated at admission time.

6.3 Validation of the carcinoid model

In order to determine if the performance of a prognostic model is satisfactory, it is important to validate the results that are obtained by means of the model (Altman and Royston, 2000). In order to validate the carcinoid model, we will use (1) a clinical database, obtained from the Netherlands Cancer Institute, containing data on 129 patients with a diagnosed low-grade midgut carcinoid tumor, and (2) more extensive data on a number of individual patients. Validation of the carcinoid model is done in a number of ways. In Section 6.3.1, we compare survival curves that were generated by the carcinoid model with Kaplan-Meier curves that have been constructed from the clinical database. In Section 6.3.2, quality of the prior and transition models is determined by means of the clinical database and a particular scoring rule, and compared with that of a proportional hazards (PH) model. Finally, in Section 6.3.3, individual patient cases are analyzed by means of the carcinoid model, which is in close correspondence with how the carcinoid model would be used in clinical practice.

6.3.1 Survival curves

Let T be a *survival random variable* where $t \in [0, \infty)$ denotes a survival time, such that the *survivor function*, given by $S(t) = 1 - P(T \leq t)$, represents the probability of survival at time t. Let t_j denote the j-th smallest survival time that occurs in the database. An estimate $\hat{S}$ of the survivor function is constructed from data $\mathcal{D}$ by means of the Kaplan-Meier method (Kaplan and Meier, 1958) as follows:

$$\hat{S}(t_j) \;=\; \prod_{i=1}^{j} \hat{P}(T > t_i \mid T \geq t_i) = \hat{S}(t_{j-1}) \cdot \frac{\mathit{risk}(t_j) - \mathit{failure}(t_j)}{\mathit{risk}(t_j)}\,,$$

where $\hat{S}(0) = 1$ by definition, $\mathit{risk}(t_j)$ denotes the number of people at risk of dying at time t_j, and $\mathit{failure}(t_j)$ denotes the number of people that has died in the period $[t_j, t_{j+1})$.

Figure 6.16 depicts the Kaplan-Meier curve as estimated from data, and a survival curve, which was generated by the carcinoid model, where we disregard patient-specific evidence. There is a salient jump in the Kaplan-Meier curve some five years after admission to the hospital, which the physician hypothesized to be due to the exhaustion of treatment options at that point. The fit of the survival curve that was generated by the model is not perfect, since it overestimates patient survival, especially for longer survival times. An analysis of the cases in the database showed that survival curves differed considerably for patients with or without hepatic metastases.

Figure 6.17 shows that patients *without* hepatic metastases have a lower survival rate in the first few years. The physician gave the following explanation of this seemingly counterintuitive result: patients that present without hepatic metastases must

have other complications, since otherwise they would not have been sent to the referral centre in the first place. In this case, we can expect the presence of other malignancies or metastatic disease in other locations.

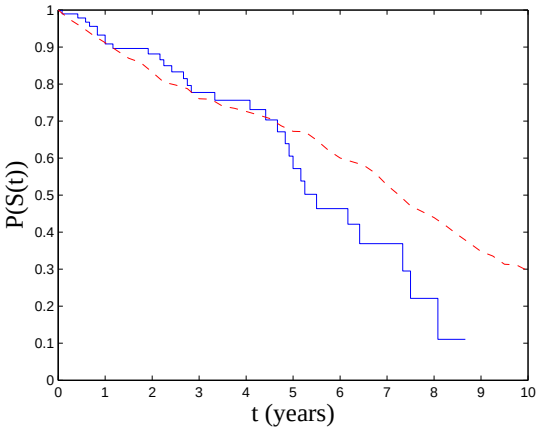


Figure 6.16: Kaplan-Meier curve (solid line) as estimated from data, and survival curve (dashed line) as predicted by the model.

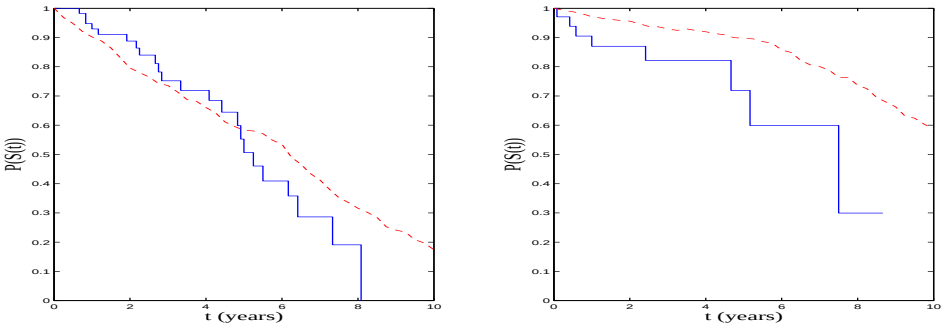


Figure 6.17: Kaplan-Meier curves (solid lines) as estimated from data, and survival curves (dashed lines) as predicted by the model, for patients that enter the hospital with hepatic metastases (left), or without hepatic metastases (right).

We have formalized this by means of the variables in Fig. 6.18, where other malignancy takes part as a cause in the causal interaction model for endurance in Eq. (6.1). The survival curve that was computed from the updated model shows a somewhat better correspondence between the Kaplan-Meier curve and the survival curve that was computed from the model. The improvement is not dramatic however, due to the fact that the patient group without hepatic metastases is relatively small.

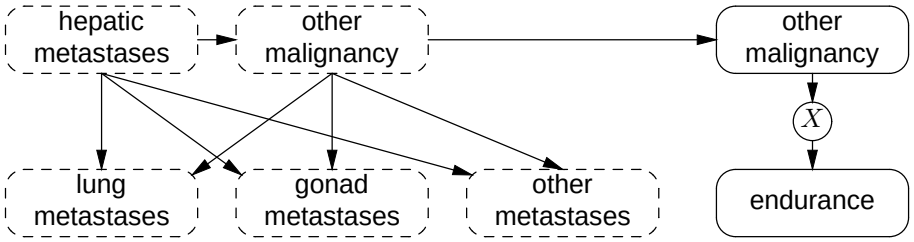


Figure 6.18: Conditioning of variables by hepatic metastases at admission time (dashed variables), and representation of patient death due to another malignancy (solid variables), where X is an intermediate variable that quantifies the effect on patient endurance.

6.3.2 Model likelihood

One way to assess model quality is by using a *scoring rule* (Murphy and Winkler, 1984), that penalizes a probability model based on a database $\mathcal{D} = \{\mathbf{u}^1, \ldots, \mathbf{u}^N\}$ with instances $\mathbf{u}^i = (u_1^i, \ldots, u_m^i)$. Let $\mathbf{X} \subseteq \mathbf{U}$ denote the variables of interest, and $\mathbf{Y} \subseteq \mathbf{U}$, $\mathbf{X} \cap \mathbf{Y} = \varnothing$ the variables for which we have evidence. We assume that instances $\mathbf{u}^i$ are independently and identically distributed, and use the *logarithmic score* (Spiegelhalter et al., 1993):

$$S = -\sum_{i=1}^{N} \log P(\mathbf{x}^i \mid \mathbf{y}^i)$$

which incurs a penalty if a low probability is assigned to events that actually occur (note that we have to account for the fact that the logarithmic score is undefined if this probability is zero). We compare the logarithmic score S of our model $\mathcal{M}$ with the logarithmic score S^{ref} of a reference model $\mathcal{M}^{\text{ref}}$, where

$$S^{\text{ref}} - S = \log\left(\frac{P(\mathcal{D} \mid \mathcal{M})}{P(\mathcal{D} \mid \mathcal{M}^{\text{ref}})}\right). \tag{6.4}$$

A positive sign of this quantity expresses that model $\mathcal{M}$ is preferred, and a negative sign expresses that model $\mathcal{M}^{\text{ref}}$ is preferred. The quantity

$$P(\mathcal{D} \mid \mathcal{M}) / P(\mathcal{D} \mid \mathcal{M}^{\text{ref}})$$

is known as the Bayes factor, where $P(\mathcal{D} \mid \mathcal{M})$ is the likelihood of model $\mathcal{M}$ given the data and $P(\mathcal{D} \mid \mathcal{M}^{\text{ref}})$ is the likelihood of model $\mathcal{M}^{\text{ref}}$ given the data. We use Eq. (6.4) in order to determine the performance of the prior and transition models of the carcinoid model.

Prior model quality

For the prior model, our interest is in variables

$$\begin{aligned}\mathbf{E}(0) = \quad \{ \quad & \mathsf{gender}(0), \mathsf{age}(0), \mathsf{5\text{-}hiaa}(0), \mathsf{cga}(0), \mathsf{diarrhea}(0), \mathsf{flushing}(0), \\ & \mathsf{bowel\ obstruction}(0), \mathsf{hepatic\ metastases}(0), \mathsf{octreoscan}(0), \\ & \mathsf{mibgscan}(0), \mathsf{primary\ localization}(0), \mathsf{mesenterial\ fibrosis}(0) \quad \}\end{aligned}$$

for which evidence at the time of admission to the hospital is available in the clinical database. The carcinoid model was then used in order to compute the logarithmic score:

$$S = -\sum_{i=1}^{N} \log \sum_{j=1}^{m} P(u_j^i \mid \mathbf{y}^i)$$

where $\mathbf{y}^i$ represents the evidence for instance $\mathbf{u}^i$. This logarithmic score is compared with that of a reference model that assigns a uniform probability to each possible value of the goal variable. The results are listed in Table 6.5, where the junction tree algorithm was used to compute the posterior distributions. Results show that most variables are predicted better by the carcinoid model. Exceptions are 5-hiaa and hepatic metastases, which is most likely caused by the fact that the model overestimates the causal relation between the presence of hepatic metastases and increased 5-hiaa levels.

Table 6.5: Bayes factors for the prior model.

Variable	Bayes factor	Variable	Bayes factor
gender(0)	1.7	bowel obstruction(0)	$1.7 \cdot 10^{14}$
age(0)	$5.3 \cdot 10^{7}$	hepatic metastases(0)	$7.0 \cdot 10^{-5}$
5-hiaa(0)	$2.2 \cdot 10^{-3}$	octreoscan(0)	1.0
cga(0)	8.4	mibgscan(0)	14
diarrhea(0)	30	primary localization(0)	$6.0 \cdot 10^{7}$
flushing(0)	$4.4 \cdot 10^{3}$	mesenterial fibrosis(0)	$6.4 \cdot 10^{2}$

Transition model quality

In order to determine the quality of the prediction of patient survival from patient specific covariates by the model, we compute the logarithmic score for the prediction of ten-year survival (in terms of three-month follow-up times), given covariates $\mathbf{y}^i$:

$$S_{\text{survival}} = -\sum_{i=1}^{N} \log P(\mathsf{survival}^i(0:40) \mid \mathbf{y}^i)\,. \tag{6.5}$$

We compare this score with the score of a PH model (Cox, 1972; Cox and Oakes, 1984): $S^{\text{PH}}_{\text{survival}}$, where baseline hazard and coefficients were estimated from data. The

variable age was discretized into $\mathsf{age} < 58$ and $\mathsf{age} > 58$, with 58 being the average age of patients in the database, and an imputation scheme was used that imputed missing values based on their prior probability, since the large number of missing values caused numerical instability of the algorithm. The obtained coefficients are shown in Table 6.6.

Table 6.6: Estimated coefficients $\boldsymbol{\theta}$ of the PH model $S^{\text{PH}}_{\text{survival}}$.

Variable	Coefficient	Variable	Coefficient
gender(0)	0.8660	bowel obstruction(0)	0.5314
age(0)	1.0454	hepatic metastases(0)	−0.3146
5-hiaa(0)	0.0043	octreoscan(0)	1.2103
cga(0)	0.7091	mibgscan(0)	−0.3952
diarrhea(0)	0.3371	primary localization(0)	−0.1812
flushing(0)	−0.5039	mesenterial fibrosis(0)	−1.2263

Note that some of the coefficients are negative, which indicates that, contrary to expectations, the presence of that particular "risk factor" is beneficial for patient survival *according to the database* $\mathcal{D}$. The computation of Eq. (6.5) for the PH model needs to take into account that there are missing values for some patient cases. Furthermore, in order to compare with the carcinoid model, we look at discrete times $t \in \{0, \ldots, 40\}$. Let $s^i_t = 1_{\mathsf{survival}^i_t = \mathit{yes}}$ and $\bar{s}^i_t = 1 - s^i_t$. We use the following equation, as an estimate of the logarithmic score for the PH model:

$$S^{\text{PH}}_{\text{survival}} \quad = \quad -\sum_{i=1}^{N} \log \prod_{t=0}^{40} \left(1^{\bar{s}^i_t} + (-1)^{s^i_t} \sum_{\mathbf{z}^i} S_0(t)^{\exp(\boldsymbol{\theta}\mathbf{y}^i)} P(\mathbf{z}^i) \right)^{1-c^i_t}$$

where $c^i(t) = 1$ ($c^i(t) = 0$) indicates that patient i is censored (uncensored) at time t, and $\mathbf{z}^i$ are instantiations of variables $\mathbf{Z}^i \subseteq \mathbf{Y}^i$ that have missing values for patient i. The contribution of each such instantiation is weighted by its prior probability in the database under the assumption that missing covariates are independent. For the carcinoid model, we simply instantiate the covariates $\mathbf{y}^i$ for which values are known, and compute

$$S_{\text{survival}} = -\sum_{i=1}^{N} \log \prod_{t=0}^{40} P(\mathsf{survival}^i(t) \mid \mathbf{y}^i)^{1-c^i_t}$$

using particle filtering with 3 000 particles. As a result, we have found that $S^{\text{PH}}_{\text{survival}} = 1.190 \cdot 10^3$ and $S_{\text{survival}} = 1.229 \cdot 10^3$, which gives a Bayes factor of $1.155 \cdot 10^{-17}$ that is significantly in favour of the PH model.

We remark that the PH model did have the advantage that its parameters were learnt from the data on which it was tested, whereas the carcinoid model is fully estimated from expert knowledge. Furthermore, although the carcinoid model is outperformed by the PH model in this respect, the carcinoid model has the advantage that

(1) it can make use of evidence that becomes available over time, (2) it may answer other types of queries, such as the expected cause of death, or the expected future treatment, and (3) since the carcinoid model is an explicit causal model of disease progression, the drawn conclusions are more understandable. These features considerably improve both the quality and detail of the prognosis, as will be demonstrated.

6.3.3 Patient specific predictions

In this section, we intend to show that having an explicit model of medical domain knowledge at ones disposal has additional benefits that cannot be obtained by means of standard proportional hazards models. In order to demonstrate this, we focus on individual patients where data about these patients, as taken from the database, is supplemented with more specific clinical evidence as found in the physician's paper records.

Patient A

Patient A is a 70 year old male that came into the hospital with a small-bowel tumor and some health-related problems. The patient had elevated 5-hiaa levels, and suffered from diarrhea, flushing, and obstruction, but it was found that the patient was free from hepatic metastases and other malignancies. There was no indication of carcinoid heart disease, and both the octreoscan and mibgscan were positive. The patient eventually died of wasting five years and two months after admission.

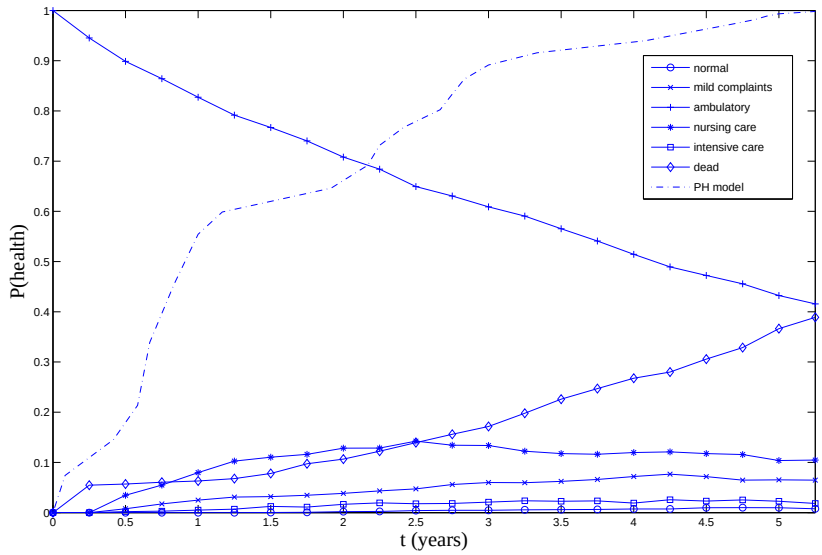


Figure 6.19: Predictions of patient A's health.

The predictions of the carcinoid model as well as the PH model for patient health are shown in Fig. 6.19. According to the carcinoid model, the patients starts with an ambulatory health status, where over time the chance of needing nursing care first increases and then decreases since the patient's chance of dying increases. In contrast, the PH model can only predict the probability of patient death over time and due to the negative contribution of the covariates an unrealistically high probability of patient death is assigned.

During hospitalization, the patient was given several treatments. He received bowel resection at admission due to obstruction. After ten months, farmacological somatostatin treatment was initiated due to the development of serotonin-producing metastases. Thirteen months after admission, the patient received farmacological MIBG for four months since deteriorations in health precluded other treatments. After three years and nine months, the patient received another bowel resection due to the development of mesenterial fibrosis.

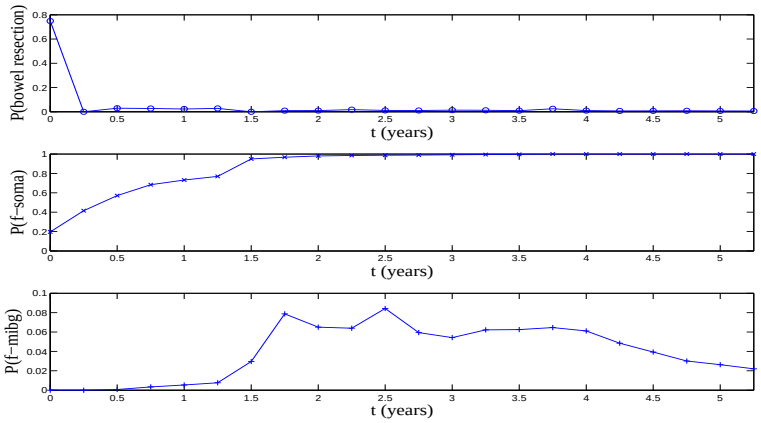


Figure 6.20: Predictions of treatment for patient A.

Figure 6.20 depicts the predictions of the carcinoid model for the treatments which patient A will receive. We condition here on the evidence that is present during admission, and on observations that are made over time; namely, the development of serotonin-producing metastases after one year (which we take here to be hepatic metastases), a deterioration in health after 18 months, and the development of mesenterial fibrosis after 45 months. The figure is in accordance with the physician's expectations. At admission, the model suggests bowel resection with high probability. This probability drops to zero at 18 months (since health has deteriorated), and shows a small increase at 45 months (due to the development of mesenterial fibrosis). The model also predicts that farmacological somatostatin is administered early on and continued indefinitely. Finally, the probability that farmacological MIBG is administered increases when it is found that health has deteriorated.

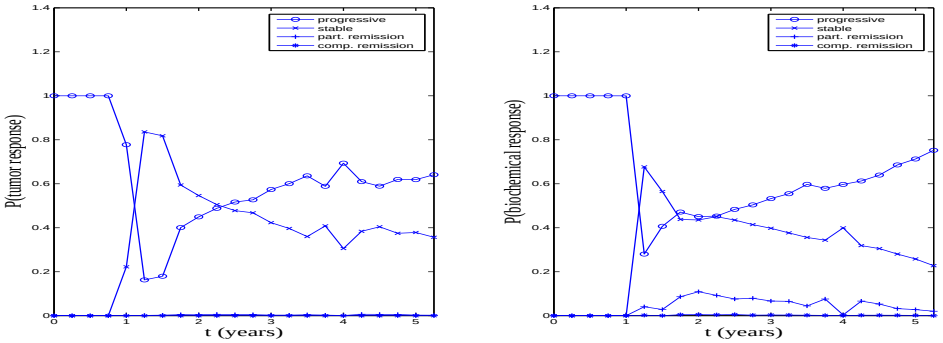


Figure 6.21: Predictions of tumor response and biochemical response for patient A.

If we instantiate the projected treatment for this patient, then we can examine the predicted tumor response and biochemical response (Fig. 6.21). Both for tumor and biochemistry, the model predicts an initial stabilization, which over time has a higher chance of becoming progressive. This is in agreement with the physician's expectation, although progression was expected to occur more rapidly.

Patient B

Patient B is a 59 year old male that came into the hospital with a small-bowel tumor, all the symptoms of carcinoid syndrome, and minor health-related problems. It was found during admission that the patient suffered from cardiac valve thickening together with moderate fibrosis as well as mesenterial fibrosis. The patient eventually died fourteen months after admission due to complications after cardiac surgery at thirteen months. An important question, would be to determine at admission time the probability that the patient will receive cardiac surgery. Figure 6.22 depicts this probability for the coming five years, and shows that this probability is at a reasonably high level after thirteen months.

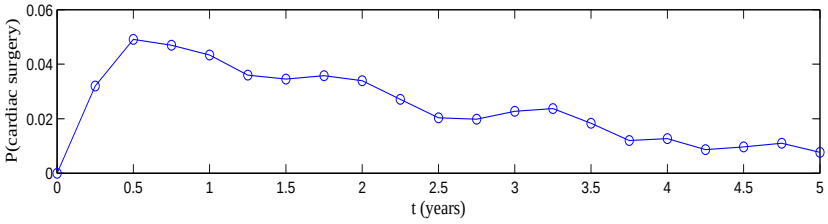


Figure 6.22: Predicting cardiac surgery for patient B.

Next to predicting future patient health and projected treatments, we may em-

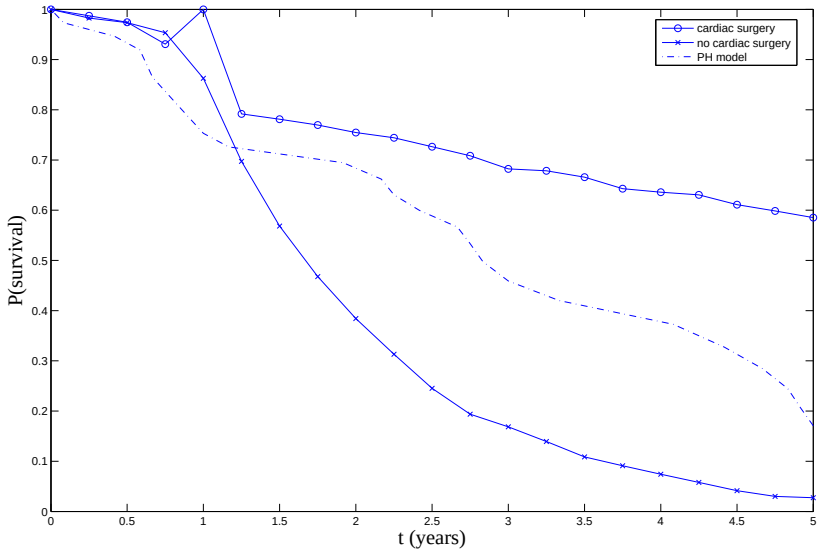


Figure 6.23: Comparing survival of patient B given cardiac surgery and no cardiac surgery.

ploy the model in order to distinguish between different scenarios. For instance, for a patient that has developed carcinoid heart disease after one year, we may compare the expected course of events in case the patient receives cardiac surgery between twelve and fifteen months with the expected course of events in case the patient never receives cardiac surgery. This comparison is shown in Fig. 6.23 and motivates the physician's choice of performing cardiac surgery since this is expected to improve long-term survival. However, the figure also shows that performing this type of surgery may lead to patient death in a minority of cases and, unfortunately, patient B also died after surgery. The sudden increase in survival probability after one year is implied by the treatment which the patient received at that time. The PH model is unable to distinguish between the treatment and no-treatment conditions and its estimate is located in between both scenarios.

With respect to mesenterial fibrosis, the model predicts that there is a 78% chance that bowel resection is immediately performed. It was found however that the patient did not receive such a surgical intervention. After some deliberation with the physician, it was found that the operationalization of mesenterial fibrosis in the model differed from that in the database. In the model, the presence of mesenterial fibrosis indicates severe fibrosis, which warrants the intervention, whereas in the database, presence of mesenterial fibrosis also indicates mild fibrosis, which does not warrant such an intervention.

Patient C

Patient C is a 68 year old male of which the primary localization is unknown. He came in with extreme CgA levels, no signs of the carcinoid syndrome or other malignancies, and only minor health problems. The patient had a negative octreoscan and a positive mibgscan. After five months the patient started to receive farmacological somatostatin. From eleven to fifteen months, the patient received interferon. After fourteen months, it was found that the patient had elevated NT-pro-BNP levels. Currently, seventeen months after diagnosis, the patient starts to receive radiolabeled MIBG. The patient remains alive today, and we wish to predict patient health and projected treatment for the next five years.

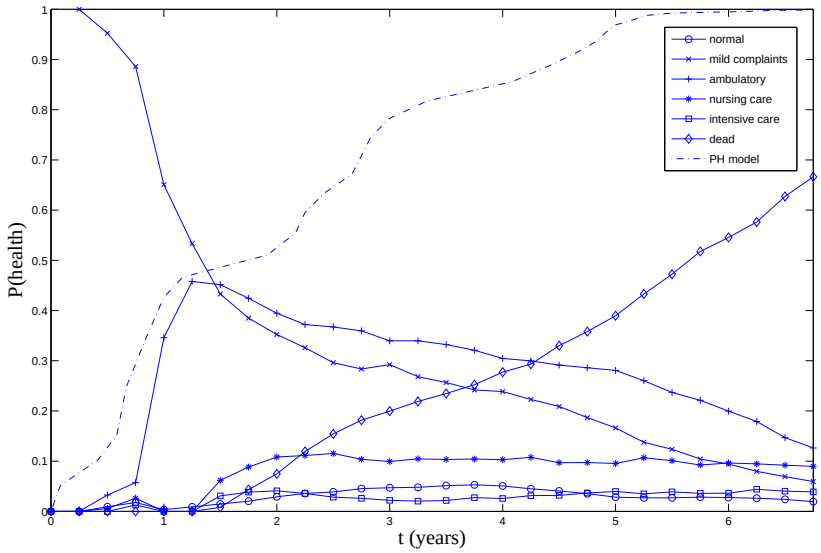


Figure 6.24: Predicting future health of patient C.

Figure 6.24 shows the predictions for patient health. Note that at seventeen months, the carcinoid model predicts that the patient either has mild complaints, or is ambulatory, since treatment with radiolabeled MIBG requires an acceptable health status. Over time, the probability of being in these states decreases, and the probability of requiring nursing care/dying increases. Five years later, the patient is predicted to have a 34% chance of remaining alive. Note that, similar to patient A, the PH model assigns an unrealistically high probability of patient death due to the negative influence of the covariates.

Even if NT-pro-BNP levels were elevated, the model assigned a low probability to the development of carcinoid heart disease. This is consistent with the physician's expectations, since diarrhea and flushing were absent at time of admission (indicating

that CHD due to elevated serotonin levels is unlikely), and no cardiac surgery was performed immediately after elevated NT-pro-BNP levels were noticed. The model also assigned low probabilities to the development of other complications such as a crisis or mesenterial fibrosis, and therefore did not require treatments specific to these complications.

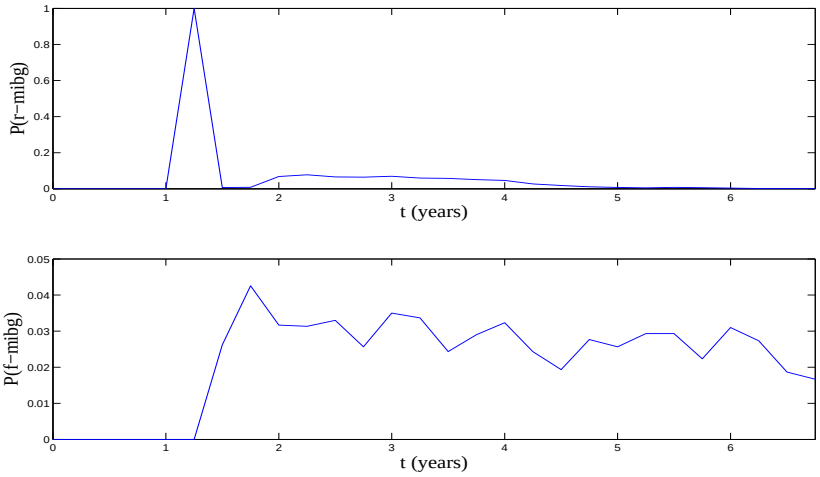


Figure 6.25: Predicting future MIBG treatment of patient C.

For f-soma treatment, the data in the database is not consistent with the model; since the octreoscan of the patient is negative, the model predicts that farmacological somatostatin is not administered. In reality the patient was given this treatment since his condition at that time precluded other treatment. Upon entering this evidence, the model responds by giving no biochemical or tumor response, which is in accordance with the observed progressive disease of the patient, and by the discontinuation of this treatment for the remaining time slices. For the same reason, the model predicts that radiolabeled octreotide will not be administered in the future. The only remaining applicable systemic treatments are then radiolabeled MIBG and farmacological MIBG, the predictions of which are shown in Fig. 6.25. The figure shows that the patient has a chance of receiving radiolabeled MIBG once more, where this chance is smeared out over a longer period, since patient health should be acceptable. The patient also has a small chance of receiving farmacological MIBG at each time slice, since this does not require any conditions other than a normal blood pressure.

6.4 Discussion

In this section, the results which have been obtained from this study are discussed.

6.4.1 Quality of the carcinoid model

In Section 6.3.2, the prior model has been shown to be of better quality than a reference model with uniform posteriors in terms of logarithmic score. The validation results for the transition model have shown that the carcinoid model, as constructed from expert knowledge, is outperformed by the proportional hazards model, in terms of logarithmic score with respect to patient survival. It is found that predictions by the carcinoid model are miscalibrated in the sense that survival is often overestimated, which is in accordance with the general observation that physicians tend to overestimate patient survival (Glare et al., 2003; Christakis and Lamont, 2000). Since the carcinoid model is a prototype, we expect that predictive performance can be improved by refining the model and its probability estimates.

An advantage of the carcinoid model is that it is not restricted to the evidence variables that are known at admission, since it allows for the inclusion of evidence that becomes available as a patient progresses. This leads to more accurate predictions, as was demonstrated in Section 6.3.3. Another advantage of the carcinoid model is that it may answer queries other than patient survival, such as expectations regarding the cause of death, health status, projected treatment, and treatment effects. In fact, with minor modifications, the carcinoid model could also be used for patient monitoring (comparing expected and observed patient status) or treatment selection. Finally, the carcinoid model can explain its predictions in terms of a semantics that captures cause-effect relations between domain variables.

6.4.2 Characteristics of the carcinoid database

Since the treatment protocol for carcinoid tumors is still under development, the database included sequences of treatments that were impossible according to the model. Furthermore, some treatments that were present in the database are no longer used in clinical practice. For instance, chemotherapy is currently considered too aggressive as a treatment option for low-grade carcinoid patients. Also, sometimes the operationalization of variables in the database was not clear, as was the case with mesenterial fibrosis, and abdominal pain, which was excluded for this reason. Additionally, Table 6.6 shows that the presence of some risk factors had a positive effect on patient survival in the database. The presence of mesenterial fibrosis, for instance, had a very strong positive effect on patient survival, and was in fact the strongest effect found. Clearly, this does not match with the carcinoid model, which predicts that mesenterial fibrosis has a negative effect on patient survival.

6.4.3 Encountered difficulties

Even though this prototype has demonstrated that disease progression for complex domains can be modeled succesfully by means of dynamic Bayesian networks, there are also some lessons to be learned from this study. During the development of the

structure of the carcinoid model, it was found that sometimes, the physician had difficulty in determining the causal structure of the domain. For example, in the early stages of modeling, a negative octreoscan (absence of tumor mass on the scan) was associated with a *negative* effect on patient survival as based on clinical expertise. Later on, when the causal structure of the domain was made explicit, it became clear that a negative scan implies that radiolabeled somatostatin cannot be given as treatment, therefore reducing the chances of survival. At these early stages of modeling, it was clearly hard for the physician to structure the domain, which frequently led to the claim that *everything is connected to everything*. However, as domain variables became consolidated, the task became easier; especially when pathophysiology was distinguished from the treatment protocol, and modeling focused on individual submodels for the various complications. Another problem that was encountered is that sometimes the physician was unsure of certain (in)dependencies. For example, the formation of mesenterial fibrosis is still under debate, thereby making model construction and parameter estimation difficult.

During the estimation of probabilities, it was found that the physician was not very sure about the point estimates that she provided. Therefore, it might have been advisable to model the physician's uncertainty explicitly in terms of hyperparameters, although this would also have increased model complexity considerably. Various kinds of biases have also been observed during the estimation process. For instance, the physician sometimes claimed initially that some events never occur (while in reality they had a small chance of ocurring) or always occur (while in reality they had a small chance of not occurring). It seemed to be the case that the physician conditioned her estimates on the average situation, without taking into account possible exceptions. The physician also noticed that she tended to base her estimates more strongly on patients that stood out in one sense or another. These are examples of the *availability heuristic* (Tversky and Kahneman, 1973). Another observed bias was the *recency effect* (Atkinson and Shiffrin, 1971), where knowledge about patients that were seen most recently was used disproportionally for belief estimation.

Sometimes, difficulties arose due to the discretization of continuous variables. As a simple example, consider the variable age. By modeling age progression by means of a small probability that patients advance one discrete state at a time (e.g., from *50-60* to *60-70* and from *60-70* to *70-80*), we have the bizarre effect that a very small patient group ends up in much older age groups after a few time-slices. Although we can still approximate the effect of age on patient survival to a reasonable degree, this behavior is clearly undesirable.

A general problem that was encountered is the fact that carcinoid disease as a whole is still not well-understood and disease progression is subject to much variation, which made model construction a difficult task. Also, due to the highly complex pathophysiology of carcinoid tumors, and the large number of treatments that are used, model complexity grew considerably, leading to a long development time.

6.5 Summary

In this chapter, we have demonstrated that prognostic models can be constructed with dynamic Bayesian networks that take causal, temporal, and decision-making characteristics into account. Although the more realistic carcinoid model does not achieve the quality in terms of logarithmic score that was obtained by a proportional hazards model that was learnt from data, this performance could be improved by subsequent model refinement. The carcinoid model also has additional benefits, such as the incorporation of evidence over time, the possibility to answer different queries, and an explicit representation of the problem domain. It is our hope that the discussed carcinoid model demonstrates the potential of probabilistic models in medicine, and guides the future development of other clinical decision support systems.

Chapter 7

Bayesian Classifiers for Clinical Decision Support

The problem of representing and reasoning with medical knowledge has attracted considerable attention during the last decades. In particular, ways of dealing with the uncertainty involved in medical decision making has been identified again and again as one of the key issues in this area and Bayesian networks are nowadays considered as standard tools for representing and reasoning with uncertain biomedical knowledge (Lucas et al., 2004). However, although possible, manually constructing a Bayesian network for a realistic medical domain is a laborious and time-consuming task.

Another approach to the construction of Bayesian networks is to learn the structure and parameters of a Bayesian network from data (Cooper and Herskovits, 1992; Buntine, 1994; Heckerman et al., 1995; Bouckaert, 1995). Parameters can be efficiently computed as the maximum likelihood estimates of the parameters given the data, but learning the correct graph structure requires a search in the space of possible acyclic directed graphs which grows superexponentially with the number of nodes (Robinson, 1973; McKay et al., 2004), and exhaustive search is therefore generally infeasible.

One approach to the problem of learning a Bayesian network from data is to search for an optimal graph structure in a restricted search space. Although the resulting Bayesian networks are not expected to represent the joint probability distribution over random variables accurately, they may still be used for computing a MAP estimate $\mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{x} \mid \mathbf{y})$ as long as the (possibly inaccurate) estimate of $P(\mathbf{x}^* \mid \mathbf{y})$ exceeds the estimate of $P(\mathbf{x} \mid \mathbf{y})$ for all $\mathbf{x} \neq \mathbf{x}^*$. If we regard $\mathbf{x}$ to be a class assignment based on the available evidence $\mathbf{y}$ then this approach can be interpreted as solving a classification problem. Since we use inference in a Bayesian network with a restricted graph structure in order to solve the classification problem, we will call these networks Bayesian classifiers. One example of Bayesian classifiers

This chapter is based on (van Gerven and Lucas, 2004a; van Gerven, 2007b; van Gerven et al., 2007a; van Gerven and Lucas, 2007).

are the naive Bayes classifier (Duda and Hart, 1973), where evidence variables are assumed to be conditionally independent given the class variable.

Classification is an important concept in current medical practice. The (differential) diagnosis of disease, the selection of appropriate treatment, and the prediction of patient survival can all be cast in a framework that selects the correct class from a set of possible classes given observed patient data. In case of Bayesian classification, each class has an associated posterior probability that represents the belief in that particular class. In this chapter, we focus on Bayesian classification for clinical decision support, and address three different Bayesian classification methods. In Section 7.1 we describe the *maximum mutual information* algorithm and its application to the diagnosis of liver disease. The aim here is to learn classifier structures that retain some of the (in)dependence structure that holds between variables in the domain. In Section 7.2 we develop *tensor decompositions* as a novel Bayesian classification technique and show that it performs well for the diagnostic problem that has already been addressed in Section 7.1. In Section 7.3 we analyze the semantics and performance of the *noisy-threshold* classifier (Jurgelenaite and Heskes, 2006) in the context of a prognostic problem in clinical oncology. We end with conclusions about our research in Section 7.4.

7.1 Maximizing mutual information

Bayesian classifiers are a valuable tool for the automation of clinical tasks. However, most Bayesian classifiers place very heavy restrictions on the form of the underlying Bayesian network structure. The naive Bayes classifier, for instance, allows no freedom in the graph structure. These constraints disallow many (in)dependence statements, such as the encoding of higher-order dependencies, where the *order* of a dependency is the size of the conditioning set $\pi(X)$ of the conditional probability $P(X \mid \pi(X))$ associated with the dependency (van Dijk et al., 2003). Furthermore, the constraints lead to classifier structures which may be unintelligible to the physician. It is felt that intelligible classifier structures may increase the acceptance of the use of Bayesian classifiers in medical practice because of an improved accordance with a physician's domain knowledge. Classifier performance will also benefit from such an agreement, since the physician may now aid in identifying counter-intuitive (in)dependence statements.

Alternative classification algorithms have been devised that focus on lifting the independence assumptions of the naive Bayes model (Spiegelhalter and Knill-Jones, 1984). The tree-augmented naive (TAN) classifier (Friedman et al., 1997) represents correlations between evidence variables as arcs between evidence variables in the form of a tree, the forest-augmented naive (FAN) classifier generalizes the TAN classifier by representing correlations between evidence variables as a forest of trees (Sacha et al., 2002; Lucas, 2004), and the limited-dependence classifier (Sahami,

1996), allows each evidence variable to have k incoming arcs, where k is chosen beforehand.

In this section, we introduce a new algorithm to construct Bayesian network classifiers. This so-called *maximum mutual information* (henceforth MMI) algorithm builds a structure which favors those features showing maximum (conditional) mutual information and resembles the limited-dependence classifier in the sense that evidence variables are allowed to have multiple incoming arcs. Next to the problems arising from constraints on classifier structure, Bayesian classifiers perform poorly in the face of small databases. (In)dependence statements may have only little support from the database (in terms of number of records) and yet are encoded within the classifier structure. The MMI algorithm incorporates a solution by making use of a heuristic during structure learning which penalizes quantities that are estimated from few data samples.

Structure learning algorithms that use information-theoretical measures such as mutual information are known as *constraint-based* algorithms. They have been researched extensively in the context of learning arbitrary Bayesian network structures (Cheng et al., 2002; Spirtes et al., 1993; Chickering and Meek, 2006). In contrast, in this research we do not aim to build arbitrary Bayesian network structures, but instead aim to build a structure learning algorithm for Bayesian classifiers that provides a balance between the complexity issues associated with general structure learning algorithms and the highly restrictive structural assumptions of classifier structure learning algorithms. In order to determine the performance of the MMI algorithm we make use of a clinical dataset of hepatobiliary (liver and biliary) disorders, the reputation of which has been firmly established. Classification accuracy of the algorithm is compared with that of an existing system for diagnosis of hepatobiliary disorders, as well as with that of FAN classifiers, of which the naive Bayes classifier and TAN classifier are special cases.

7.1.1 Probabilistic classification

One way to determine the performance of a Bayesian classifier is to compute its classification accuracy. Let $\mathcal{D}$ be a dataset consisting of N cases and let c^k be the value of the class variable C given the k-th example $\mathbf{e}^k$. The *classification accuracy* is defined as the percentage of correctly classified cases:

$$\eta(\mathcal{D}) = \frac{1}{N} \sum_{k=1}^{N} \Big(1 - L(\mathbf{e}^k)\Big) \times 100\%\,, \tag{7.1}$$

where $L(\mathbf{e}^k)$ is the *loss function*, which equals zero if $\arg\max_c\{P(C=c \mid \mathbf{E}=\mathbf{e}^k)\}$ $= c^k$ and equals one otherwise.

In order to assess the classification accuracy of the MMI algorithm, we compare it with the classification accuracy of the forest-augmented naive (FAN) classifier (Fig. 7.1). A FAN classifier is a modification of the tree-augmented naive (TAN)

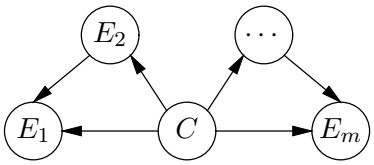


Figure 7.1: A forest-augmented naive (FAN) classifier. For each evidence variable E_i there is at most one incoming arc allowed from $\mathbf{E} \setminus \{E_i\}$ and exactly one incoming arc from the class variable C. Both the naive classifier and the tree-augmented naive classifier are extreme cases of the forest-augmented naive classifier

classifier, where the topology of the resulting graph over evidence variables is restricted to a forest of trees (Lucas, 2004). The algorithm to construct FAN classifiers, that is used in this chapter, is based on a modification of the algorithm to construct TAN classifiers (Friedman et al., 1997), where the conditional mutual information, as computed from a dataset $\mathcal{D}$, is used to build a minimum cost spanning tree between evidence variables $\mathbf{E} = \{E_1, \ldots, E_m\}$.

7.1.2 The maximum mutual information algorithm

The maximum mutual information algorithm is a classifier construction algorithm that is less restrictive than the discussed FAN algorithm. It uses both the computed mutual information between evidence variables and the class-variable, and the computed conditional mutual information between evidence-variables as a basis for constructing a Bayesian classifier. The mutual information (MI) between an evidence variable E and the class-variable C is given by:

$$I(E, C) = \sum_{e,c} P(e, c) \log \frac{P(e, c)}{P(e)P(c)}\,, \tag{7.2}$$

whereas the conditional mutual information between evidence variables, given other evidence variables and/or the class variable is given by:

$$I(E, E' \mid \mathbf{A}) \;\; = \;\; \sum_{e,e',\mathbf{a}} P(e, e', \mathbf{a}) \log \frac{P(e, e' \mid \mathbf{a})}{P(e \mid \mathbf{a})P(e' \mid \mathbf{a})} \tag{7.3}$$

with $\mathbf{A} \subset \mathbf{E} \cup \{C\}$. Contrary to naive and TAN classifiers, the MMI algorithm makes no assumptions about the initial network structure. It starts from a fully disconnected graph, whereas the FAN algorithm starts with a naive classifier structure such that $(C, E) \in A(G)$ for all evidence variables $E \in \mathbf{E}$. Since redundant attributes are not encoded, network structures are sparser, at the same time indicating important information about independence between class and evidence variables. In this sense, the MMI algorithm can be said to resemble *selective Bayesian classifiers* (Langley and Sage, 1994). The algorithm iteratively selects the arc with highest (conditional)

mutual information from the set of candidates and adds it to the Bayesian network $\mathcal{B} = (G, P)$. It starts by computing $I(C, E)$ for the set $\mathbf{S} = \mathbf{E}$. From this set, the evidence variable E having highest mutual information with the class variable C is selected. This candidate is removed from $\mathbf{S}$ and (C, E) is added to the arcs in G. Subsequently, it will construct all candidates of the form (E', E) and add them to (an initially empty) set $\mathbf{A}$ if E' was added later to G than E. The conditional mutual information $I(E', E \mid \pi(E))$ is computed for these candidates. Now, the algorithm iteratively selects the candidate of $\mathbf{S} \cup \mathbf{A}$ having the highest (conditional) mutual information. If a candidate $(E', E) \in \mathbf{A}$ is chosen, then $I(E'', E \mid \pi(E))$ is recomputed for all pairs $(E'', E) \in \mathbf{A}$, since the parent set of E has changed. By directing evidence arcs to attributes that show high mutual information with the class variable, we enforce that the resulting graph remains directed and acyclic. The full algorithm is shown in Algorithm 7.1.

Algorithm 7.1 MMI construction algorithm.

input: empty graph G, database $\mathcal{D}$,
class variable C, evidence variables $\mathbf{E}$, number of required arcs M
$\mathbf{S} \leftarrow \mathbf{E}$
$\mathbf{A} \leftarrow$ an initially empty set of pairs of evidence variables
for $i = 0$ to M **do**
  let $E = \arg\max_{E' \in \mathbf{S}} \{I(C, E')\}$
  let $(E_1, E_2) = \arg\max_{(E', E) \in \mathbf{A}} \{I(E', E \mid \pi(E))\}$
  if $I(C, E) > I(E_1, E_2 \mid \pi(E_2))$ **then**
    $\mathbf{S} \leftarrow \mathbf{S} \setminus \{E\}$
    $A(G) \leftarrow A(G) \cup \{(C, E)\}$
    for all $E' \in \mathbf{S}$ **do**
      $\mathbf{A} \leftarrow \mathbf{A} \cup \{(E', E)\}$
    end for
  else
    $A(G) \leftarrow A(G) \cup \{(E_1, E_2)\}$
    $\mathbf{A} \leftarrow \mathbf{A} \setminus \{(E_1, E_2)\}$
  end if
end for
return G

Figure 7.2 shows an example of how the algorithm builds a Bayesian classifier structure. The final structure incorporates feature selection, orientation of arcs in the direction of evidence variables that show high mutual information with the class variable, and the encoding of a third-order dependency $P(E_2 \mid C, E_1, E_3)$.

Looking back at Eq. (7.3) a possible complication is identified. Since the set $\pi(E)$ of an evidence variable E may grow indefinitely and the number of parent configurations grows exponentially with n, the network may become victim of its own unrestrictedness. Note that since one has a finite (and often small) database at ones disposal, this means that the actual conditional probability $P(E \mid \pi(E))$ will become

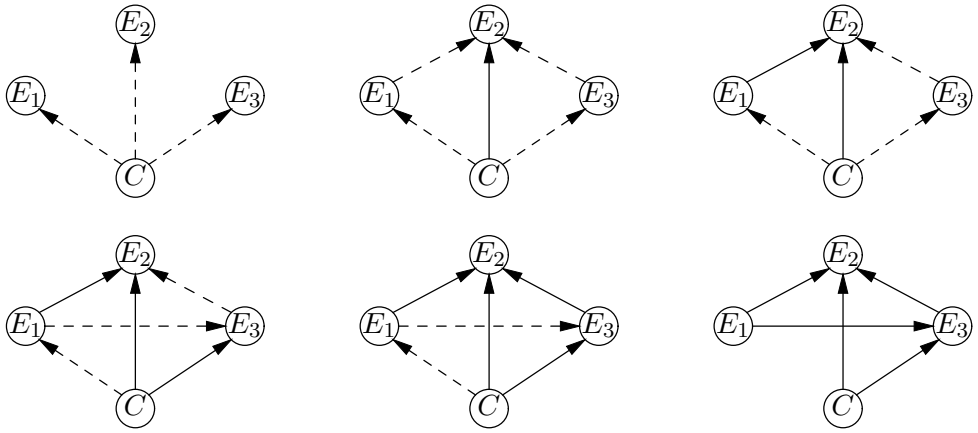


Figure 7.2: An example of the MMI algorithm building a Bayesian classifier structure from the top left to the bottom right with dashed arrows representing candidate dependencies.

increasingly inaccurate when the number of parents grows; configurations associated with large parent-sets cannot be reliably estimated from moderate size databases, introducing what may be termed *spurious dependencies*. In order to prevent the occurrence of spurious dependencies, we make use of the following heuristic. We use

$$\tilde{P}(E, E' \mid \mathbf{a}) = \frac{N_{\mathbf{a}}}{N_{\mathbf{a}} + \beta} P(E, E' \mid \mathbf{a}) + \frac{\beta}{N_{\mathbf{a}} + \beta} P(E \mid \mathbf{a}) P(E' \mid \mathbf{a}) \qquad (7.4)$$

as the expression for the conditional probability of E and E' given that $\mathbf{A} = \mathbf{a}$, *during the computation of conditional mutual information* according to Eq. (7.3). In Eq. (7.4), $N_{\mathbf{a}}$ is the number of times the configuration $\mathbf{a}$ occurs in $\mathcal{D}$, and β is a parameter that we choose as $\beta = 500$ throughout our experiments, unless indicated otherwise. $P(E, E' \mid \mathbf{a})$, $P(E \mid \mathbf{a})$, and $P(E' \mid \mathbf{a})$ are computed from $\mathcal{D}$ and smoothed using Laplace smoothing. This heuristic ensures that the conditional mutual information computed according to Eq. (7.3) will be small when the number of occurrences $N_{\mathbf{a}}$ of the conditioning case is small, since, in the limit $N_{\mathbf{a}} \to 0$, we obtain

$$\begin{aligned} I(E, E' \mid \mathbf{A}) &= \sum_{e, e', \mathbf{a}} P(e, e', \mathbf{a}) \log \frac{\tilde{P}(e, e' \mid \mathbf{a})}{P(e \mid \mathbf{a}) P(e' \mid \mathbf{a})} \\ &= \sum_{e, e', \mathbf{a}} P(e, e', \mathbf{a}) \log \frac{P(e \mid \mathbf{a}) P(e' \mid \mathbf{a})}{P(e \mid \mathbf{a}) P(e' \mid \mathbf{a})} = 0. \end{aligned}$$

7.1.3 The COMIK dataset

In order to validate classifier performance we made use of the COMIK dataset, which was collected by the Copenhagen Computer Icterus (COMIK) group and consists of

data on 1002 jaundiced patients. The COMIK group has been working for over a decade on the development of a system for diagnosing liver and biliary disease which is known as the Copenhagen Pocket Diagnostic Chart (Malchow-Møller et al., 1986). Using a set $\mathbf{E}$ of 21 evidence variables, the system classifies patients into one of four diagnostic categories: *acute non-obstructive*, *chronic non-obstructive*, *benign obstructive* and *malignant obstructive*. The chart offers a compact representation of three logistic regression equations, where the probability of *acute obstructive jaundice*, for instance, is computed as follows: $P(\textit{acute obstructive jaundice} \mid \mathbf{E}) = P(\textit{acute} \mid \mathbf{E}) \cdot P(\textit{obstructive} \mid \mathbf{E})$. The performance of the system has been studied using retrospective patient data and it has been found that the system is able to produce a correct diagnostic conclusion (in accordance with the diagnostic conclusion of expert clinicians) in about $75 - 77\%$ of jaundiced patients (Lindberg et al., 1987).

7.1.4 Classification results and network interpretation

In this section we will demonstrate the usefulness of the β parameter that was introduced in Eq. (7.4), compare the classification performance of both the FAN and MMI classifiers on the COMIK dataset and give a medical interpretation of the resulting structures.

Table 7.1: Effects of varying parameter β for a model consisting of 30 arcs.

β	$\eta(\mathcal{D})$	$\mathrm{F}(\mathcal{B})$	β	$\eta(\mathcal{D})$	$\mathrm{F}(\mathcal{B})$	β	$\eta(\mathcal{D})$	$\mathrm{F}(\mathcal{B})$
1	74.75 %	87	102	75.95 %	65	800	76.25 %	59
4	74.75 %	77	290	75.95 %	63	900	76.25 %	59
36	74.85 %	71	610	75.95 %	61	2000	76.25 %	57
56	75.15 %	67	660	76.25 %	61			

First we present the results of varying the parameter β in order to determine whether this has an effect on the classification performance and network structure of our classifiers. To this end, we focused on a Bayesian classifier $\mathcal{B} = (G, P)$ that allows 30 arcs (the parameter M in Algorithm 7.1). For this classifier, we have determined the classification accuracy, and summed squared fan-in

$$\mathrm{F}(\mathcal{B}) = \sum_{i \in V(G)} |\pi(i)|^2 \,,$$

where $|\pi(i)|$ denotes the cardinality of the parent set of a vertex i. Table 7.1 clearly shows that the summed squared fan-in decreases when β increases; indicating that spurious dependencies are removed. This removal also has a beneficial effect on the classification accuracy, which rises from 74.75% for $\beta = 1$ to 76.25% for $\beta = 660$.

We have compared the performance of the MMI algorithm with that of the FAN algorithm, using leave-one-out cross-validation, using $\beta = 500$. Figure 7.3 shows

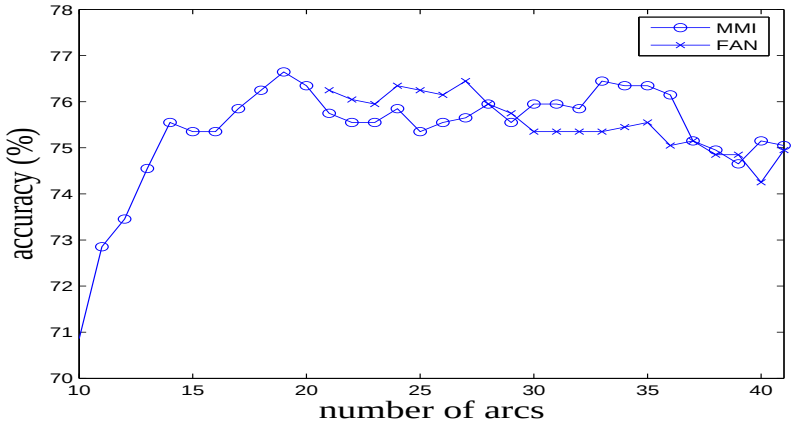


Figure 7.3: Classification accuracy for Bayesian classifiers with a varying number of arcs learnt using the FAN algorithm or the MMI algorithm for the COMIK dataset.

that both algorithms perform comparably and within the bounds of the Copenhagen Pocket Diagnostic Chart. Both algorithms show a small performance decrease for dense network structures, which may be explained in terms of overfitting artifacts. Maximal classifier accuracy for the MMI algorithm is 76.65% for a network of 19 arcs versus 76.45% for a network of 27 arcs for the FAN algorithm.

In terms of classifier structure, one can observe that both algorithms represent similar dependencies, with the difference that those of the MMI algorithm form a subset of those of the FAN algorithm. The best FAN classifier has a structure with an arc from the class variable to every evidence variable and the following arcs between evidence variables: *biliary-colics-gallstones* $\rightarrow$ *upper-abdominal-pain* $\rightarrow$ *leukemia-lymphoma* $\rightarrow$ *gall-bladder*, *history-ge-2-weeks* $\rightarrow$ *weight-loss*, *ascites* $\rightarrow$ *liver-surface* and *ASAT* $\rightarrow$ *clotting-factors*. The MMI algorithm has left *leukemia-lymphoma,congestive-heart-failure* and *LDH* independent of the class-variable and shows just the dependency *liver-surface* $\rightarrow$ *ascites* between evidence variables.

Given our aim of learning Bayesian classifiers that not only display good classification performance, but are comprehensible to medical doctors as well, we have carried out a qualitative comparison between two of the Bayesian networks learned from the COMIK data: Figure 7.4 shows a FAN classifier which was learned using the FAN algorithm described previously (Lucas, 2004), whereas Figure 7.5 shows an MMI network with the same number of arcs. Clearly, the restriction imposed by the FAN algorithm that the arcs between evidence variables form a forest of trees does have implications with regard to the understandability of the resulting networks. Yet, parts of the Bayesian network shown in Figure 7.4 can be given a clinical interpretation. Similar remarks can be made for the MMI network, although one would hope that giving an interpretation is at least somewhat easier.

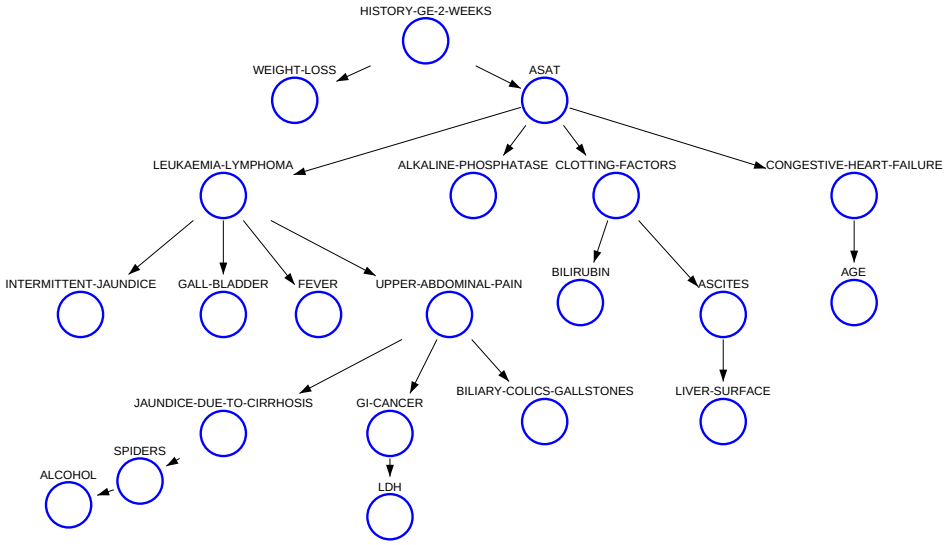


Figure 7.4: Arcs between evidence variables for a FAN classifier containing 41 arcs. The class variable was connected with all evidence variables (not shown).

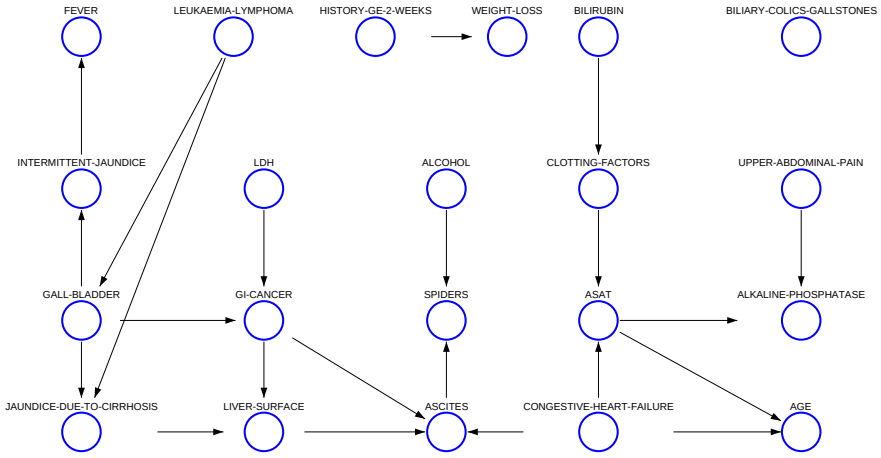


Figure 7.5: Arcs between evidence variables for an MMI classifier containing 41 arcs. The class variable was connected with all evidence variables (not shown).

If we ignore the arcs between the class vertex and the evidence vertices, there are 20 arcs between evidence vertices in the FAN and 22 arcs between evidence vertices in the MMI network. Ignoring arc orientation, 9 of the arcs in the MMI network are

shared by the FAN classifier. As the choice of the direction of arcs in the FAN is arbitrary, it is worth noting that in 4 of these arcs the direction is different; in 2 of these arcs it is medically speaking impossible to establish the right direction of the arcs, as hidden variables are involved, in 1 the arc direction is correct (*congestive-heart-failure* → *ASAT*), whereas in the remaining arc (*GI-cancer* → *LDH*) the direction is incorrect. Some of the 13 non-shared arcs of the MMI network have a clear clinical interpretation. For example, the arcs *GI-cancer* → *ascites*, *congestive-heart-failure* → *ascites* and *GI-cancer* → *liver-surface* are arcs that can be given a causal interpretation, as gastrointestinal (GI) cancer and right-heart failure do give rise to the accumulation of fluid in the abdomen (i.e. ascites), and liver metastases due to GI cancer may change the liver surface. Observe that the multiple causes of ascites cannot be represented in the FAN due to its structural restrictions. The path *gall-bladder* → *intermittent-jaundice* → *fever* in the MMI network offers a reasonably accurate picture of the course of events of the process giving rise to fever; in contrast, the situation depicted in the FAN, where *leukemia-lymphoma* acts as a common cause, does not reflect clinical reality. However, the arc from *upper-abdominal-pain* to *biliary-colics-gallstones* in the FAN, which is correct, is missing in the MMI network. Overall, the MMI network seems to reflect clinical reality somewhat better than the FAN, although not perfectly.

7.2 Decomposed tensor classifiers

In this section, we present a novel probabilistic classification technique which is based on the decomposition of a multiway array, also known as a *tensor* (de Lathauwer, 1997), by means of a set of components, often taking the form of vectors. We call classifiers that use this technique *decomposed tensor classifiers*, and test their performance by means of a database that contains data about 1002 patients that present with hepatic disease. The goal is to diagnose the correct disease for each of the patients from a set of four distinct diseases. The performance of this new technique is analyzed and compared with the performance of the naive Bayes classifier (Maron, 1961).

We proceed as follows. In Sections 7.2.1 and 7.2.2 the theoretical background of tensors and their decompositions is described. Subsequently, in Section 7.2.3, we address how tensor decompositions can be used for probabilistic classification. The clinical database and the techniques used to evaluate classification performance are described in Section 7.2.5. We end with an analysis of the experimental results in Section 7.2.6.

7.2.1 Tensors

A tensor is a concept taken from multilinear algebra which generalizes the concepts of vectors and matrices, and is defined as follows.

Definition 7.1. *Let* $I_1, \ldots, I_N \in \mathbb{N}$ *denote index upper bounds. A* tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ *is an N-way array where elements* $a_{i_1 \cdots i_n}$ *are indexed by* $i_j \in \{1, \ldots, I_j\}$ *for* $1 \leq j \leq N$.

We call N the *order* of a tensor, such that a tensor of order one denotes a vector $\mathbf{a} \in \mathbb{R}^{I_1}$, and a tensor of order two denotes a matrix $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$. The nth *mode* of a tensor refers to the nth dimension of a tensor. A tensor can be expressed in terms of a matrix using the concept of a matrix unfolding.

Definition 7.2. *The* matrix unfolding $\mathbf{A}_{(j)} \in \mathbb{R}^{I_j \times (I_{j+1} I_{j+2} \cdots I_N I_1 I_2 \cdots I_{j-1})}$ *of an* N*th order tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ *is the matrix that has element* $a_{i_1 \cdots i_N}$ *at row number* i_j *and column number*

$$1 + \sum_{\substack{1 \leq k \leq N \\ k \neq j}} (i_k - 1) \prod_{\substack{k+1 \leq m \leq N \\ m \neq j}} I_m \, .$$

Example 7.1. The matrix unfolding $\mathbf{A}_{(2)}$ of a third-order tensor

$$\mathcal{A} = \begin{pmatrix} (a,b)^T & (c,d)^T \\ (e,f)^T & (g,h)^T \end{pmatrix} \text{ is given by } \mathbf{A}_{(2)} = \begin{pmatrix} a & b & e & f \\ c & d & g & h \end{pmatrix} .$$

A tensor may be multiplied by a matrix by means of the *n-mode product*.

Definition 7.3. *The* n-mode product $\mathcal{A} \times_n \mathbf{B}$ *of a tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ *and a matrix* $\mathbf{B} \in \mathbb{R}^{J_N \times I_N}$*, is a tensor* $\mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ *with elements:*

$$c_{i_1 \cdots i_{n-1} j_n i_{n+1} \cdots i_N} = \sum_{i_n} a_{i_1 \cdots i_N} b_{j_n i_n} \, .$$

Example 7.2. Let $\mathcal{A}$ be a third-order tensor as in example 7.1 and let $\mathbf{B}$ denote a square matrix with $b_{11} = u$, $b_{12} = v$, $b_{21} = w$, $b_{22} = x$. The 2-mode product $\mathcal{A} \times_2 \mathbf{B}$ is then given by

$$\begin{pmatrix} (a(u+v), b(u+v))^T & (c(w+x), d(w+x))^T \\ (e(u+v), f(u+v))^T & (g(w+x), h(w+x))^T \end{pmatrix} .$$

We also define, for tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, the *inner product*

$$\langle \mathcal{A}, \mathcal{B} \rangle \equiv \sum_{i_1, \ldots, i_N} a_{i_1 \cdots i_N} b_{i_1 \cdots i_N}$$

and *Frobenius norm* $||\mathcal{A}|| \equiv \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$. The *outer product* $\mathcal{A} \circ \mathcal{B}$ of two tensors $\mathcal{A} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ and $\mathcal{B} \in \mathbb{R}^{J_1 \times \cdots \times J_n}$ is defined as the tensor $\mathcal{C} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times J_1 \times \cdots \times J_n}$ such that $c_{i_1 \cdots i_n j_1 \cdots j_n} = a_{i_1 \cdots i_n} \cdot b_{j_1 \cdots j_n}$ for all elements of $\mathcal{C}$. The rank of a tensor is then defined as follows (Håstad, 1990).

Definition 7.4. *A tensor of order N has* rank *one if it can be written as an outer product $\mathbf{a}^{(1)} \circ \cdots \circ \mathbf{a}^{(N)}$ of vectors. The rank of a tensor $\mathcal{A}$ is defined as the minimal number of tensors $\mathcal{A}_1, \ldots, \mathcal{A}_K$ of rank one such that*

$$\mathcal{A} = \sum_{k=1}^{K} \mathcal{A}_k \,. \tag{7.5}$$

Example 7.3. The third-order tensor

$$\mathcal{A} = \begin{pmatrix} (\ \ 6, -3)^T & (\ \ 8, -4)^T \\ (-12,\ \ 6)^T & (-16,\ \ 8)^T \end{pmatrix}$$

has rank one since it can be written as the outer product of $(1, -2)^T$, $(3, 4)^T$, and $(2, -1)^T$.

7.2.2 Tensor decompositions

Equation (7.5) is known as a *rank-K decomposition* of $\mathcal{A}$. A more general kind of decomposition is the *Tucker decomposition* (Tucker, 1966), which can be interpreted as a multilinear formulation of the singular value decomposition (de Lathauwer et al., 2000a):

$$T_{\mathbf{J}}(\mathcal{A}) = \mathcal{C} \times_1 \mathbf{B}^{(1)} \times_2 \cdots \times_N \mathbf{B}^{(N)} \tag{7.6}$$

with $\mathbf{J} = (J_1, \ldots, J_N)$, *core tensor* $\mathcal{C} = (c_{j_1 \cdots j_N})$ and matrices $\mathbf{B}^{(n)} \in \mathbb{R}^{I_n \times J_n}$. Elements of $\mathcal{A}$ are then computed as follows:

$$a_{i_1 \cdots i_N} = \left(\sum_{j_1, \ldots, j_N} c_{j_1 \cdots j_N} \cdot b^{(1)}_{i_1 j_1} \cdots b^{(N)}_{i_N j_N} \right) + r_{i_1 \cdots i_N} \,, \tag{7.7}$$

where $(r_{i_1 \cdots i_N})$ denotes a residual tensor $\mathcal{R}$. The parameters of the Tucker decomposition can be found using *higher-order orthogonal iteration* (de Lathauwer et al., 2000b). A special case of the Tucker decomposition is obtained when one assumes that the core tensor $\mathcal{C}$ is a superdiagonal tensor with $c_{j_1 \cdots j_N} = 0$ if there are $u, v \in \{1, \ldots, N\}$ such that $j_u \neq j_v$. Hence, we obtain:

$$a_{i_1 \cdots i_N} = \left(\sum_{k=1}^{K} \lambda_k \cdot b^{(1)}_{i_1 k} \cdots b^{(N)}_{i_N k} \right) + r_{i_1 \cdots i_N} \tag{7.8}$$

for some suitably chosen K. Equation (7.8) is known as the *canonical decomposition* (Carroll and Chang, 1970), or *parallel factors decomposition* (Harshman, 1970). In general, the decomposition of Eq. (7.8) is not necessarily minimal nor exact, and can be interpreted as a sum of rank-1 approximations. One way of finding a rank-1 approximation is by means of the *higher-order power method* (HOPM) (de Lathauwer et al., 2000b), as shown in Algorithm 7.2.

Algorithm 7.2 Higher-Order Power Method (HOPM).

input: $\mathcal{A}$
initialize $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(N)}$
repeat
  for $n = 1$ to N **do**
    $\tilde{\mathbf{b}}^{(n)} = \mathcal{A} \times_1 \mathbf{b}^{(1)^T} \times_2 \cdots \times_{n-1} \mathbf{b}^{(n-1)^T} \times_{n+1} \mathbf{b}^{(n+1)^T} \times_{n+2} \cdots \times_N \mathbf{b}^{(N)^T}$
    $\lambda_n = ||\tilde{\mathbf{b}}^{(n)}||$
    $\mathbf{b}^{(n)} = \tilde{\mathbf{b}}^{(n)} / \lambda_n$
  end for
until convergence
return $\hat{\mathcal{A}} = \lambda_N \cdot \mathbf{b}^{(1)} \circ \cdots \circ \mathbf{b}^{(N)}$

The higher-order power method finds a tensor $\hat{\mathcal{A}} = \lambda \cdot \mathbf{b}^{(1)} \circ \cdots \circ \mathbf{b}^{(N)}$, with scalar λ and unit-norm vectors $\mathbf{b}^{(n)}$, $1 \leq n \leq N$, that minimizes the least-squares cost function $C(\mathcal{A}, \hat{\mathcal{A}}) \equiv ||\mathcal{A} - \hat{\mathcal{A}}||^2$. A greedy approach to finding the sum of rank-1 terms in Eq. (7.8) is to apply the higher-order power method to the residuals that remain after obtaining a rank-1 approximation. This technique has been employed successfully in Ref. (Wang and Ahuja, 2004) in order to achieve high compression rates for image sequences. By defining $\mathcal{A}^1 \equiv \mathcal{A}$ and $\mathcal{A}^k \equiv \mathcal{A}^{k-1} - \text{HOPM}(\mathcal{A}^{k-1})$ the following rank-K approximation of a tensor $\mathcal{A}$ is obtained:[1]

$$R_K(\mathcal{A}) \equiv \sum_{k=1}^{K} \text{HOPM}(\mathcal{A}^k) \,. \tag{7.9}$$

In order to initialize matrices and vectors in Algorithm 7.2, various schemes can be used. One approach is to repeat the algorithm for several random initializations and to choose that decomposition which maximizes the fit between the original tensor and the approximation. Another approach, which has proven to work well in practice, is to choose the first dominant left singular vector of the matrix unfolding $\mathbf{A}_{(j)}$, as an initial estimate of $\mathbf{b}^{(j)}$ (de Lathauwer et al., 2000b,a). The algorithm has converged when the increase in fit between the tensor and its approximation that is gained after one iteration drops below a small error criterion ϵ. In the following section, we will use decompositions of tensors $\mathcal{A} \in [0,1]^{I_1 \times \cdots \times I_N}$ for the task of probabilistic classification.

7.2.3 Classification with tensor decompositions

In this section, we focus on a multiset $\mathbf{A} = \{\mathbf{a}^1, \ldots, \mathbf{a}^n\}$ that represents our data, and where an instance $\mathbf{a}^i = (x_1^i, \ldots, x_N^i)$ consists of evidence $(x_1^i, \ldots, x_{N-1}^i)$ and a class label x_N^i. We assume that all variables are discrete and use I_j with $1 \leq j \leq N$

[1] This procedure is only guaranteed to find the optimal rank-K approximation if the tensor $\mathcal{A}$ is orthogonally decomposable (Zhang and Golub, 2001).

to denote the finite number of values x_j of a variable X_j. The basic idea is to obtain an approximation of an *incomplete* tensor $\mathcal{A}$ using a tensor decomposition. Let $\mathbf{x}$ denote the evidence and let $n(\mathbf{x}, x_N)$ stand for the number of times $(\mathbf{x}, x_N)$ occurs in $\mathbf{A}$. We transform $\mathbf{A}$ into an incompletely specified tensor $\mathcal{A} \in [0,1]^{I_1 \times \cdots \times I_N}$, such that

$$a_{x_1 \cdots x_N} = \frac{1}{n} n(\mathbf{x}, x_N) \tag{7.10}$$

for all $(\mathbf{x}, x_N)$ for which some $(\mathbf{x}, j)$ with $1 \leq j \leq I_N$ occurs in $\mathbf{A}$. Hence, $a_{x_1 \cdots x_N}$ is undefined for unseen evidence $\mathbf{x}$ (as indicated by $*$), which implies that the tensor is incomplete. The element $a_{x_1 \cdots x_N}$ is used to represent an estimate of the joint probability $P(\mathbf{x}, x_N)$. For incomplete tensors, we interpret undefined elements as zero in Algorithm 7.2. Since zero elements have no contribution, we may use a sparse representation of tensors $\mathcal{A} \in [0,1]^{I_1 \times \cdots \times I_N}$ with very large N, provided that only some of the elements are defined.

Example 7.4. Consider a dataset $\mathbf{A} = \{(1,1,1),\ (1,2,1),\ (1,2,1),\ (1,2,2),\ (2,1,2),\ (1,1,2)\}$. By applying the transformation (7.10) to the example dataset, we obtain the third-order tensor

$$\mathcal{A} = \begin{pmatrix} (\frac{1}{6}, \frac{1}{6})^T & (\frac{2}{6}, \frac{1}{6})^T \\ (\frac{0}{6}, \frac{1}{6})^T & (*, *)^T \end{pmatrix}.$$

The basic idea is then to obtain an approximation of an *incomplete* tensor $\mathcal{A}$ using a tensor decomposition. For incomplete tensors, we may use a sparse representation, by interpreting undefined elements as zero in Algorithm 7.2. Since zero elements have no contribution, we may use tensors $\mathcal{A} \in [0,1]^{I_1 \times \cdots \times I_N}$ with very large N, provided that only some of the elements are defined.

Example 7.5. When applying the higher-order power method to our exemplary tensor, we have the following matrix unfoldings:

$$\mathbf{A}_{(1)} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{2}{6} & \frac{1}{6} \\ \frac{0}{6} & \frac{1}{6} & * & * \end{pmatrix} \quad \mathbf{A}_{(2)} = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{0}{6} & \frac{1}{6} \\ \frac{2}{6} & \frac{1}{6} & * & * \end{pmatrix} \quad \mathbf{A}_{(3)} = \begin{pmatrix} \frac{1}{6} & \frac{2}{6} & \frac{0}{6} & * \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & * \end{pmatrix}$$

Assuming that $\mathbf{b}^{(1)} = (a,b)^T, \mathbf{b}^{(2)} = (c,d)^T$, and $\mathbf{b}^{(3)} = (e,f)^T$, one cycle of Algorithm 7.2 for variable A_1 would give us

$$\begin{aligned} \tilde{\mathbf{b}} &= \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{2}{6} & \frac{1}{6} \\ \frac{0}{6} & \frac{1}{6} & * & * \end{pmatrix} \times_2 (c,d)^T \times_3 (e,f)^T \\ &= \begin{pmatrix} \frac{1}{6}ce + \frac{1}{6}cf + \frac{2}{6}de + \frac{1}{6}df \\ \frac{0}{6}ce + \frac{1}{6}cf + *de + *df \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{6}ce + \frac{1}{6}cf + \frac{2}{6}de + \frac{1}{6}df \\ \frac{1}{6}cf \end{pmatrix} \end{aligned}$$

with $\lambda_1 = ||\tilde{\mathbf{b}}_1|| = \sqrt{(\frac{1}{6}ce + \frac{1}{6}cf + \frac{2}{6}de + \frac{1}{6}df)^2 + (\frac{1}{6}cf)^2}$. Hence, zero and undefined elements have no effect.

In case of probabilistic classification, our interest is in computing the posterior probability $P(x_N \mid \mathbf{x})$ based on our estimate of $P(\mathbf{x}, x_N)$. Although $P(\mathbf{x}, x_N)$ is approximated by $R_K(\mathcal{A})_{x_1 \cdots x_N}$, we have no guarantee that the tensor approximation represents a proper probability distribution for unseen evidence (which is the goal of probabilistic classification), since the approximation may be unnormalized or even lying outside the unit interval. Therefore, we use the following transform when computing the conditional probability of X_N given $\mathbf{x}$:

$$P(x_N \mid \mathbf{x}) = \frac{R_K^+(\mathcal{A})_{x_1 \cdots x_N}}{\sum_{1 \leq j \leq I_N} R_K^+(\mathcal{A})_{x_1 \cdots x_{N-1} j}} \tag{7.11}$$

where

$$R_K^+(\mathcal{A})_{x_1 \cdots x_N} \equiv R_K(\mathcal{A})_{x_1 \cdots x_N} - \min \left\{ 0, \min_j (R_K(\mathcal{A})_{x_1 \cdots x_{N-1}, j}) \right\}$$

ensures that we sum over positive terms by making (small) negative terms non-negative. Alternatively, a log transform together with a suitable prior may be used in order to guarantee that we obtain a proper conditional probability distribution. However, experiments in that direction led to less optimal classification results.

We use the term *decomposed tensor classifier* to denote a classifier that uses the approximation $R_K(\mathcal{A})_{x_1 \cdots x_N}$ for the purpose of classification. In this chapter, we use the rank-K approximation, although other tensor decompositions such as the Tucker decomposition could also be used. Furthermore, we require that variables are discrete (or discretized a priori), and data is complete (or completed using an imputation scheme). The classification procedure is shown in Algorithm 7.3.

Algorithm 7.3 Decomposed tensor classification.

input: $\mathbf{A}_{\text{train}}$, $\mathbf{A}_{\text{test}}$, R_K
transform the dataset $\mathbf{A}_{\text{train}}$ into the tensor $\mathcal{A}_{\text{train}}$ using Eq. (7.10)
learn the approximation $R_K(\mathcal{A}_{\text{train}})$ using Algorithm 7.2
for all rows $(\mathbf{x}) \in \mathbf{A}_{\text{test}}$ **do**
  for $j = 1$ **to** I_N **do**
    compute $P(j \mid \mathbf{x})$ using Eq. (7.11)
  end for
  assign class label $\mathcal{L}(\mathbf{x}) = \arg\max_j \{P(j \mid \mathbf{x})\}$
end for
return class labels $\mathcal{L}$

7.2.4 Graphical model interpretation

If the approximation $R_K(\mathcal{A})_{i_1 \cdots i_N}$ is exact, then we may interpret a rank-K approximation in terms of a graphical model structure, as noticed in (Savický and Vomlel, 2006), and shown in Fig. 7.6.

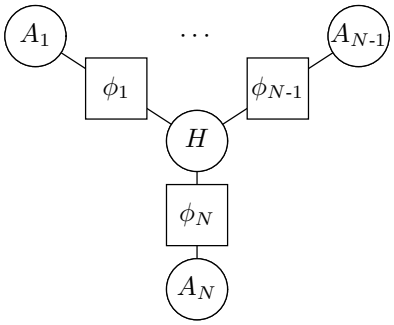


Figure 7.6: Representation of a tensor rank-K approximation as a graphical model, with (possibly negative) real-valued functions ϕ_j, and where the hidden variable H has states $\{1, \ldots, K\}$.

According to Eq. (7.8), a rank-K approximation can be written as

$$R_K(\mathcal{A})_{x_1 \cdots x_N} = \sum_{h=1}^{K} \lambda_h \cdot b^{(1)}_{x i_1 h} \cdots b^{(N)}_{i_N k} \,. \tag{7.12}$$

We define functions $\phi_j(x_j, h) \equiv b^{(j)}_{x_j h}$ for $1 \leq j < N$ and absorb λ into the function $\phi_N(x_N, h) \equiv \lambda_h \cdot b^{(N)}_{x_N h}$. We now define

$$P(x_1, \ldots, x_N, h) = \frac{1}{Z} \prod_{j=1}^{N} \phi_j(x_j, h) \tag{7.13}$$

with partition function $Z \equiv \sum_{x_1, \ldots, x_N, h} \prod_{j=1}^{N} \phi_j(x_j, h)$ as the joint probability distribution for random variables $X_1, \ldots, X_N, H$. Equation (7.12) can then be interpreted as marginalization over the hidden variable H:

$$P(x_1, \ldots, x_N) = \frac{1}{Z} \sum_{h} \prod_{j=1}^{N} \phi_j(x_j, h) \,, \tag{7.14}$$

and the computation of $P(x_N \mid \mathbf{x})$ can therefore be interpreted in terms of probabilistic inference in the graphical model of Fig. 7.6.

7.2.5 Classifier evaluation

In order to examine the performance of decomposed tensor classifiers, we have made use of the COMIK dataset, which was collected by the Copenhagen Computer Icterus (COMIK) group and consists of data on 1002 jaundiced patients that may be classified into one of four diagnostic categories: *acute non-obstructive*, *chronic non-obstructive*, *benign obstructive* and *malignant obstructive* given 21 evidence variables (Malchow-Møller et al., 1986). Earlier classification studies have shown that, typically, the correct diagnostic conclusion (in accordance with the diagnostic conclusion of expert clinicians) is found for about $75 - 77\%$ of jaundiced patients (Lindberg et al., 1987; van Gerven and Lucas, 2004a). As a preprocessing step, we have computed the mutual information between evidence variables and the class variable, and selected the eighteen evidence variables that show highest mutual information (MI) with the class variable as the basis for classification, since the three remaining evidence variables give relatively small contributions (Fig. 7.7).

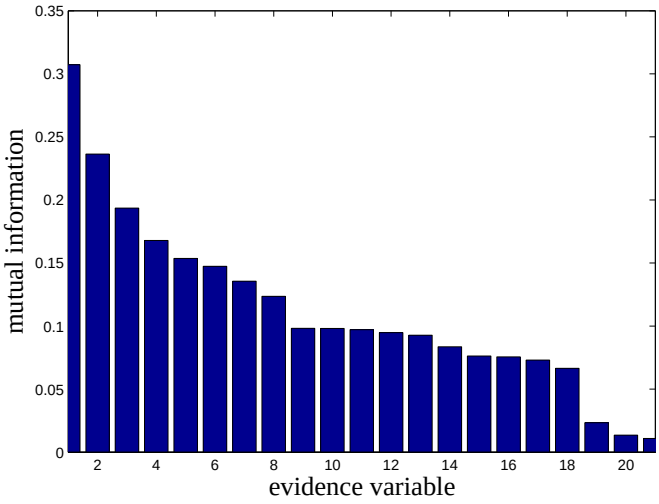


Figure 7.7: Mutual information between the class variable and evidence variables.

Classification performance of the decomposed tensor classifiers is compared with that of a naive Bayes classifier using a ten-fold cross-validation scheme. Empirical estimates of the required probabilities are smoothed using Laplace smoothing. The naive Bayes classifier typically reaches high classification accuracies, and uses the (naive) assumption that evidence variables are independent given the class label:

$$P(x_N \mid \mathbf{x}) \propto P(x_N) \prod_{j=1}^{N-1} P(x_j \mid x_N) \,.$$

Since the COMIK dataset contains missing values, and the decomposed tensor classifiers require complete data, we have used multiple imputation to create three complete datasets from the incomplete dataset. Since we have no knowledge about the missing data mechanism, we make the (admittedly unrealistic) assumption that data is missing completely at random, and use the prior probabilities of the evidence variables to determine the imputed values. This allows a comparison in terms of classification performance between the naive Bayes classifier and the decomposed tensor classifiers, where performance is averaged over folds and datasets.

Classification performance is quantified by means of classification accuracy and logarithmic score. Classification accuracy is defined as the percentage of correctly classified cases, as in (7.1). The logarithmic score (Spiegelhalter et al., 1993) is a scoring rule which penalizes a probability model based on a database consisting of m instances $(\mathbf{x}^i, x_N^i)$ where $\mathbf{x}^i$ denotes the evidence and x_N^i denotes the class value. Assuming that instances are independently sampled and identically distributed, the logarithmic score is defined as:

$$S = -\sum_{i=1}^{m} \log P(x_N^i \mid \mathbf{x}^i)$$

which incurs a penalty if a low probability is assigned to events that actually occur. The logarithmic score of the decomposed tensor classifier is compared with that of the naive Bayes classifier in order to determine how well actual posterior probabilities are approximated.

7.2.6 Experimental results

In order to use the rank-K approximation for classification, the first question is which initialization procedure to use in Algorithm 1. Therefore, we have conducted a preliminary experiment in order to compare different initialization schemes in terms of classification accuracy and least squares error. To this end, we have chosen the five most informative evidence variables as the basis for classification, and compared the performance on the test set of classifiers R_K, with $1 \leq K \leq 30$, for 1, 5, and 10 random initializations, and for the initialization with dominant left singular vectors.

The results shown in Fig. 7.8 indicate that there is not much difference in classification accuracy or least squares error for the different initialization schemes. Differences in standard deviations were also negligible (not shown). Therefore, we have chosen to use just one random initialization since this uses the least computational resources.

We have learnt decomposed tensor classifiers based on the eighteen most informative evidence variables for $1 \leq K \leq 30$ components. The comparison of the classification accuracy of the decomposed tensor classifier with that of the naive Bayes classifier is shown in Fig. 7.9. The highest average accuracy for the decomposed tensor classifier is reached at nineteen components with an accuracy of 76.75%, whereas

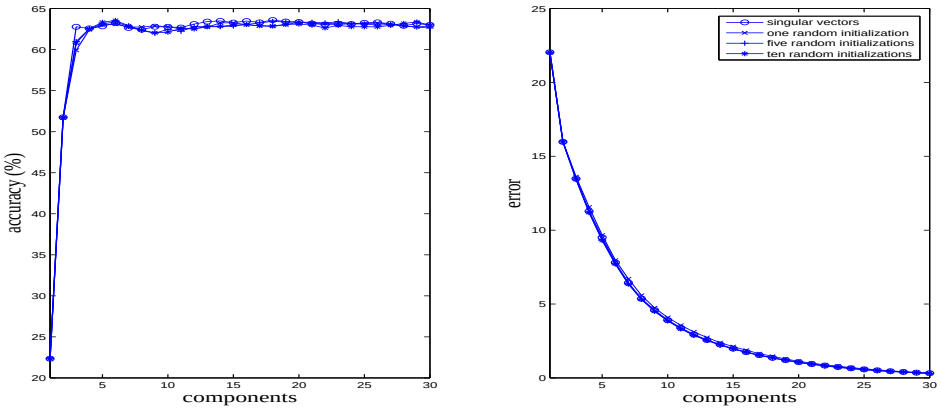


Figure 7.8: Average classification accuracy on the test set (left) and least squares error of the tensor approximation (right) based on five evidence variables with different initializations.

for the naive Bayes classifier, the average classification accuracy is 77.25%. At that point, the standard deviation of the classification accuracy of the decomposed tensor classifier is 3.24%, whereas that of the naive Bayes classifier is 3.40%. Although the naive Bayes classifier performs somewhat better than the decomposed tensor classifier in terms of classification accuracy, differences are negligible.

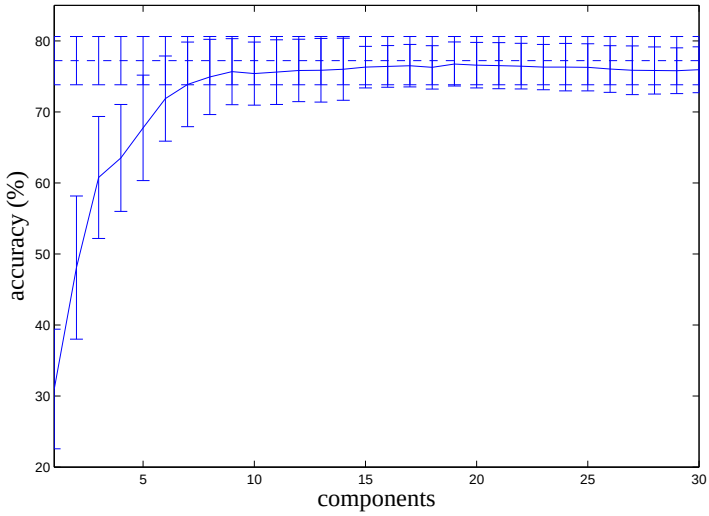


Figure 7.9: Average classification accuracy and standard deviations on the test set for the decomposed tensor classifier (solid line) and the naive Bayes classifier (dashed line).

Figure 7.10 depicts the average logarithmic scores for the decomposed tensor classifier and the naive Bayes classifier (where we have added a small term to Eq. (7.11) in order to prevent numerical problems). It shows that the logarithmic score of the decomposed tensor classifier decreases as more components are added and eventually becomes lower than that of the naive Bayes classifier.[2]

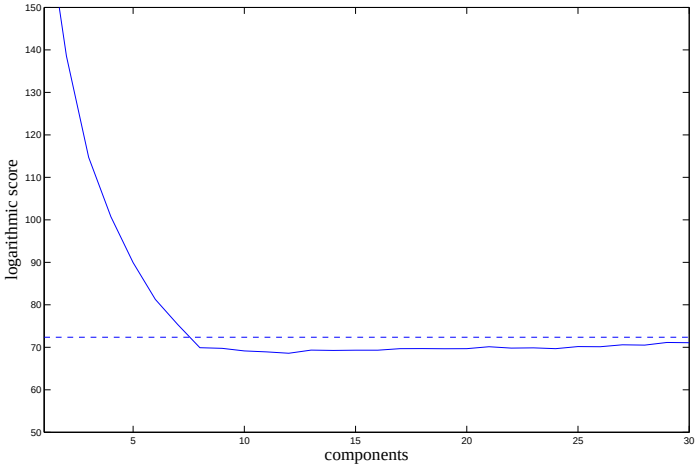


Figure 7.10: Average logarithmic score on the test set for the decomposed tensor classifier (solid line) and the naive Bayes classifier (dashed line).

Figure 7.11 shows a Hinton diagram, depicting the contribution of each component for each of the four classes for a decomposed tensor classifier containing nineteen components. The large white block that can be found in each column indicates that each of the components improves the approximation by focusing mainly on one class.

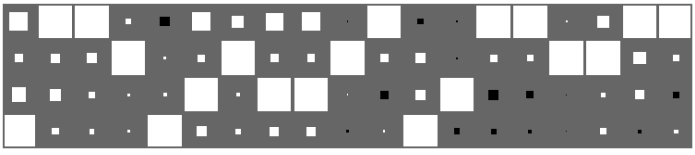


Figure 7.11: Hinton diagram, showing the magnitude of positive contributions (white blocks) and negative contributions (black blocks) of nineteen rank-1 components (horizontal axis) for the four classes (vertical axis).

For the decomposed tensor classifier, the transform of Eq. (7.11) assigns distributions skewed towards zero for incorrect classes and skewed towards one for the

[2] In practice, the appropriate number of components is selected by means of cross-validation on a hold-out set.

correct class, although not as well as the naive Bayes classifier, as shown in Fig. 7.12.

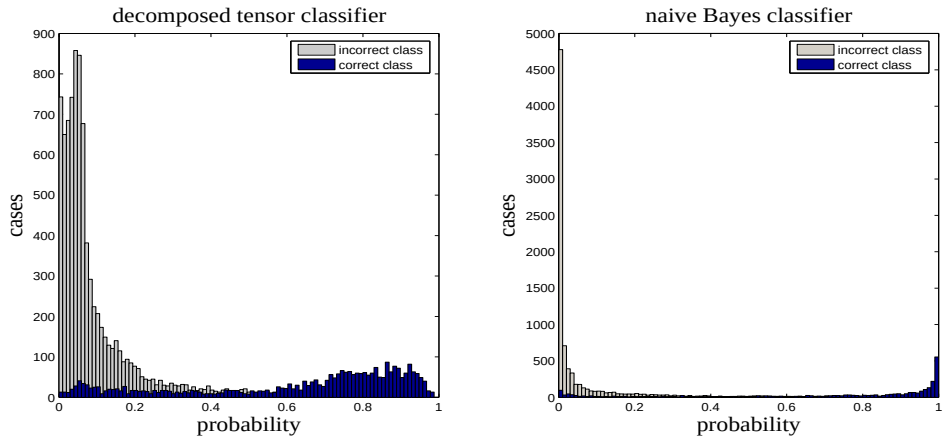


Figure 7.12: Distribution of posterior probabilities of correct and incorrect classes for the decomposed tensor classifier and the naive Bayes classifier.

The analysis of the classification accuracy and the difference in logarithmic score show that although both classifiers operate differently, they perform comparably with respect to classification accuracy. If we inspect the classifications that were made by the classifiers then it is interesting to see that only 254 out of a total of 2955 cases (8.60%) have been classified differently by the two classifiers. Out of these 254 cases, the naive Bayes classifier assigned 107 cases to the correct class, whereas the decomposed tensor classifier assigned 93 cases to the correct class. Hence, the classifiers are able to classify different cases correctly, suggesting that there are certain problems for which the naive Bayes classifier is more suitable, and other problems for which the decomposed tensor classifier is more suitable.

7.3 Predicting CHD with a noisy-threshold classifier

In this section, we employ a novel Bayesian classifier, introduced in (Jurgelenaite and Heskes, 2006), that facilitates medical interpretation as it explicitly provides for a semantics in terms of cause and effect relationships (Heckerman and Breese, 1994). This *noisy-threshold classifier* is based on a generalization of the well-known *noisy-or* model, which has already been used for the purpose of text classification in (Vomlel, 2002). In order to demonstrate the merits of the noisy-threshold classifier in a medical context, we apply the technique to the prediction of *carcinoid heart disease* (chd); a serious condition that arises as a complication of certain neuroendocrine tumors (Zuetenhorst et al., 2003). We demonstrate that the noisy-threshold classifier performs competitively with state-of-the art classification techniques for this medi-

cally relevant problem. Furthermore, an expert physician at the Netherlands Cancer Institute (NKI) was consulted, and it is demonstrated how her knowledge concerning chd relates to the parameters that were estimated for the noisy-threshold classifier.

7.3.1 Semantics of the noisy-threshold model

We will show how to arrive at the noisy-threshold model, by introducing a number of assumptions that are motivated by the semantics in terms of causes and effects, that is taken to hold for causal independence models. Causal independence is a popular way to specify interactions among cause variables (Pearl, 1988; Heckerman and Breese, 1994; Zhang and Poole, 1996; Díez, 1993; Lucas, 2005). The global structure of a causal independence model is shown in Figure 4.2 and expresses the idea that causes $\mathbf{C} = \{C_1, \ldots, C_n\}$ influence a common effect E through hidden variables $\mathbf{H} = \{H_1, \ldots, H_n\}$ and a deterministic function f, called the *interaction function*. The causal independence assumption does not refer to independence between causes, but rather to the assumption that hidden variables H_i are independent of causes $\mathbf{C} \setminus \{C_i\}$ given C_i. Causal independence is therefore also known as *independence of causal influence* or *exception independence*. In practice, causes in a causal independence model can be dependent; for instance, when the model is embedded within a larger network, or if there are direct dependencies between causes. However, if causes are completely observed then it is not necessary to model the dependence structure between cause variables.

We assume that causes are either *present* or *absent*. We use x^+ and x^- for $X = \top$ (true) and $X = \bot$ (false) respectively, and interpret $\top$ as 1 and $\bot$ as 0 in an arithmetic context. The individual contribution of a cause C_i to the effect E is realized by the parameter $P(H_i \mid C_i)$ associated with the hidden variable H_i; if $P(h_i^+ \mid c_i^+) < 1$ then H_i is said to inhibit the cause C_i. The assumption of *accountability* states that absent causes do not contribute to the effect which implies that $P(h_i^+ \mid c_i^-) = 0$ (Pearl, 1988). The interaction function f represents in which way the hidden variables H_i, and indirectly also the causes C_i, interact *deterministically* to yield the final effect E. Since variables are binary, f reduces to a Boolean function. It is also useful to introduce a *leak term* whenever it is infeasible to identify all the variables that influence the effect. We model this leak term by postulating a cause C_l that is always present and with which is associated a leak probability $P(h_l^+ \mid c_l^+)$ (Pradhan et al., 1994). In this manner, we maintain the *closed-world assumption* (Reiter, 1978). It follows from these assumptions that the conditional probability of the effect e^+ given a configuration $\mathbf{c}$ of the causes $\mathbf{C}$ can be obtained from the parameters $P(h_i \mid c_i)$ as follows (Zhang and Poole, 1996):

$$P_f(e^+ \mid \mathbf{c}) = \sum_{\mathbf{h}:\ f(\mathbf{h})} \prod_{i=1}^{n} P(h_i \mid c_i)\,, \tag{7.15}$$

where $P_f(e^+ \mid \mathbf{h}) = 1 \Leftrightarrow f(\mathbf{h}) = \top$.

As there are 2^{2^n} different n-ary Boolean functions (Enderton, 1972; Wegener, 1987), the potential number of causal independence models that is admitted by Eq. (7.15) is huge. However, if we assume that the order of the cause variables does not matter, the Boolean functions become *symmetric* and the number of such functions reduces to 2^{n+1} (Wegener, 1987). An important symmetric Boolean function is the *exact* Boolean function ϵ_m, which is defined as:

$$\epsilon_m(h_1, \ldots, h_n) = \top \Leftrightarrow \sum_{j=1}^{n} h_j = m\,.$$

Any symmetric Boolean function can be decomposed in terms of the exact functions ϵ_m as follows (Wegener, 1987):

$$f(h_1, \ldots, h_n) = \bigvee_{m=0}^{n} \epsilon_m(h_1, \ldots, h_n) \wedge \gamma_m \tag{7.16}$$

where γ_m are Boolean constants dependent on the choice of the symmetric function f. A particularly useful type of symmetric Boolean function is the *threshold* function τ_k, which simply checks whether there are at least k values $\top$ among the arguments, i.e.:

$$\tau_k(h_1, \ldots, h_n) = \top \Leftrightarrow \sum_{j=1}^{n} h_j \geq k\,.$$

In terms of causes and effects, the use of the threshold function as the interaction function of a causal independence model expresses the notion that a *sufficient* number of causes should be present in order to induce the effect. Then, the *noisy-threshold model*, as defined in (Jurgelenaite et al., 2006), is given by:

$$P_{\tau_k}(e^+ \mid \mathbf{c}) = \sum_{j=k}^{n} \sum_{\mathbf{h}\colon \epsilon_j(\mathbf{h})} \prod_{i=1}^{n} P(h_i \mid c_i)\,. \tag{7.17}$$

To express a threshold function in terms of Eq. (7.16) we use $\gamma_0 = \cdots = \gamma_{k-1} = \bot$ and $\gamma_k = \cdots = \gamma_n = \top$. Note that the noisy-or model, with $f(h_1, \ldots, h_n) \Leftrightarrow h_1 \vee \cdots \vee h_n$, corresponds to threshold function τ_1, and the noisy-and model, with $f(h_1, \ldots, h_n) \Leftrightarrow h_1 \wedge \cdots \wedge h_n$, corresponds to threshold function τ_n. Hence, these two commonly used causal independence models are the extremes of a spectrum of causal independence models that are defined by the noisy-threshold function.

The parameters $P(h_i^+ \mid c_i^+)$ of the model can be learned using an *expectation-maximization* (EM) algorithm (Dempster et al., 1977). EM is a method for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on (unobserved) hidden variables. Every iteration of an EM algorithm consists of two steps: the expectation step (E-step), which computes the expected value of the hidden variables, and a maximization step (M-step), which computes the

maximum likelihood estimates of the parameters given the data. To learn the parameters in the noisy-threshold classifier we use the EM algorithm for noisy-threshold models (Jurgelenaite and Heskes, 2006). Generally, the expectation and maximization steps are alternated repeatedly until convergence. However, for small data sets, this may result in overfitting artifacts; an issue to which we return later.

The analysis in this section has shown that causal independence models such as the noisy-threshold model have an interesting semantics in terms of causes and effect, and can be learned using the EM algorithm, given a symmetric Boolean interaction function. The next section describes the medical problem that is used to illustrate the usefulness of the noisy-threshold model as a classifier.

7.3.2 Carcinoid heart disease

Carcinoid tumors belong to the group of neuroendocrine tumors, which are known for the production of vasoactive agents in the presence of metastatic disease; usually hepatic (liver) metastases. Among these agents, serotonin is the most important agent, leading to the characteristic carcinoid syndrome of flushes and diarrhea. The other main characteristic feature of neuroendocrine tumors is the slow progression of most tumors if the histology shows a low-grade pattern (Zuetenhorst and Taal, 2005).

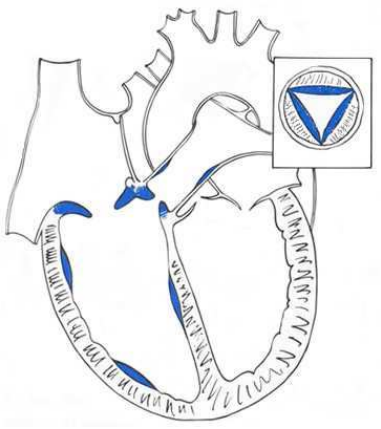


Figure 7.13: chd is characterized by heart valve fibrosis as shown in the overlay.

Serotonin overproduction may also cause carcinoid heart disease (chd), which is characterized by fibrosis of the right sided heart valves as shown in Fig. 7.13. Fibrosis induces thickening and retraction of the tricuspid valve, leading to tricuspid insufficience and ultimately heart failure, which is the cause of death in as much as half of carcinoid patients (Zuetenhorst et al., 2003; Zuetenhorst and Taal, 2003). Since so many carcinoid patients die of chd, it is important to distinguish patients that are admitted to the clinic into patients that are prone to develop a severe form of carcinoid heart disease, and those that do not develop this severe form. In this way,

patients that are at risk can be given more aggressive treatment in order to reduce the probability of the development of chd. Hence, the classification task for this medical problem will be to classify the patients into these two groups, depending on the attributes that are known at the time of admission to the clinic. We use chd^+ to denote the development of moderate to extreme tricuspid valve insufficience and chd^- to denote the absence, or development of mild tricuspid valve insufficience during patient follow-up.

Table 7.2: Patient attributes that are measured at admission.

Name	Definition	Name	Definition
hia	5-HIAA levels	gil	General illness
cga	Chromogranin A levels	bob	Bowel obstruction
dia	Diarrhea	ibl	Internal bleeding
whe	Wheezing	fev	Fever
flu	Flushing	hme	Hepatic metastases
apa	Abdominal pain		

In principle, the physician can make use of the attributes that are measured at admission (Table 7.2), in order to predict the development of chd. However, in practice, in order to determine the probability of developing moderate to severe tricuspid valve insufficience, the physician makes use of the following decision rule:

$$P(\text{chd}^+ \mid \mathbf{c}) = \begin{cases} 0.50 & \text{if } \text{hia}^+ \wedge \text{dia}^+ \wedge \text{hme}^+ \\ 0.25 & \text{if } \text{hia}^+ \wedge (\text{dia}^- \wedge \text{hme}^+ \vee \text{dia}^+ \wedge \text{hme}^-) \\ 0.10 & \text{if } \text{hia}^+ \wedge \text{dia}^- \wedge \text{hme}^- \vee \text{hia}^- \wedge \text{dia}^+ \wedge \text{hme}^+ \\ 0.03 & \text{otherwise.} \end{cases}$$

The aim of this section, is to show that a noisy-threshold model can be used as a Bayesian classifier where performance is compared with the physician's classification performance, as well as with standard classification techniques such as the naive Bayes classifier, logistic regression, and decision-trees. Patient attributes are used as cause variables in the definition of a noisy-threshold model, and it is assumed that independence of causal influence, accountability, symmetry and sufficiency hold. As required, variables are binary, and positive states of variables are perceived to be less favorable than negative states, such that they could be responsible for carcinoid heart disease. To train and test Bayesian classifiers for this medical problem, we have used a clinical database consisting of fifty-four patients that suffered from a neuroendocrine tumor, and for which the grade of tricuspid valve insufficience was known. Twenty-two patients developed moderate or worse tricuspid valve insufficience during follow-up.

We have not yet touched upon the most important assumption of causal independence models. That is, can the variables be regarded as causes of carcinoid heart disease? For some attributes this is questionable. Diarrhea for instance is a symptom of other processes and is therefore not likely to be a cause of carcinoid heart disease.

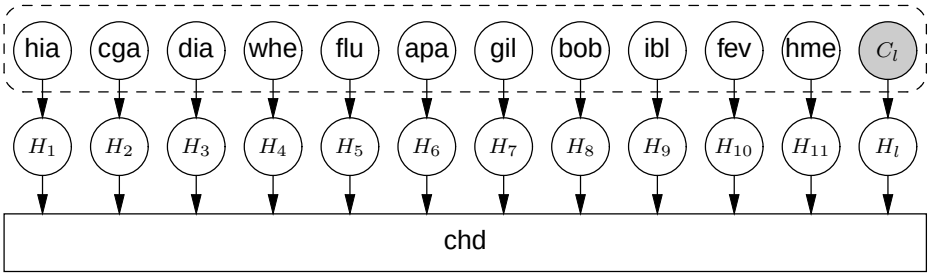


Figure 7.14: A noisy-threshold model for carcinoid heart disease, where the dashed region represents the total tumor burden for the patient. Note the use of the leak cause C_l in order to model possible hidden causes.

However, we *can* interpret the attributes as risk factors that act as components of the total *tumor burden*, as depicted in Fig. 7.14. Since the causes are assumed to be completely observed, we refrain from adding additional dependencies between cause variables.

7.3.3 The noisy-threshold classifier

Classifier construction

Construction of a noisy-threshold classifier (NTC) proceeds as follows. We first determine the cause variables $\mathbf{C}$ and effect variable E that are used in the classifier. In the context of a classifier, the cause variables stand for the attributes and the effect variable stands for the class-variable. Secondly, we need to determine the positive states of the variables. In the chd domain, the positive states are simply defined as the presence of attributes that affect the presence of the class-variable chd. Once the cause and effect variables have been defined, we need to find both the optimal values for the parameters $P(h_i^+ \mid c_i^+)$ using an EM algorithm, as well as the correct threshold function τ_k.

The parameters and threshold function are learned from a database $\mathcal{D} = \{\mathbf{u}^1, \ldots, \mathbf{u}^N\}$ where instances $\mathbf{u}^j = \{\mathbf{c}^j, e^j\} = \{c_1^j, \ldots, c_n^j, e^j\}$ with $j = 1, \ldots, N$ consist of realizations of causes and the effect. Let $\mathcal{D}^+ \subseteq \mathcal{D}$ denote those instances $\{\mathbf{c}^j, e^j\}$ for which $e^j = \top$, and let $\mathcal{D}^- \subseteq \mathcal{D}$ denote those instances $\{\mathbf{c}^j, e^j\}$ for which $e^j = \bot$. We define the following measures with respect to a fixed database $\mathcal{D}$ and model M. Let the *true positives* (tp) stand for the number of instances $\mathbf{u}^j \in \mathcal{D}^+$ for which $P(e^+ \mid \mathbf{c}^j) \geq 0.5$ and let the *false negatives* (fn) stand for the number of instances $\mathbf{u}^j \in \mathcal{D}^+$ for which $P(e^+ \mid \mathbf{c}^j) < 0.5$. Likewise, we define the *true negatives* (tn) as the number of instances $\mathbf{u}^j \in \mathcal{D}^-$ for which $P(e^+ \mid \mathbf{c}^j) < 0.5$ and the *false positives* (fp) as the number of instances $\mathbf{u}^j \in \mathcal{D}^-$ for which $P(e^+ \mid \mathbf{c}^j) \geq 0.5$. In order to learn the parameters of the noisy-threshold model, we used a training set

$\mathcal{D}_{\text{train}}$ and a validation set $\mathcal{D}_{\text{validate}}$. The validation set is used to counterbalance the overfitting that may occur when learning model parameters. The aim of the learning phase is to maximize both the *classification accuracy*

$$\eta(\mathcal{D}) = \frac{tp + tn}{tp + tn + fn + fp} \times 100\%$$

as a measure of the number of correctly classified cases, and the F_1 *measure*

$$F_1(\mathcal{D}) = \frac{2\pi\rho}{\pi + \rho}$$

as a measure that takes into account the tradeoff between *precision* $\pi = \frac{tp}{tp+fp}$ and *recall* $\rho = \frac{tp}{tp+fn}$, which is also known as *sensitivity*. We use these two measures since the accuracy is the obvious measure but may convey the wrong intuition when the classes are not equal in size (van Rijsbergen, 1979). Finding the optimal noisy-threshold classifier then proceeds as follows:

1. Divide the data set $\mathcal{D}$ into the disjoint sets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{validate}}$ and $\mathcal{D}_{\text{test}}$.
2. For all noisy-threshold models $P_{\tau_1}, \ldots, P_{\tau_n}$ with $n = |\mathbf{C}|$, use the training data $\mathcal{D}_{\text{train}}$ and the EM-algorithm in (Jurgelenaite and Heskes, 2006) to learn the parameters $P(h_i^+ \mid c_i^+)$.
3. Select the noisy-threshold model and the number of iterations of the EM-algorithm that maximizes $w_1 \cdot \eta(\mathcal{D}_{\text{validate}}) + w_2 \cdot F_1(\mathcal{D}_{\text{validate}})$ with equal weights $w_1 = w_2$, as the optimal noisy-threshold classifier.

With regard to the clinical data set $\mathcal{D}$ we have used a leave-one-out cross-validation scheme to implement the above algorithm. $\mathcal{D}$ contains too many missing values to simply remove the instances that contain missing data. We have used *mean substitution* (Kline, 1998) as an imputation scheme, and note that *multiple imputation* (Rubin, 1987) produced similar results.

Classifier evaluation

In order to evaluate the performance of the noisy-threshold classifier, we compared its classification accuracy with the accuracy of a number of other well-known algorithms. For the comparison we have used the naive Bayes classifier (NBC), for which

$$P(e^+ \mid \mathbf{c}) \propto P(e^+) \prod_{i=1}^{n} P(c_i \mid e^+)\,,$$

logistic regression (LG), where the posterior probability of developing carcinoid heart disease is given by

$$P(e^+ \mid \mathbf{c}) = \frac{1}{1 + \mathrm{e}^{-(a_0 + a_1 c_1 + \cdots + a_n c_n)}}\,,$$

and a decision-tree learning algorithm (C4.5), as implemented by the WEKA machine learning tool (Witten and Frank, 2005).[3] Furthermore, we compare the performance of the optimal noisy-threshold classifier with that of the noisy-or classifier P_{τ_1} as a special case (Vomlel, 2002). Parameters are estimated from data and for the probabilistic algorithms classification proceeds by selecting the class value that has highest posterior probability $P(e^+ \mid \mathbf{c})$. For the decision-tree learning algorithm, no posterior probability is computed and classification proceeds by traversing the tree and selecting the class value that is associated with the leaf node.

As pointed out in (Salzberg, 1997), when comparing two classification algorithms, the approach preferred to a standard t-test, is to use a binomial test, which uses the number of cases n for which the two classifiers produce a different output, and the number of cases s where the output of the examined classifier was correct, while the output of the reference classifier was wrong. Under the null hypothesis that the two classifiers perform equally well, we compute:

$$q = \sum_{i=s}^{n} \frac{n!}{i!(n-i)!}(0.5)^n$$

for a one-tailed test, and $p = 2q$ for a two-tailed test.

Since the classification accuracy assumes equal costs between false positives and false negatives, we use the *Receiver Operating Characteristics* (ROC) curve to compare the performance of some of the classifiers in terms of the trade off between *sensitivity* $\rho = \frac{tp}{tp+fn}$ and *specificity* $\sigma = \frac{tn}{tn+fp}$ for every possible cutoff (Egan, 1975), where ρ is shown on the y-axis, and $1 - \sigma$ is shown on the x-axis. This performance can be quantified by computing the area under the ROC curve (AUC), which has been shown to equal the outcome of the Mann-Whitney U statistic (Bamber, 1975):

$$\text{AUC} = \frac{\sum_{\mathbf{c}^i \in \mathcal{D}^+} \sum_{\mathbf{c}^j \in \mathcal{D}^-} u(\mathbf{c}^i, \mathbf{c}^j)}{|\mathcal{D}^+||\mathcal{D}^-|}$$

where

$$u(\mathbf{c}^i, \mathbf{c}^j) = \begin{cases} 1 & \text{if } P(e^+ \mid \mathbf{c}^i) > P(e^+ \mid \mathbf{c}^j) \\ \frac{1}{2} & \text{if } P(e^+ \mid \mathbf{c}^i) = P(e^+ \mid \mathbf{c}^j) \\ 0 & \text{if } P(e^+ \mid \mathbf{c}^i) < P(e^+ \mid \mathbf{c}^j) \end{cases}$$

We can interpret this statistic as follows. We assume that there is a ranking between instances in $\mathcal{D}$ such that any deviation from the perfect ranking that ranks all positive examples higher than all negative examples leads to a decrease in the AUC (Cortes and Mohri, 2004). If $P(e^+ \mid \mathbf{c}^i) > P(e^+ \mid \mathbf{c}^j)$ then we produce a correct ranking, if $P(e^+ \mid \mathbf{c}^i) = P(e^+ \mid \mathbf{c}^j)$ then we break ties at random and produce a correct ranking half of the time, and if $P(e^+ \mid \mathbf{c}^i) < P(e^+ \mid \mathbf{c}^j)$ then we produce an incorrect ranking.

[3] We use WEKA's default parameter settings; the default imputation method is to interpret a missing value for X as a separate value $x \in \Omega_X$.

7.3.4 Results

Classification performance

Table 7.3 lists the classification accuracy for noisy-threshold classifiers P_{τ_1} to $P_{\tau_{12}}$. The noisy-threshold classifier P_{τ_6} is selected, based on the validation set $\mathcal{D}_{\text{validate}}$, and shows the best classification accuracy of 72% on the test set $\mathcal{D}_{\text{test}}$. Note that this exceeds considerably the classification accuracy of 54% for the noisy-or classifier P_{τ_1}.

Table 7.3: Classification accuracy on $\mathcal{D}_{\text{test}}$ for noisy-threshold classifiers $P_{\tau_1}, \ldots, P_{\tau_{12}}$.

NTC	$\eta(\mathcal{D}_{\text{test}})$	**NTC**	$\eta(\mathcal{D}_{\text{test}})$	**NTC**	$\eta(\mathcal{D}_{\text{test}})$
P_{τ_1}	54 %	P_{τ_5}	69 %	P_{τ_9}	59 %
P_{τ_2}	65 %	P_{τ_6}	**72** %	$P_{\tau_{10}}$	59 %
P_{τ_3}	65 %	P_{τ_7}	65 %	$P_{\tau_{11}}$	59 %
P_{τ_4}	70 %	P_{τ_8}	57 %	$P_{\tau_{12}}$	59 %

In order to test how well the NTC performs compared with the physician, and with the other classification algorithms that were previously discussed, we have determined the classification accuracy. Table 7.4 describes the classification accuracy on $\mathcal{D}_{\text{test}}$ for the physician's decision rule, NBC, LG, C4.5 and noisy-or, and p-values for the null-hypothesis that the classifier accuracy is comparable to that of the NTC P_{τ_6}.

Table 7.4: Classification accuracy and p-values for classification of $\mathcal{D}_{\text{test}}$.

Classifier	$\eta(\mathcal{D}_{\text{test}})$	p
physician	69 %	$7.0 \cdot 10^{-1}$
LG	67 %	$6.3 \cdot 10^{-1}$
NBC	63 %	$2.3 \cdot 10^{-1}$
noisy-or	54 %	$6.4 \cdot 10^{-3}$
C4.5	44 %	$6.2 \cdot 10^{-5}$

Note that the expert physician's classification accuracy is reasonably high, outperforming all but the noisy-threshold classifier. The noisy-threshold classifier P_{τ_6} shows the best classification accuracy, although the difference is significant only for C4.5 and the noisy-or classifier at a confidence level of $p = 0.05$. For the physician's decision rule, the naive Bayes classifier, and logistic regression, we cannot reject the null hypothesis that the algorithms may in fact be equally accurate for this data set.

It is well-known that classifiers that show large bias tend to outperform classifiers that show high variance for small data sets, since this reduces the risk of overfitting. For this reason, the naive Bayes classifier tends to perform well on many data sets (Kohavi and Wolpert, 1996). However, although not always reflected in its classification accuracy (Domingos and Pazzani, 1997), the assumption of independence

between attributes given the class-variable, is a strong assumption which does not hold in general. In contrast, the noisy-threshold classifier's assumptions are motivated by a cause-effect semantics as described in Section 7.3.1, and hold for domains where the presence of a sufficient number of causes is sufficient to induce the effect.

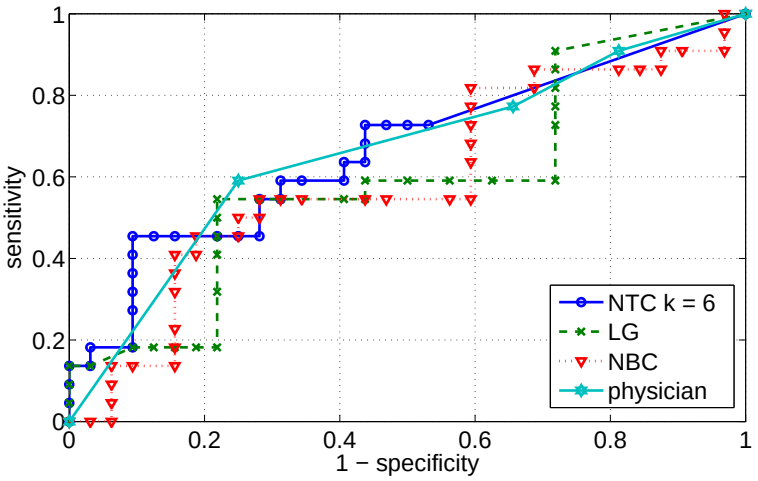


Figure 7.15: ROC curve for the noisy-threshold classifier, logistic regression, and the naive Bayes classifier, where the straight line segment in the NTC curve is a consequence of the assumption that absent causes do not contribute to the effect.

Figure 7.15 presents the ROC curves for the physician's decision rule, the noisy-threshold classifier P_{τ_6}, the naive Bayes classifier, and logistic regression, where the area under the curve equals 0.66, 0.66, 0.60 and 0.59 respectively. Although the performance in terms of AUC is mediocre, both the physician's decision rule, and the noisy-threshold classifier show a considerably better performance than the other standard classification techniques. The ROC curve does demonstrate a potential danger of using the noisy-threshold classifier, especially when the causal assumptions are not satisfied. Whereas the naive Bayes classifier is able to gradually increase the true positive rate at the expense of increasing the true negative rate, the noisy-threshold classifier fails to accomplish this for all true positive rates. This is a consequence of the assumption that absent causes cannot contribute to the effect; the probability $P_{\tau_k}(e^+ \mid \mathbf{c}^i)$ of assigning an instance to the positive class equals zero whenever the number of present causes is less than the threshold k.

Medical interpretation

In this section we look at the noisy-threshold classifier for chd from a medical point of view. Prior to presenting the resulting classifier, we have asked the physician to indicate how important the individual attributes were felt to be with respect to predicting the development of carcinoid heart disease.

According to the physician, progressive carcinoid disease is often accompanied by the carcinoid syndrome, which is characterized by diarrhea (dia) caused by increased bowel motility due to serotonin overproduction, by periodical flushing attacks (flu) due to the synergistic interaction between various vasoactive agents, and sometimes by wheezing (whe). As discussed in Section 7.3.2, serotonin overproduction is thought to play a key role in the etiology of chd and it can be measured indirectly by means of the urinary 5-HIAA level (hia) since this is a metabolite of serotonin. Hence, the variables related to the carcinoid syndrome are indicative of serotonin overproduction and ultimately chd. It is therefore assumed that the variables hia, dia, flu and to a lesser extent whe have a high predictive value. Serotonin overproduction is itself caused by the carcinoid tumor in the presence of particular metastases; hormones released by carcinoid tumors are often destroyed by the liver before they reach the general circulation to cause symptoms. Therefore, only hepatic metastases (hme), or metastases that can release hormones directly into the general circulation, can produce the carcinoid syndrome. According to the physician, the presence of hepatic metastases (hme) during hospitalization is indicative of chd development, since this is a requirement for serotonin overproduction, The plasma chromogranin A (cga) level is used as a general marker of neuroendocrine activity and tumor extensiveness (Nobels et al., 1998). Although not regarded as important as the previously discussed attributes, the physician expected cga to have a high predictive value since extensive tumors with high neuroendocrine activity are more likely to cause chd. In contrast, the variables ibl, fev, apa and bob were not thought to predict chd very well. Local progression of hyper-vascular primary tumors into the lumen of the small bowel is often the cause of internal bleeding (ibl), but is not thought to be related to metastatic disease. Fever (fev) can be caused by hepatic metastases, as captured by the variable hme, but it is also a non-specific symptom that is not necessarily caused by carcinoid disease in the first place. Abdominal pain (apa) and bowel obstruction (bob) are often caused by complications due to the primary tumor and were assumed to be unrelated to the development of chd. According to the physician, general illness (gil) could be indicative of the development of carcinoid heart disease; a poor condition is often due to extensive metastases and therefore a high probability of serotonin overproduction. In general, the physician expected that at least some of the risk factors should occur together in order to cause chd.

Figure 7.16 depicts the estimates of prior probabilities $P(c_i^+)$ and conditional probabilities $P(h_i^+ \mid c_i^+)$ for the noisy-threshold classifier that was used for predicting chd. The predictive value of the variables hia, dia, flu and whe is reflected in

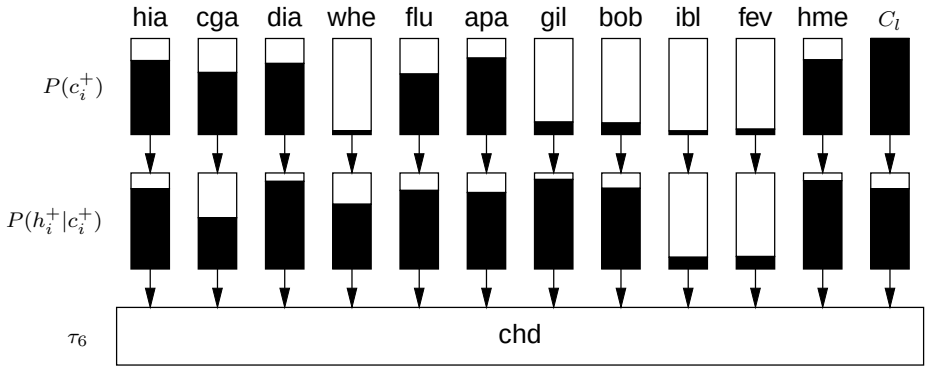


Figure 7.16: Estimates of priors $P(c_i^+)$, and conditional probabilities $P(h_i^+ \mid c_i^+)$, for the noisy-threshold classifier with threshold function τ_6.

the reasonably high associated probabilities $P(h_i^+ \mid c_i^+)$ with $i \in \{1, 3, 4, 5\}$, which range from 0.67 to 0.91, where wheezing is indeed seen to be of less predictive value than the other attributes. The presence of hepatic metastases (hme) is also an important predictor of chd, as is indicated by the high probability $P(h_{11}^+ \mid c_{11}^+) = 0.92$. Notice that most patients that are admitted already present with such metastases, which is reflected by the high prior probability $P(c_{11}^+) = 0.78$. Contrary to the physician's expectations, cga was not a very good predictor of chd, with $P(h_i^+ \mid c_i^+) = 0.53$. In hindsight, this may be explained by the fact that cga overproduction does not necessarily reflect serotonin overproduction, and if it does, it may be redundant information given that we have observed hia, which is a metabolite of serotonin. Internal bleeding (ibl) and fever (fev), with $P(h_i^+ \mid c_i^+) = 0.12$ and $P(h_i^+ \mid c_i^+) = 0.13$ respectively, did not contribute much to the effect. Unexpectedly, both abdominal pain (apa) and bowel obstruction (bob) had relatively high probability values $P(h_i^+ \mid c_i^+)$ of 0.80 and 0.84 respectively. After some deliberation, the physician gave the following possible explanation. Since abdominal pain and bowel obstruction are often caused by complications due to the primary tumor, both apa and bob indicate a midgut tumor with possible mesenterial fibrosis. A midgut localization is a prerequisite for serotonin overproduction, and mesenterial fibrosis is thought to be related to tricuspid valve fibrosis (Modlin et al., 2004). Therefore, the presence of these variables could have been indicative of the development of chd. General illness (gil) had a high probability value of $P(h_i^+ \mid c_i^+) = 0.93$. Five out of seven patients that suffered from general illness indeed developed chd. The threshold function τ_6 corresponds to the physician's opinion that the presence of just one risk factor is generally insufficient to cause chd, whereas the presence of all risk factors is much too strict a requirement as a cause for chd; demonstrating that the noisy-threshold model as a generalization of both the noisy-or and noisy-and model can be the proper choice for realistic domains.

7.4 Summary

In this chapter, we have described three different probabilistic classification techniques. We discuss each of the techniques seperately.

Maximizing mutual information

The MMI algorithm makes few structural assumptions and iteratively builds classifier structures that reflect higher-order dependencies between evidence variables. In this sense, the MMI algorithm resembles Sahami's limited-dependence classifier (Sahami, 1996) with the difference that we do not require the addition of an arc between the class variable and each evidence variable. Furthermore, the heuristic that was used during the estimation of conditional mutual information prevents the construction of overly complex network structures and the introduction of spurious dependencies. As is shown, the number of higher-order dependencies will only increase if this is warranted by sufficient data. The experimental results show that classification performance of the MMI classifier is comparable with that of the FAN classifier while the weaker assumptions allow for a network structure that is less ad-hoc and somewhat better to interpret from a medical point of view.

Decomposed tensor classifiers

In this chapter, we have also shown that tensor decompositions can be used for the purpose of probabilistic classification. The classification accuracy of this novel classification method on a problem in medical diagnosis is comparable to that of the naive Bayes classifier and other methods which have been specifically developed to solve this classification problem. The logarithmic score of decomposed tensor classifiers suggests that the method is less suitable for obtaining accurate posterior probabilities, although the different mode of operation, together with the results concerning correctly classified cases, suggest that there may be particular problems for which this new technique performs better than the naive Bayes classifier. Current limitations of the technique are the requirements that data is discrete and complete, and the fact that learning the classifiers requires more computational resources than the (easy to learn) naive Bayes classifier.

Decomposed tensor classifiers are a new way of employing tensor decompositions, the usefulness of which we have demonstrated in this research using a classification problem in medical diagnosis. Dealing with current limitations and validation of the technique by means of multiple datasets are future research goals.

The noisy-threshold classifier

The noisy-threshold classifier is a novel type of classifier that has a well-defined semantics in terms of causes and effect. Due to the independence assumptions that are

made by the classifier, parameters can be reliably estimated without needing to resort to huge amounts of data. This is an important feature since many domains are characterized by limited amounts of data, as discussed in (van Gerven and Lucas, 2004b). Learning Bayesian classifiers from data is to be contrasted with the construction of a full Bayesian network that captures available domain knowledge, which, although possible, can be very resource intensive for realistic domains.

We have demonstrated that the noisy-threshold classifier performs comparably with the decision rule that is used by an expert physician, and competitively with state-of-the-art classifiers, on an important classification task in oncology. Furthermore, it significantly outperforms the noisy-or classifier, as a special case of the noisy-threshold classifier, for this data set. The semantics of the noisy-threshold classifier enables an interpretation in terms of available domain knowledge, as is illustrated by the physician's interpretation of classifier parameters. Nevertheless, one should be cautious when defining the positive states of the cause variables since negative states cannot contribute to the effect, as reflected by the straight line segment of the ROC curve. The competitive classification performance and well-defined semantics make the noisy-threshold classifier a promising new machine learning technique, as was demonstrated here in the context of medical prognosis.

Chapter 8

Conclusion

The goal of this thesis has been to examine how graphical models for clinical decision support (such as Bayesian networks and influence diagrams) can be constructed in order to deal with large and complex dynamic decision problems that require reasoning under uncertainty and are characterized by limited availability of data. In this concluding chapter, we describe the scientific contributions of this thesis (Section 8.1), consider the strengths and limitations of our approach (Section 8.2), and reflect on the subject matter of this thesis (Section 8.3).

8.1 Scientific contributions

Chapter 3: Clinical decision support with Bayesian networks

The construction of Bayesian networks for clinical decision support often proceeds in an ad-hoc fashion. Therefore, in Chapter 3, we addressed the problem of how to construct Bayesian networks for clinical decision support by considering how a clinical problem together with practical considerations translate into particular Bayesian network designs. We have shown that insight into the nature of the clinical problem can be obtained by:

- describing the clinical task in terms of abstract problem solving,
- distinguishing non-temporal and temporal forms of problem solving,
- differentiating between a patient model and a physician model, and
- dividing clinical concepts into different categories.

Practical considerations relate to the amount of time one is willing to spend on model construction. The following factors reduce modeling effort at the expense of model expressiveness:

- using restricted associative instead of unrestricted causal models,
- using non-temporal instead of temporal models,

- taking a restricted number of clinical categories into account,
- enforcing conditional independence assumptions between clinical categories, and
- externally imposing, instead of incorporating, a decision-making strategy.

By making the nature of the clinical task more explicit and by taking into account possible practical considerations, we have pointed out a more principled approach when choosing a Bayesian network design in order to solve a clinical problem. Finally, we have shared some of the insight that has been gained during the development of Bayesian networks for clinical decision support, which we have divided into variable definition, structure specification, factor association, and parameter estimation.

Chapter 4: A qualitative characterization of causal independence

The manual construction of Bayesian networks, especially probability estimation, is a difficult task. This motivates the development of techniques that reduce the effort when specifying a Bayesian network. In Chapter 4, we have focused on causal independence models, which offer one way to facilitate Bayesian network construction. The theory developed in this chapter allows one to identify whether a particular causal independence model with a chosen interaction function can fulfill the specified qualitative properties in principle. This is a useful development since without the theory one would need to estimate the conditional probabilities $P(\hat{m} \mid \hat{c})$ for each of the causes and exhaustively compute the influences and synergies for the model as in Section 4.1.2. By means of the theory, the qualitative behavior can be read off directly from the underlying interaction function. The developed theory can also be employed for placing direct constraints on the structure of the underlying interaction function *given* a qualitative specification in terms of influences and synergies, as demonstrated by Tables 4.1, 4.2, and 4.3. These results can be used to generate the set of interaction functions that respect the constraints which facilitates the selection of a suitable interaction function for problems that can be represented as causal independence models. Given the fact that probability estimation is time-consuming, and since causal independence models allow for efficient inference and have a semantics that is understandable by the physician, the presented approach (i.e., using a qualitative specification to identify a suitable causal independence model) is seen as a valuable contribution.

Chapter 5: Dynamic decision making with DLIMIDs

Chapter 3 described the steps that need to be taken when constructing a Bayesian network for clinical decision support, where it was assumed that the physician's treatment strategy is either externally imposed or explicitly represented. However, when

dealing with (clinical) decision problems, the ultimate goal of decision theory is to find the optimal (treatment) strategy in the first place. This was the topic of Chapter 5, where dynamic limited-memory influence diagrams (DLIMIDs) were described as a novel approach to the representation of dynamic decision problems. DLIMIDs provide an alternative to the solution strategies offered by partially-observable Markov decision processes (Monahan, 1982) for the solution of (infinite-horizon) dynamic decision problems. We have developed new solution algorithms, where simulated annealing combined with single rule updating is shown to perform well on a realistic clinical decision problem. The alternative representation of complex dynamic decision problems together with the definition of algorithms that find acceptable solutions motivates the usefulness of our approach.

Chapter 6: A probabilistic model for carcinoid prognosis

In Chapter 6, we embarked on the manual construction of a dynamic Bayesian network (DBN) for prognosis of low-grade carcinoid tumors of the midgut. With 218 variables and 74 342 CPT entries for the prior and transition model, the so-called carcinoid model is one of the largest DBNs for clinical decision support that has been developed to date. The resulting model was created from domain knowledge that was provided by an expert physician at the Netherlands Cancer Institute. It captures state-of-the-art knowledge about treatment and prognosis of carcinoid tumors. The predictive performance of the carcinoid model was not as good as that of a proportional hazards model, but it has to be noted that the latter model was allowed to learn from the data on which it was tested. Furthermore, the quality of the database itself can be questioned, as evident from Table 6.6 and Section 6.4.2. In Section 6.3.3, it was shown that the carcinoid model can make very specific predictions for individual patients, which is the carcinoid model's projected mode of operation. Even though the carcinoid model is an initial prototype, it has already demonstrated that DBNs are suitable for the representation of complex pathophysiological processes as they are influenced by the physician, whereas approximate inference allows for the online computation of prognostic outcome at future points in time.

Chapter 7: Bayesian classifiers for clinical decision support

A different approach was taken in Chapter 7, where Bayesian networks were learnt from data instead of expert knowledge. We focused on clinical decision making as a classification problem and used Bayesian networks with a restricted graph structure for the purpose of probabilistic classification. In the chapter, three varieties of these so-called Bayesian classifiers have been described. In Section 7.1 the maximum mutual information (MMI) algorithm was developed. In contrast to the limited-dependence classifier (Sahami, 1996), the MMI classifier is a selective method that uses a heuristic in order to automatically determine the number of incoming arcs

for the evidence variables. The algorithm performs well on a diagnostic problem in hepatology and allows graph structures that are more informative than that of, say, the naive Bayes classifier. In Section 7.2, tensor decompositions (a technique taken from multilinear algebra) were used for the purpose of probabilistic classification. In particular cases, these decompositions can be described in terms of a graphical model structure. They are shown to perform about as well as the naive Bayes classifier on the diagnostic problem of Section 7.1. Its good classification performance, along with the fact that it classifies other instances correctly when compared with the naive Bayes classifier, warrants further research on this new and promising technique for probabilistic classification. In Section 7.3, we analyzed the performance of a recently described type of Bayesian classifier. The noisy-threshold classifier is a causal independence model that is employed for the purpose of classification. It compares favorably with state-of-the-art classifiers on the prediction of carcinoid heart disease in carcinoid patients and due to the nice semantics in terms of causes and effects is also more interpretable by the physician. The described techniques offer new directions for learning Bayesian networks from a limited amount of data. We have demonstrated their usefulness using clinical datasets, but remark that their applicability extends beyond the medical domain.

8.2 Strengths and limitations

It is well-known that the manual construction of realistic Bayesian networks is difficult and time-consuming. Contrary to learning a Bayesian network from data (as in Chapter 7), where general purpose algorithms can be used to automatically construct a model, there are no off-the-shelf recipes for the manual construction of a Bayesian network. Chapter 3 provides a partial solution to this problem by coupling problem solving and a characterization of clinical tasks with particular Bayesian network designs. However, actual model construction must still be done by the knowledge engineer on a case-by-case basis. As yet, the large scale deployment of Bayesian networks (and expert systems in general) is not realized in practice, and we view the knowledge elicitation bottleneck as the main reason for this failure to deliver. It would therefore be a major improvement if the knowledge engineer has access to often used network designs. These designs may be specified at the task level, as was done in Section 3.2 and in (Murphy, 2002) for dynamic Bayesian networks, or at the level of node-node interactions, as was done in Section 3.3.3 and in (Neil and Fenton, 2000). For clinical purposes, we envision reusable network fragments for the functioning of organs, main pathophysiological processes, and often occurring complications, that can be reused when modeling different disorders.

The analysis of causal independence models in Chapter 4 allows for the identification of qualitative properties of a causal independence model based on its interaction function. A limitation of the current approach is that the theory is defined

for binary variables only, and a generalization to non-binary variables would extend the applicability of the theory. There is also a need to further research the identification of the set of interaction functions that fulfills a given qualitative characterization. For n causes, there are $5^n \cdot 5^{\binom{n}{2}} \cdot 5^{\binom{n}{2}}$ different qualitative characterizations (in terms of possible combinations of qualitative influences, additive synergies, and product synergies) and the size of the the set of Boolean functions that is associated with each qualitative characterization may become huge since we have 2^{2^n} possible Boolean functions. However, preliminary results indicate that sets with many ambiguous qualitative influences and synergies contain many functions, whereas sets with few ambiguous qualitative influences and synergies contain few functions. Since we expect real-world specifications to contain many unambiguous qualitative influences and synergies, the approach may still be feasible. As a final note, since causal independence models allow for efficient inference, it may be useful to approximate arbitrary probability distributions with causal independence models. By computing the qualitative properties of the target distribution, the described approach may aid in identifying the causal independence models that offer the best approximation.

The DLIMIDs of Chapter 5 allow us to find *acceptable* treatment strategies using algorithms such as single rule updating and simulated annealing but it is not guaranteed that the strategy found is the *optimal* strategy. However, alternatives such as partially-observable Markov decision processes (POMDPs) can only find solutions for small problems, which makes such an approach infeasible for complex medical decision problems. Therefore, any strategy that is found by a DLIMID and improves upon the accepted strategy that is used in current clinical practice (in the sense that expected utility increases) is regarded to be acceptable.

DLIMIDs require the a priori specification of the informational predecessors (observable variables) that are assumed to influence the treatment strategy. For example, the variable cga in Fig. 5.9 is excluded from consideration by the physician, even though its inclusion may well lead to better treatment strategies. Therefore, an interesting research direction would be to devise a procedure that adds observable variables automatically, based on the utility that is gained by its inclusion. As a specific example, consider a memory variable M that captures the history of a finding F, based on which we decide to treat or not to treat a patient. One way to search for better strategies is to automatically increase the length of the history that is represented by M. There are situations in which a full history is needed to make the optimal decision, which precludes this approach, but for real-world problems, changes in expected utility should decrease for older observations. Therefore, by focusing on more recent observations, the size of memory variables can be kept relatively small. For the same reason, it may be useful to adapt Algorithm 5.3 such that decision rules that change recent observations are selected more often than decision rules that change older observations, in order to speed up the approximation of the optimal treatment strategy.

In Chapter 6 we have constructed the carcinoid model for prognosis of low-grade midgut carcinoid tumors. Although the model did not perform as well as a proportional hazards model when predicting survival for patients taken from a clinical database, the model was better at making patient specific predictions due to the explicit representation of how domain variables interact. Furthermore, due to this explicit representation, the range of questions that can be answered by the carcinoid model exceeds that of the proportional hazards model since the latter is optimized for prediction only. The discussion of the carcinoid model in Section 6.4 has made clear that the construction of a dynamic Bayesian network for clinical decision support is hard. We expect that the quality of (dynamic) Bayesian networks is improved by taking the following considerations into account:

- A clear understanding of the clinical task and the a priori selection of a suitable Bayesian network design, based on the nature of the clinical task.
- A focus on clearly defined disorders that show limited variability in progression, where the treatment protocol is fixed and not subject to much change, thereby facilitating model construction and parameter estimation by the physician.
- The a priori availability of a high quality database that guides the identification of relevant domain variables, allows for automated learning of model parameters, and/or makes preliminary evaluation of model components possible.
- To retain the continuous nature of random variables as much as possible, or to use holding times in order to model that random variables remain in specific discrete states for a prolonged time. In the latter case, the resulting dynamic Bayesian network can be interpreted as a *semi-Markov* decision process (e.g., (Leong, 1994; Murphy, 2002)).
- The explicit representation of a physician's uncertainty about probability estimates in terms of hyper-parameters.

In Chapter 7 we focused on learning from limited amounts of data. The developed algorithms are not to be used for the accurate representation of a joint probability distribution, but rather for the purpose of probabilistic classification. The maximum mutual information (MMI) algorithm is useful when one desires to retain part of the independence structure that is present in the domain. Contrary to Sahami's limited dependence classifier (Sahami, 1996), the MMI algorithm does not require an upper bound for the number of incoming arcs to each evidence variable, since this is determined through Eq. (7.4), although it does require that we make a suitable choice for the parameter β, which is not straightforward.

In Section 7.2, we focused on the use of rank-K approximations as the basis for decomposed tensor classification which, at the moment, is restricted to discrete

and complete data. The rank-K approximation is a special case of the more general Tucker decomposition, which may also be used for the purpose of probabilistic classification and can be learnt using *higher-order orthogonal iteration* (de Lathauwer et al., 2000b). Preliminary results suggest that this is possible, albeit much harder, since we are now required to search for the optimal sizes of matrices $\mathbf{B}^{(n)} \in \mathbb{R}^{I_n \times J_n}$, $1 \leq n \leq N$, as shown in Eq. (7.6). Once found, the Tucker decomposition has the advantage that it is a more natural decomposition since it does not necessarily require the repeated approximation of residual tensors. The core tensor $\mathcal{C} = (c_{j_1 \cdots j_N})$ of the Tucker decomposition gives additional insight into how the original tensor and hence the problem decomposes.

We also mention that tensor decompositions such as the rank-K approximation of Section 7.2 can be useful for approximate inference. Earlier work (Savický and Vomlel, 2006; van Gerven, 2006) has shown that each family of nodes in a Bayesian network can be replaced by the graphical model equivalent of a tensor decomposition, as shown in Fig. 7.6. This replacement leads to sparser networks and therefore more efficient probabilistic inference (Fig. 8.1). The increase in efficiency depends on the size of the hidden node, which in turn depends on the quality of the tensor approximation. This approach to approximate inference is currently under investigation (van Gerven, 2007a).

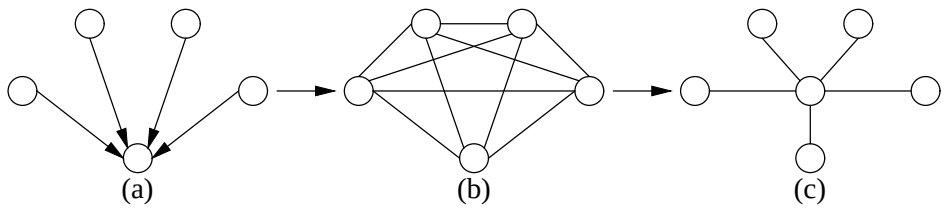


Figure 8.1: Moralization of (a) leads to the dense network (b) whereas a tensor decomposition of (b) leads to the sparse network (c).

In Section 7.3 we looked at the noisy-threshold classifier. At the moment the technique is restricted to binary variables and a threshold function as the interaction function. Various extensions that increase the applicability of the noisy-threshold classifier are possible. One can think here of the incorporation of graded or continuous variables that allow for a more natural representation of risk factors such as abdominal pain or fever, a focus on more general interaction functions, or the incorporation of time, analogous to the generalization of noisy-or models to temporal noisy-or models as was realized in (Galán and Díez, 2002). Furthermore, lifting the assumption of independence of causal influence by allowing multiple causes to influence the same hidden variable may lead to more realistic models. We leave these extensions as topics for further research.

8.3 Concluding remarks

In this thesis, we have advocated the use of (Bayesian) probability theory as the method of choice for reasoning under uncertainty in medicine while the decision-theoretic notion of utility motivates the clinical decisions that are being made. This begs the question of how physicians reason in practice. Do they act according to probability theory when making an inference and do they act according to decision theory when making a decision? In other words, are probability and decision theory just normative (describing optimal problem solving for a rational agent) or descriptive as well (describing optimal problem solving in humans)? The literature about the cognitive biases and heuristics displayed in humans in general (Kahneman et al., 1982) and physicians in particular (Chapman and Elstein, 2000; Borstein and Emier, 2001) suggests no. However, recent research has also demonstrated that some of the biases disappear when questions are posed in a less artificial way (Cosmides and Tooby, 1996; Gigerenzer, 2000). The emerging framework of naturalistic decision-making (Klein et al., 1993) recognizes the importance of these observations, and dictates that we should consider decision-making in a natural setting, where we need to deal with stress, time pressure, fatigue, and communication patterns, as well as with the bounded rationality of humans due to information-processing constraints (Simon, 1955). Under that interpretation, heuristics are not viewed as erroneous, but rather as effective strategies for real-world decision-making (Patel et al., 2002).

One particularly influential view of clinical problem solving is the *hypothetico-deductive approach* (Elstein et al., 1978), which is an iterative process where hypotheses are generated according to the available data, and hypotheses in turn guide the selection of new data. It is found that expert physicians generate the correct hypothesis early on and use the remaining time to confirm and refine the hypothesis, whereas less experienced physicians take longer to decide upon the final hypothesis due to an inability to eliminate incorrect alternatives (Joseph and Patel, 1990). Another observation is that although expert physicians have more extensive knowledge about pathophysiological processes, they tend to make less use of it than non-experts, and base themselves more on clinical experience. One explanation of this effect is the notion of *knowledge encapsulation*, which suggests that explicit pathophysiological knowledge is represented by the expert in compiled form, while still being retrievable if necessary (Boshuizen and Schmidt, 1992). The picture which emerges, is one where expert physicians rapidly recognize the correct hypothesis (possibly aided by heuristics) while still being able to give a causal explanation of how they arrive at a hypothesis. Our experiences during the construction of the carcinoid model of Chapter 6 suggest that expert physicians may indeed operate in this way. During the initial phase of knowledge elicitation the physician often jumped to conclusions, associating findings with expected outcomes, whereas after requiring a causal explanation, it became possible to explain these associations in terms of cause-effect relations.

These two modes of operation also relate to the difference between Bayesian clas-

sifiers and causal Bayesian networks, which has been stressed throughout this thesis. This distinction between associative and causal models had already been recognized by knowledge engineers of the 1970s, where early associative expert systems such as Internist-1 (Miller and Pople, 1982) were observed to suffer from the lack of pathophysiological knowledge (Schwartz et al., 1987), leading to the development of causal expert systems such as CasNet/Glaucoma (Weiss et al., 1978b; Kulikowski and Weiss, 1982). The distinction between associative and causal modes of operation also has computational consequences. Associative models have the benefit that they can be both learnt as well as operated more efficiently than causal models, albeit at the expense of offering a less accurate representation of the underlying domain knowledge. This suggests a computational strategy for Bayesian networks, where a Bayesian classifier is used to quickly generate a small set of possible hypotheses which can be subsequently fed into a causal Bayesian network of higher complexity in order to generate more accurate probability estimates. In fact, the strategy of reasoning at multiple levels of detail has already been used in the Abel expert system (Patil, 1981; Szolovits and Pauker, 1993). In earlier work (van Gerven and Lucas, 2004b), we have shown how the causal Bayesian network depicted in Fig. 3.7 can be transformed into a forest-augmented naive Bayes classifier (Lucas, 2004); a process reminiscent of knowledge encapsulation in domain experts.

From the point of view of knowledge engineering, we emphasize once more that the translation of a physician's knowledge into a graphical model is difficult and time-consuming, which implies a trade-off between the amount of time one is willing to spend and the quality of the resulting system. When one believes intervention to be the ultimate goal of clinical reasoning, associative models can perform as well as causal models provided that they lead to the same actions. Associative models, such as those of Section 3.2 and Chapter 7, can show acceptable performance and can be constructed with minimal effort. However, expert systems research has shown that not just intervention but also the *explanation* of intervention is a necessary ingredient of clinical decision support systems, since drawn conclusions must be justifiable to the physician who is responsible for patient care (Teach and Shortliffe, 1984). Furthermore, it is a characteristic of associative models that they are difficult to extend as new knowledge becomes available (Schwartz et al., 1987). Therefore, if the aim is to create a flexible system that represents domain knowledge with a high degree of detail, then one should consider building a causal model and follow the construction steps of Section 3.3 as illustrated by the carcinoid model of Chapter 6.

At the start of the twenty-first century, artificial intelligence finds itself in an exciting position, where the traditional analysis of human problem solving can finally be combined with mathematically sound inference techniques in order to create high-quality expert systems for complex domains. This thesis presented ideas concerning the use of decision-theoretic principles for the purpose of clinical decision support. It is hoped that these ideas find their way from proof of concept to the actual improvement of quality of life.

Appendix A

The Carcinoid Model

In this appendix, we show a full representation of the carcinoid model of Chapter 6. In order to depict the full model, we use an object-oriented representation. Figure A.1 makes clear that such a representation becomes a necessity for complex domains, since otherwise model construction becomes infeasible. Figure A.2 shows this object-oriented representation of the carcinoid model. The 62 depicted nodes encapsulate a total of 218 variables and 74 342 CPT entries.

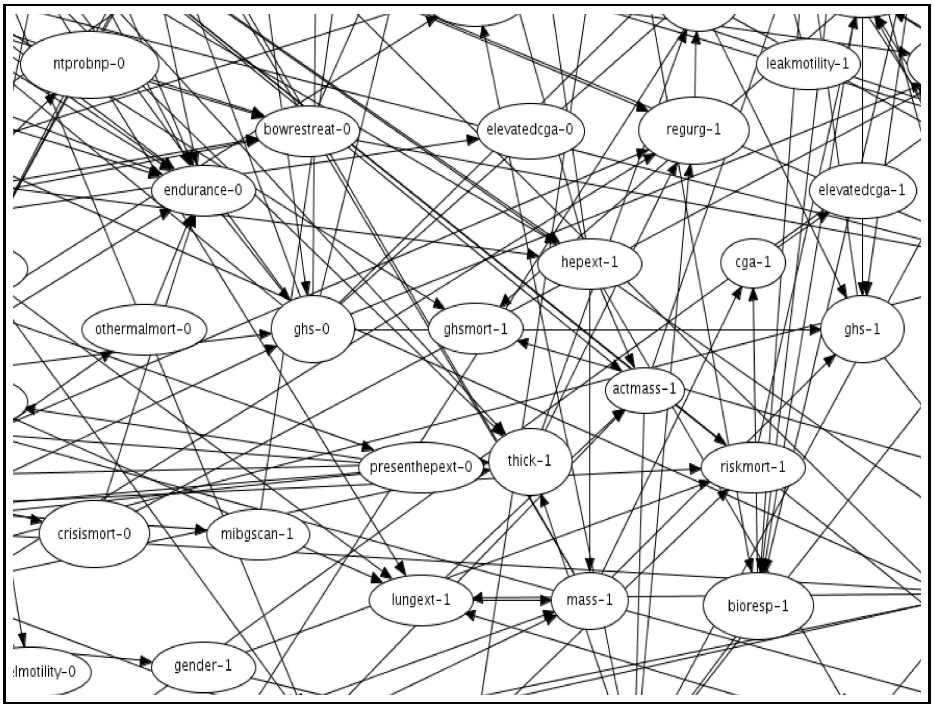


Figure A.1: Fragment of the carcinoid model.

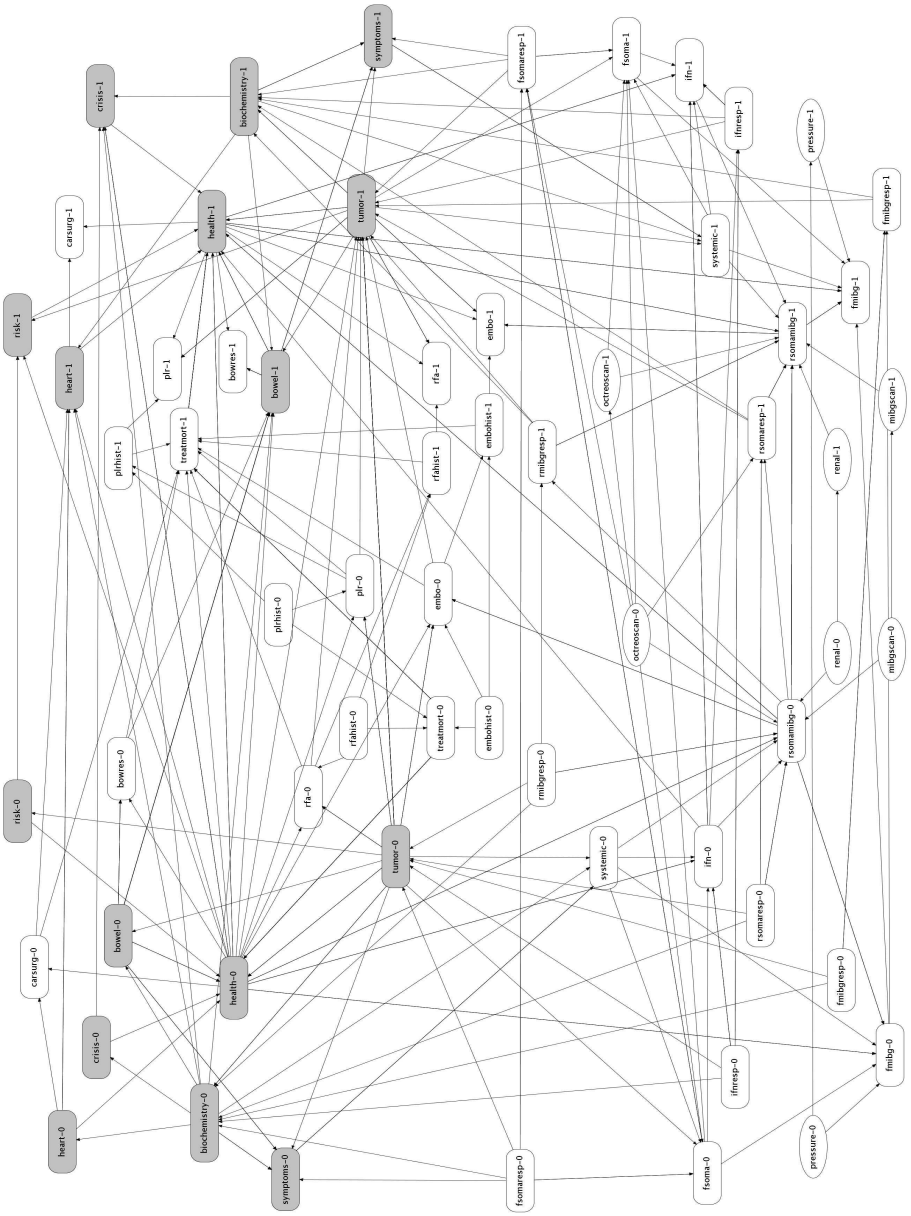


Figure A.2: The carcinoid model, as given by an object-oriented representation of the prior and transition model, where shaded nodes represent the pathophysiological component and unshaded nodes represent the treatment component of the carcinoid model.

References

Abramson, B. and Ng, K.-C. (1993). Toward an art and science of knowledge engineering: A case for belief networks. *IEEE Trans Knowl Data Eng*, 5(4):705–712.

Abu-Hanna, A. and Lucas, P. J. F. (2001). Prognostic models in medicine: AI and statistical approaches. *Meth Inform Med*, 40:1–5.

Aikins, J. S. (1983). Prototypical knowledge for expert systems. *Artif Intell*, 20(2):163–210.

Aikins, J. S., Kunz, J. C., Shortliffe, E. H., and Fallat, R. J. (1983). PUFF: an expert system for interpretation of pulmonary function data. *Comput Biomed Res*, 16:199–208.

Allen, J. F. (1984). Towards a general theory of action and time. *Artif Intell*, 23:123–154.

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Stat Med*, 19:453–473.

Andreassen, S., Benn, J. J., Hovorka, R., Olesen, K. G., and Carson, E. R. (1994). A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. *Comput Meth Programs Biomed*, 41:153–165.

Andreassen, S., Woldbye, M., Falck, B., and Andersen, S. (1987). MUNIN – a causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of Tenth International Joint Conference on Artificial Intelligence*, pages 366–372, Milan, Italy.

Aström, K. J. (1965). Optimal control of Markov decision processes with incomplete state estimation. *J Math Anal Appl*, 10:174–205.

Atkinson, R. C. and Shiffrin, R. M. (1971). The control of short-term memory. *Sci Am*, 225:82–90.

Augusto, J. C. (2005). Temporal reasoning for decision support in medicine. *Artif Intell Med*, 33:1–24.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol*, 12:387–415.

Baron, J. (1994). *Thinking and Deciding*. Cambridge University Press, Cambridge, UK, 2nd edition.

Bast-Jr, R. C., Kufe, D. W., Pollock, R. E., Weichselbaum, R. R., Holland, J. F., Frei III, E., and Gansler, T. S. (2000). *Cancer Medicine e.5*. B.C. Decker Inc.

Beck, J. R. and Pauker, S. G. (1983). The Markov process in medical prognosis. *Med Decis Making*, 3(4):420–434.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Ben-Bassat, M., Carlson, V. K., Puri, V. K., Davenport, M. D., Schriver, J. A., Latif, M. M., Smith, R., Lipnick, E. H., and Weil, M. H. (1980). Pattern-based interactive diagnosis of multiple disorders: The MEDAS system. *IEEE Trans Pattern Anal Mach Intell*, 2:148–160.

Berwick, D. M., Fineberg, H. V., and Weinstein, M. C. (1981). When doctors meet numbers. *Am J Med*, 71:991–998.

Bielza, C. and Shenoy, P. P. (1999). A comparison of graphical techniques for asymmetric decision problems. *Manag Sci*, 45(11):1552–1569.

Birkhoff, G. and Mac Lane, S. (1997). *A Survey of Modern Algebra*. A.K. Peters, 5th edition.

Boehm, B. W. (1988). A spiral model of software development and enhancement. *ACM IEEE Computer*, 21(5):61–72.

Borstein, B. H. and Emier, A. C. (2001). Rationality in medical decision making: A review of the literature on doctors' decision-making biases. *J Eval Clin Pract*, 7:97–107.

Boshuizen, H. P. A. and Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognit Sci*, 16:153–184.

Bouckaert, R. (1995). *Bayesian Belief Networks: From Construction to Inference*. PhD thesis, Utrecht University, Utrecht, The Netherlands.

Boutilier, C., Dean, T., and Hanks, S. (1996a). Planning under uncertainty: Structural assumptions and computational leverage. In Ghallab, M. and Milani, A., editors, *New Directions in AI Planning*, pages 157–171. IOS Press, Amsterdam, The Netherlands.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996b). Context-specific independence in Bayesian networks. In Horvitz, E. and Jensen, F.,

editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 115–123, San Francisco, CA. Morgan Kaufmann.

Boutilier, C. and Poole, D. (1996). Computing optimal policies for partially observable decision processes using compact representations. In Clancey, W. J. and Weld, D., editors, *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1168–1175, Portland, OR. AAAI Press.

Brown, G. W. (1972). Recursive sets of rules in statistical decision processes. In Bancroft, T. A. and Brown, S. A., editors, *Statistical Papers in Honor of George W. Snedecor*, pages 59–76. Iowa State University Press, Ames, Iowa.

Buchanan, B. G. and Shortliffe, E. H., editors (1984). *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA.

Buntine, W. L. (1994). Operations for learning with graphical models. *J Artif Intell Res*, 2:159–225.

Cao, C., Leong, T., Leong, A., and Seow, F. (1998). Dynamic decision analysis in medicine: A data driven approach. *Int J Med Informat*, 51:13–28.

Capella, C., Heitz, P. U., Hofler, H., Solcia, E., and Kloppel, G. (1995). Revised classification of neuroendocrine tumours of the lung, pancreas and gut. *Virchows Archiv*, 425(6):547–560.

Carroll, J. D. and Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35:283–319.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Springer-Verlag, New York, NY.

Centraal Bureau voor de Statistiek (2005). Overlevingstafels naar geslacht. *C. B. S. Periodiek*.

Chapman, G. B. and Elstein, A. S. (2000). Cognitive processes and biases in medical decision making. In Chapman, G. B. and Sonnenberg, F. S., editors, *Decision making in health care: Theory, psychology and applications*, pages 183–210, Cambridge, UK. Cambridge University Press.

Charitos, T., van der Gaag, L. C., Visscher, S., Schurink, K., and Lucas, P. J. F. (2005). A dynamic Bayesian network for diagnosing ventilator-associated pneumonia in icu patients. In Holmes, J. H. and Peek, N., editors, *Proceedings of the Tenth Intelligent Data Analysis in Medicine and Pharmacology Workshop*, pages 32–37.

Charniak, E. (1983). The Bayesian basis of common sense medical diagnosis. In Genesereth, M. R., editor, *Proceedings of the National Conference on Artificial Intelligence*, pages 70–73, Menlo Park, CA. AAAI Press.

Cheeseman, P. (1985). In defense of probability. In Joshi, A. K., editor, *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 1002–1009, San Francisco, CA. Morgan Kaufmann.

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

Chickering, D. M. (2002). Optimal structure identification with greedy search. *J Mach Learn Res*, 3:507–554.

Chickering, D. M. and Meek, C. (2006). On the incompatibility of faithfulness and monotone DAG faithfulness. *Artif Intell*, 170:653–666.

Christakis, N. A. and Lamont, E. B. (2000). Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *Br Med J*, 320:469–473.

Clancey, W. J. (1983). The advantages of abstract control knowledge in expert system design. In *Proceedings of the Third National Conference on Artificial Intelligence*, pages 74–78, Los Altos, CA. Morgan Kaufmann.

Cobb, B. R., Shenoy, P. P., and Rumí, R. (2006). Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials. *Stat Comput*, 16(3):293–308.

Coiera, E. (2003). *A Guide to Health Informatics*. Hodder & Stoughton Educational, London, UK.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, London, UK, 2nd edition.

Console, L., Dupré, D. T., and Torasso, P. (1989). A theory of diagnosis for incomplete causal models. In *International Joint Conference on Artificial Intelligence*, pages 1311–1317, Detroit, MI.

Console, L., Dupr, D., and Torasso, P. (1991). On the relationship between abduction and deduction. *Journal of Logic and Computation*, 1(5):661–690.

Cooper, G. F. (1984). *NESTOR: A Computer-based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge*. PhD thesis, Medical Information Sciences, Stanford University, Stanford, CA.

Cooper, G. F. (1988). A method for using belief networks as influence diagrams. In Shachter, R., Levitt, T., Kanal, L. N., and Lemmer, J. F., editors, *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 55–63, New York, NY. Elsevier Science.

Cooper, G. F. (1990). Probabilistic inference using belief networks is NP-hard. *Artif Intell*, 42:393–405.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from databases. *Mach Learn*, 9:309–347.

Cortes, C. and Mohri, M. (2004). AUC optimization vs. error rate minimization. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Cosmides, L. and Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58(1):1–73.

Coupé, V. H. M., van der Gaag, L. C., and Habbema, J. D. F. (1999). Sensitivity analysis: An aid for belief-network quantification. Technical Report UU-CS-1999-13, Utrecht University, Utrecht, The Netherlands.

Covaliu, Z. and Oliver, R. M. (1995). Representation and solution of decision problems using sequential decision diagrams. *Manag Sci*, 41(12):1860–1881.

Cowell, R., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York, NY.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J Roy Stat Soc B*, 34:197–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London, UK.

Cox, R. T. (1946). Probability frequency and reasonable expectation. *Am J Phys*, 14(1):1–13.

Cozman, F. G. (2004). Axiomatizing noisy-OR. In López de Mántaras, R. and Saitta, L., editors, *European Conference on Artificial Intelligence*, pages 979–980. IOS Press, Amsterdam.

Cruz, J. A. and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Canc Informat*, 2:59–78.

Cullen, J. and Bryman, A. (1988). The knowledge acquisition bottleneck: Time for reassessment. *Expet Syst*, 5(3):216–225.

Dagum, P. and Galper, A. (1993). Forecasting sleep apnea with dynamic network models. In Heckerman, D. and Mamdani, A., editors, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 64–71, San Francisco, CA. Morgan Kaufmann.

Dagum, P. and Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artif Intell*, 60(1):141–153.

Dawid, A. P. (1979). Conditional independence in statistical theory. *J Roy Stat Soc*, 41:1–31.

de Dombal, F. T., Leaper, D. J., Staniland, J. R., Horrocks, J. C., and McCann, A. P. (1972). Computer aided diagnosis of acute abdominal pain. *Br Med J*, 2:9–13.

de Lathauwer, L. (1997). *Signal Processing Based on Multilinear Algebra*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

de Lathauwer, L., de Moor, B., and Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM J Matrix Anal Appl*, 21:1253–1278.

de Lathauwer, L., de Moor, B., and Vandewalle, J. (2000b). On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM J Matrix Anal Appl*, 21(4):1324–1342.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Comput Intell*, 5(3):142–150.

Dean, T. and Wellmann, M. (1991). *Planning and Control*. Morgan Kaufmann, San Mateo, CA.

Dechter, R. and Rish, I. (1994). Directional resolution: The Davis-Putnam procedure, revisited. In Doyle, J., Sandewall, E., and Torasso, P., editors, *Principles of Knowledge Representation and Reasoning*, pages 134–145. Morgan Kaufmann, San Francisco, CA.

Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med*, 34(2):113–127.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc*, 39:1–38.

D'Herbomez, M. and Gouze, V. (2002). Chromogranin: A marker of neuroendocrine tumours. *Ann Biol Clin*, 60:641–646.

Diestel, R. (2000). *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, Germany, 2nd edition.

Díez, F. J. (1993). Parameter adjustment in Bayes networks. The generalized noisy OR-gate. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 99–105, San Francisco, CA. Morgan Kaufmann.

Díez, F. J., Mira, J., Iturralde, E., and Zubillaga, S. (1997). DIAVAL: a Bayesian expert system for electrocardiography. In *Artif Intell Med*, volume 10, pages 59–73.

Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach Learn*, 29:103–130.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, San Francisco, CA.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., and Stoddart, G. L. (2005). *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, Oxford, UK, 3rd edition.

Druzdzel, M. J. (1997). Five useful properties of probabilistic knowledge representations from the point of view of intelligent systems. *Fundamenta Informaticae*, 30(3–4):241–254.

Druzdzel, M. J. and Díez, F. J. (2003). Combining knowledge from different sources in probabilistic models. *J Mach Learn Res*, 4:295–316.

Druzdzel, M. J. and Henrion, M. (1993a). Efficient reasoning in qualitative probabilistic networks. In Fikes, R. and Lehnert, W., editors, *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 548–553, Menlo Park, CA. AAAI Press.

Druzdzel, M. J. and Henrion, M. (1993b). Intercausal reasoning with uninstantiated ancestor nodes. In Heckerman, D. E. and Mamdani, A., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 317–325, San Mateo, CA. Morgan Kaufmann.

Druzdzel, M. J., Oniśko, A., Schwartz, D., Dowling, J. N., and Wasyluk, H. (1999). Knowledge engineering for very large decision-analytic medical models. Technical Report CMBI-99-26, University of Pittsburgh, Pittsburgh, PA.

Druzdzel, M. J., van der Gaag, L. C., Henrion, M., and Jensen, F. V. (1995). Building probabilistic networks: where do the numbers come from? - a guide to the literature. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, New York, NY.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press, New York, NY.

Eglese, R. W. (1990). Simulated annealing: A tool for operational research. *Eur J Oper Res*, 46:271–281.

Elstein, A. S., Shulman, L. S., and Sprafka, S. A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press, Cambridge, MA.

Elston, R. and Stewart, A. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered*, 21:523–542.

Enderton, H. B. (1972). *A Mathematical Introduction to Logic*. Academic Press, Inc., New York, NY.

Eriksson, B. K., Larsson, E. G., Skogseid, B. M., Löfberg, A. M., Lörelius, L. E., and Öberg, K. E. (1998). Liver embolizations of patients with malignant neuroendocrine gastrointestinal tumors. *Cancer*, 83:2293–2301.

Feelders, A. and van der Gaag, L. C. (2006). Learning Bayesian network parameters under order constraints. *Internat J Approx Reason*, 42(1–2):37–53.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Mach Learn*, 29:131–163.

Fryback, D. G. (1978). Bayes' theorem and conditional nonindependence of data in medical diagnosis. *Comput Biomed Res*, 11:423–434.

Galán, S. F. and Díez, F. J. (2002). Networks of probabilistic events in discrete time. *Internat J Approx Reason*, 30:181–202.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*, 6:721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 24:1317–1399.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the real world*. Oxford University Press, New York, NY.

Glare, P., Virik, K., Jones, M., Hudson, M., Eychmuller, S., Simes, J., and Christakis, N. (2003). A systematic review of physicians' survival predictions in terminally ill cancer patients. *Br Med J*, 327(7408):195–198.

Good, I. J. (1983). *Good Thinking: the Foundations of Probability and its Applications*. University of Minnesota Press, Mineapolis, MN.

Gorry, G. A. (1973). Computer-assisted clinical decision-making. *Meth Inform Med Supplement*, 7:215–230.

Grimmett, G. R. and Stirzaker, D. R. (1992). *Probability and Random Processes*. Clarendon Press, Oxford, UK.

Groeger, J. S., Lemeshow, S., Price, K., Nierman, D. M., White Jr, P., Klar, J., Granovsky, S., Horak, D., and Kish, S. K. (1998). Multicenter outcome study of cancer patients admitted to the intensive care unit: a probability of mortality model. *J Clin Oncol*, 16:761–770.

Guestrin, C., Koller, D., and Parr, R. (2001). Solving factored POMDPs with linear value functions. In Nebel, B., editor, *IJCAI-01 Workshop on Planning under Uncertainty and Incomplete Information*, pages 67–65, Seattle, Washington.

Haddix, A., Teutsch, S., Shaffer, P., and Dunet, D. (1996). *Prevention Effectiveness: A Guide to Decision Analysis and Economic Evaluation*. Oxford University Press, Oxford, UK.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Model and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84.

Hart, A. and Wyatt, J. C. (1990). Evaluating black boxes as medical decision-aids: issues arising from a study of neural networks. *Int J Med Informat*, 15:229–236.

Hasman, A. and Takeda, H. (2003). Quality of health care: Informatics foundations. *Meth Inform Med*, 42(5):509–518.

Håstad, J. (1990). Tensor rank is NP-complete. *Journal of Algorithms*, 11:644–654.

Hauskrecht, M. and Fraser, H. (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artif Intell Med*, 18:221–244.

Heckerman, D. E. (1989). A tractable inference algorithm for diagnosing multiple diseases. In Henrion, M., Shachter, R., Kanal, L. N., and Lemmer, J. F., editors, *Proceedings of the Fifth Conference on Uncertainty in Artificial Intelligence*, pages 174–181, New York, NY. Elsevier Science.

Heckerman, D. E. (1990). Probabilistic similarity networks. *Networks*, 20(5):607–636.

Heckerman, D. E. and Breese, J. (1994). A new look at causal independence. In López de Mántaras, R. and Poole, D., editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 286–292, San Francisco, CA. Morgan Kaufmann.

Heckerman, D. E. and Breese, J. (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE Trans Syst Man Cybern*, 26:826–831.

Heckerman, D. E., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach Learn*, 20(3):197–243.

Heckerman, D. E. and Nathwani, B. N. (1992a). Toward normative expert systems: Part I the pathfinder project. *Meth Inform Med*, 31:90–105.

Heckerman, D. E. and Nathwani, B. N. (1992b). Toward normative expert systems: Part II probability-based representations for efficient knowledge acquisition and inference. *Meth Inform Med*, 31:106–116.

Henrion, M. (1989). Some practical issues in constructing belief networks. In Lemmer, J. F., Levitt, T., and Kanal, L. N., editors, *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pages 161–173, New York, NY. Elsevier Science.

Henrion, M. and Druzdzel, M. J. (1991). Qualitative propagation and scenario-based approaches to explanation in probabilistic reasoning. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 17–32, New York, NY. Elsevier Science.

Hernando, M. E., Gómez, E. J., del Pozo, F., and Corcoy, R. (1996). Diabnet: a qualitative model-based advisory system for therapy planning in gestational diabetes. *Int J Med Informat*, 21(4):359–374.

Hilden, J. and Habbema, J. D. F. (1987). Prognosis in medicine: an analysis of its meaning and role. *J Theor Med*, 8:249–365.

Horvitz, E. and Heckerman, D. E. (1986). The inconsistent use of measures of uncertainty in artificial intelligence research. In Kanal, L. N. and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence*, pages 137–151. Amsterdam: North-Holland.

Horvitz, E., Heckerman, D. E., and Langlotz, C. P. (1986). A framework for comparing formalisms for plausible reasoning. In *Proceedings of the National Conference on Artificial Intelligence*, pages 210–214.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, MA.

Howard, R. A. and Matheson, J. E. (1984a). Influence diagrams. In Howard, R. and Matheson, J., editors, *Readings in the Principles and Applications of Decision Analysis*. Strategic Decisions Group, Menlo Park, CA.

Howard, R. A. and Matheson, J. E., editors (1984b). *Readings in the Principles and Applications of Decision Analysis*. Strategic Decisions Group, Menlo Park, CA.

Jackson, P. (1990). *Introduction to Expert Systems*. International Computer Science Series. Addison-Wesley, Harlow, UK, 2nd edition.

Janson, E. M. T. and Öberg, K. (1996). Carcinoid tumours. *J Clin Gastroenterol*, 10:589–601.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jenkinson, D. (2005). The elicitation of probabilities - a review of the statistical literature. Technical report, University of Sheffield, Sheffield, UK.

Jensen, A. L. (1995). Quantification experience of a DSS for mildew management in winter wheat. In Druzdzel, M. J., van der Gaag, L. C., Henrion, M., and Jensen, F. V., editors, *Working Notes of the Workshop on Building Probabilistic Networks: Where Do The Numbers Come From?*, pages 23–31.

Jensen, F. V. (1988). Junction trees and decomposable hypergraphs. Technical Report JUDEX Research Report, Aalborg University, Aalborg, Denmark.

Johnson, P., Duran, A., Hassebrock, F., Moller, J., Prietulla, M., Feltovich, P., and Swanson, D. (1981). Expertise and error in diagnostic reasoning. *Cognit Sci*, 5:235–283.

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

Joseph, G.-M. and Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Med Decis Making*, 10:31–46.

Jurgelenaite, R. and Heskes, T. (2006). EM algorithm for symmetric causal independence models. In *Proceedings of the Seventeenth European Conference on Machine Learning*, pages 234–245, Heidelberg, Germany. Springer-Verlag.

Jurgelenaite, R., Heskes, T., and Lucas, P. J. F. (2006). Noisy threshold functions for modelling causal independence in Bayesian networks. Technical Report ICIS-R06014, Radboud University, Nijmegen, The Netherlands.

Kahn Jr, C. E., Roberts, L. M., Shaffer, K. A., and Haddawy, P. (1997). Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med*, 27:19–29.

Kahneman, D., Slovic, P., and Tversky, A., editors (1982). *Judgment under uncertainty: Heurstics and biases*. Cambridge University Press, Cambridge, UK.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *J Am Stat Assoc*, 53:457–481.

Kaplan, R. M. and Anderson, J. P. (1988). A general health policy model: Update and applications. *Health Serv Manag Res*, 23(2):203–235.

Kappen, H. J. and Neijt, J. P. (2002). Promedas, a probabilistic decision support system for medical diagnosis. Technical report, Stichting Neurale Netwerken, Nijmegen, The Netherlands.

Kim, J. H. and Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference engines. In *Proceedings of the Eight Joint International Conference on Artificial Intelligence*, Karlsruhe, Germany.

Kirkpatrick, S., Gelatt Jr, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Klein, A., Orasanu, J., Calderwood, R., and Zsambok, C. E., editors (1993). *Decision Making in Action: Models and Methods*. Ablex Publishing Corporation, Norwood, NJ.

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. Guilford, New York, NY.

Knaus, W., Wagner, D. P., Draper, E., Zimmerman, J., Bergner, M., Bastos, P., Sirio, C., Murphy, D., Lotring, T., Damiano, A., and Harrel, F. (1991a). The APACHE III prognostic system. risk prediction of hospital mortality for critically ill hospitalised adults. *Chest*, 100:1619–1636.

Knaus, W. A., Wagner, D. P., and Lynn, J. (1991b). Short term mortality predictions for critically ill hospitalised adults: Science and ethics. *Science*, 254:389–394.

Kohavi, R. and Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. In Saitta, L., editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283, San Mateo, CA. Morgan Kaufmann.

Koller, D. and Lerner, U. (2001). Sampling in factored dynamic systems. In Doucet, A., de Freitas, N., and Gordon, N., editors, *Sequential Monte Carlo Methods in Practice*, chapter 21, pages 445–464. Springer-Verlag, San Francisco, CA.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 302–313, San Francisco, CA. Morgan Kaufmann.

Kong, A. (1991). Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. *Genet Epidemiol*, 8:81–103.

Korver, M. and Lucas, P. J. F. (1993). Converting a rule-based expert system into a belief network. *Int J Med Informat*, 18(3):219–241.

Kulikowski, C. A. and Weiss, S. M. (1982). Representation of expert knowledge for consultation: The CASNET and EXPERT projects. In Szolovits, P., editor, *Artif Intell Med*, pages 21–55. Westview Press, Boulder, CO.

Kuntz, K. M. and Weinstein, M. C. (2001). Modelling in economic evaluation. In Drummond, M. and McGuire, A., editors, *Economic Evaluation in Health Care: Merging Theory with Practice*, pages 141–171. Oxford University Press, New York, NY.

Kyburg, H. and Smokler, H. (1964). *Studies in Subjective Probability*. John Wiley & Sons, New York, NY.

Lacave, C. and Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *Knowl Eng Rev*, 17(2):107–127.

Lacave, C. and Díez, F. J. (2003). Knowledge acquisition in PROSTANET - a Bayesian network for diagnosing prostate cancer. In *Int J Knowl Base Intell Eng Syst*, pages 1345–1350, Oxford, UK. Springer.

Lamberts, S. W., van der Lely, A. J., de Herder, W. W., and Hofland, L. J. (1996). Octreotide. *New Engl J Med*, 334(4):246–254.

Langley, P. and Sage, S. (1994). Induction of selective Bayesian classifiers. In López de Mántaras, R. and Poole, D., editors, *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA. Morgan Kaufmann.

Larrañaga, P., Poza, M., Yurramendi, M., Murga, R., and Kuijpers, C. (1996). Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Trans Pattern Anal Mach Intell*, 18:912–926.

Laskey, K. and Mahoney, S. (1997). Network fragments: Representing knowledge for constructing probabilistic models. In Geiger, D. and Shenoy, P. P., editors, *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 334–341, San Francisco, CA. Morgan Kaufmann.

Lauritzen, S. L. and Nilsson, D. (2001). Representing and solving decision problems with limited information. *Manag Sci*, 47(9):1235–1251.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J Roy Stat Soc B*, 50:157–224.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann Stat*, 17:31–57.

Le Gall, J.-R., Lemeshow, S. S., and Saulnier, F. (1993). A new simplified acute physiology score (SAPS II) based on a european/north american multicenter study. *J Am Med Informat Assoc*, 270:2957–2963.

Ledley, R. S. and Lusted, L. B. (1959). Reasoning foundation of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130:9–21.

Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, H. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., Whalen, R. E., and Rosati, R. A. (1986). Predicting outcome in coronary disease: statistical models vs. expert clinicians. *Am J Med*, 80(4):553–560.

Leong, T.-Y. (1994). *An Integrated Approach to Dynamic Decision Making under Uncertainty*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Lindberg, G., Thomson, C., Malchow-Møller, A., Matzen, P., and Hilden, J. (1987). Differential diagnosis of jaundice: applicability of the Copenhagen pocket diagnostic chart proven in Stockholm patients. *Liver*, 7:43–49.

Long, W. J. (1996). Temporal reasoning for diagnosis in a causal probabilistic knowledge base. *Artif Intell Med*, 8:193–215.

Lucas, P. J. F. (1995). Logic engineering in medicine. *Knowl Eng Rev*, 10(2):153–179.

Lucas, P. J. F. (1998). Analysis of notions of diagnosis. *Artif Intell*, pages 295–343.

Lucas, P. J. F. (2001). Certainty-factor-like structures in Bayesian networks. In *Knowl Base Syst*, volume 14, pages 327–335.

Lucas, P. J. F. (2004). Restricted Bayesian network structure learning. In Gâmez, J. A., Moral, S., and Salmeron, A., editors, *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, volume 146, pages 217–232. Springer-Verlag, Berlin.

Lucas, P. J. F. (2005). Bayesian network modelling by qualitative patterns. *Artif Intell*, 163:233–263.

Lucas, P. J. F. and Abu-Hanna, A. (1999). Prognostic methods in medicine. *Artif Intell Med*, 15:105–119.

Lucas, P. J. F., Boot, H., and Taal, B. G. (1998). Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 37:206–219.

Lucas, P. J. F., de Bruijn, N. C., Schurink, K., and Hoepelman, A. (2000). A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artif Intell Med*, 19:251–279.

Lucas, P. J. F. and van der Gaag, L. C. (1991). *Principles of Expert Systems*. Addison-Wesley, Wokingham, UK.

Lucas, P. J. F., van der Gaag, L. C., and Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artif Intell Med*, 30:201–214.

Lusena, C., Goldsmith, J., and Mundhenk, M. (2001). Nonapproximability results for partially observable Markov decision processes. *Artif Intell*, 14:83–103.

Macartney, F. J. (1988). Diagnostic logic. In Philips, C., editor, *Logic in Medicine*. British Medical Journal, London, UK.

Magni, P. (1998). A new approach to optimal dynamic therapy planning. In *Proc AMIA Symp*, pages 936–940, Philadelphia, PA. Hanley & Belfus.

Magni, P. and Bellazzi, R. (1997). DT-Planner: An environment for managing dynamic decision problems. *Comput Meth Programs Biomed*, 54:183–200.

Magni, P., Quaglini, S., Marchetti, M., and Barosi, G. (2000). Deciding when to intervene: A Markov decision process approach. *Int J Med Informat*, 60:237–253.

Mahoney, S. M. and Laskey, K. B. (1996). Network engineering for complex belief networks. In Horvitz, E. and Jensen, F., editors, *Uncertainty in Artificial Intelligence, Proceedings of the Twelfth Conference*, pages 389–396, San Francisco, CA. Morgan Kaufmann.

Malchow-Møller, A., Thomson, C., Matzen, P., Mindeholm, L., Bjerregaard, B., Bryant, S., Hilden, J., Holst-Christensen, J., Johansen, T. S., and Juhl, E. (1986). Computer diagnosis in jaundice: Bayes' rule founded on 1002 consecutive cases. *J Hepatol*, 3:154–163.

Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *Journal of the ACM*, 8(3):404–417.

Mazumdar, M. and Glassman, J. (2000). Tutorial in biostatistics, categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med*, 19:113–132.

McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B. and Michie, D., editors, *Machine Intelligence*, volume 4, pages 463–502, Edinburgh, UK. Edinburgh University Press.

McKay, B. D., Oggier, F. E., Royle, G. F., Sloane, N. J. A., Wanless, I. M., and Wilf, H. S. (2004). Acyclic digraphs and eigenvalues of (0,1)-matrices. *Journal of Integer Sequences*, 7(04.3.3):1–5.

Meek, C. and Heckerman, D. E. (1997). Structure and parameter learning for causal independence and causal interaction models. In Geiger, D. and Shenoy, P. P., editors, *Proceedings of Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 366–375, San Francisco, CA. Morgan Kaufmann.

Meij, V., Zuetenhorst, J. M., van Hillegersberg, R., Kroger, R., Prevoo, W., van Coevorden, F., and Taal, B. G. (2005). Local treatment in unresectable hepatic metastases of carcinoid tumors: Experiences with hepatic artery embolization and radiofrequency ablation. *World J Surg Oncol*, 3(75).

Meuleau, N., Kim, K.-E., Kaelbling, L. P., and Cassandra, A. R. (1999). Solving POMDPs by searching the space of finite policies. In Laskey, K. and Prade, H., editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 417–426, San Francisco, CA. Morgan Kaufmann.

Miller, R. A., Masarie, F. E., and Myers, J. D. (1986). "Quick Medical Reference" for diagnostic assistance. *MD Computing*, 3:34–48.

Miller, R. A. and Pople, H. E. J. (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. *New Engl J Med*, 307:468–476.

Modlin, I. M., Kidd, M., Latich, I., Zikusoka, M. N., and Shapiro, M. D. (2005). Current status of gastrointestinal carcinoids. *Gastroenterology*, 128:1717–1751.

Modlin, I. M., Shapiro, M. D., and Kidd, M. (2004). Carcinoid tumors and fibrosis: An association with no explanation. *Am J Gastroenterol*, 99:2466–2478.

Moertel, C. G., Kvols, L. K., O'Connell, M. J., and Rubin, J. (1991). Treatment of neuroendocrine carcinomas with combined etoposide and cisplatin. Evidence of major therapeutic activity in the anaplastic variants of these neoplasms. *Cancer*, 68(2):227–232.

Monahan, G. E. (1982). A survey of partially observable Markov decision processes. *Manag Sci*, 28(1):1–16.

Morgan, M. G. and Henrion, M. (1990). *Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, UK.

Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology. *J Am Stat Assoc*, 79:489–500.

Murphy, K. P. (2002). *Dynamic Bayesian Networks*. PhD thesis, UC Berkeley, Berkeley, CA.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In Laskey, K. and Prade, H., editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 467–475, San Francisco, CA. Morgan Kaufmann.

Nease, R. F. and Owens, D. K. (1997). Use of influence diagrams to structure medical decisions. *Med Decis Making*, 17(3):263–275.

Nehar, D., Lombard-Bohas, C., Olivieri, S., Claustrat, B., Chayvialle, J.-A., Penes, M.-C., Sassolas, G., and Borson-Chazot, F. (2004). Interest of chromogranin a for diagnosis and follow-up of endocrine tumours. *Clin Endocrinol*, 60:644–652.

Neil, M. and Fenton, L. (2000). Building large-scale Bayesian networks. *Knowl Eng Rev*, 15(3):257–284.

Newell, A. and Simon, H. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.

Ng, A. Y. and Jordan, M. I. (2000). PEGASUS: A policy search method for large MDPs and POMDPs. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 406–415, San Francisco, CA. Morgan Kaufmann.

Nobels, F. R. E., Kwekkeboom, D. J., Bouillon, R., and Lamberts, S. W. J. (1998). Chromogranin A: Its clinical values as marker of endocrine tumours. *Eur J Clin Investig*, 28:431–440.

Öberg, K. and Eriksson, B. (1991). The role of interferon in the management of carcinoid tumors. *Br J Haematol*, 79(1):74–77.

Öberg, K., Theodorsson-Norheim, E., and Norheim, I. (1987). Motilin in plasma and tumor tissues from patients with the carcinoid syndrome. possible involvement in the increased frequency of bowel movements. *Scand J Gastroenterol*, 22:1041–1048.

Offringa, M., Assendelft, W. J. J., and Scholten, R. J. P. M., editors (2003). *Inleiding in Evidence-Based Medicine*. Bohn Stafleu van Loghum, Houten, The Netherlands.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, New York, NY.

Ohno-Machado, L. (1997). A comparison of Cox proportional hazards and artificial neural network models for medical prognosis. *Comput Biol Med*, 27:55–65.

Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., and Carbone, P. P. (1982). Toxicity and response criteria of the eastern cooperative oncology group. *Am J Clin Oncol*, 5:649–655.

Olesen, K. G., Kjaerulff, U., Jensen, F., Jensen, F. V., Falck, B., Andreassen, S., and Andersen, S. K. (1989). A MUNIN network for the median nerve - a case study on loops. *Appl Artif Intell*, 3:301–319.

Olmsted, S. M. (1983). *On Representing and Solving Decision Problems*. PhD thesis, Stanford University, Stanford, CA.

Osler, W. (1906). *Aequanimitas. With other addresses to medical students, nurses and practitioners of medicine*. Blakiston's Son & Co., Philadelphia, PA.

Owens, D. K., Shachter, R. D., and Nease, R. F. (1997). Representation and analysis of medical decision problems with influence diagrams. *Med Decis Making*, 17(3):241–262.

Papadimitriou, C. H. and Tsitsiklis, J. N. (1987). The complexity of Markov decision processes. *Math Oper Res*, 12(3):441–450.

Patel, V. L., Arocha, J. F., and Zhang, J. (2004). Thinking and reasoning in medicine. In Holyoak, K. J. and Morrison, R. G., editors, *Cambridge Handbook of Thinking and Reasoning*, pages 727–751. Cambridge University Press, Cambridge, UK.

Patel, V. L., Kaufman, D. R., and Arocha, J. F. (2002). Emerging paradigms of cognition in medical decision-making. *J Biomed Informat*, 35:52–75.

Patil, R., Szolovits, P., and Schwartz, W. (1982). Modelling knowledge of the patient in acid-base and electrolyte disorders. In Szolovits, P., editor, *Artificial Intelligence in Medicine*, pages 191–226. Westview Press, Boulder, CO.

Patil, R. S. (1981). *Causal Representation of Patient Illness for Electrolyte and Acid-Base Diagnosis*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Pauker, S. G., Gorry, G. A., Kassirer, J. P., and Schwartz, W. B. (1976). Toward the simulation of clinical cognition: Taking a present illness by computer. *Am J Med*, 60:981–995.

Pearl, J. (1987). Evidential reasoning using stochastic simulation of causal models. *Artif Intell*, 32(2):245–257.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 2nd edition.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY.

Pearl, J. and Paz, A. (1985). GRAPHOIDS: A graph-based logic for reasoning about relevance relations. In Du Boulay, B., Hogg, D., and Steels, L., editors, *Advances in Artificial Intelligence 2*, pages 357–363, Amsterdam, The Netherlands.

Peek, N. B. (1999). Explicit temporal models for decision-theoretic planning of clinical management. *Artif Intell Med*, 15:135–154.

Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R., Hall, R., Johnson, P., Jones, N., Kumar, A., Miksch, S., Quaglini, S., Seyfang, A., Shortliffe, E., and Stefanelli, M. (2003). Comparing computer-interpretable guideline models: a case-study approach. *J Am Med Informat Assoc*, 10(1):52–68.

Pople, H. E. (1982). Heuristic methods for imposing structure on ill-structured problems: the structure of medical diagnosis. In Szolovits, P., editor, *Artif Intell Med*, pages 119–190. Westview Press, Boulder, CO.

Pradhan, M., Provan, G., Middleton, B., and Henrion, M. (1994). Knowledge engineering for large belief networks. In López de Mántaras, R. and Poole, D., editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 484–490, San Francisco, CA. Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Mach Learn*, 1(1):81–106.

Quinlan, J. R. (1992). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA.

Reiter, R. (1978). On closed–world data bases. In Gallaire, H. and Minker, J., editors, *Logic and Databases*, pages 55–76, New York, NY. Plenum Press.

Renooij, S. (2001). *Qualitative Approaches to Quantifying Probabilistic Networks*. PhD thesis, University of Utrecht, Utrecht, The Netherlands.

Renooij, S. and Witteman, C. L. M. (1999). Talking probabilities: communicating probabilistic information with words and numbers. *Internat J Approx Reason*, 22:169–194.

Robinson, R. W. (1973). Counting labeled acyclic digraphs. In Harary, F., editor, *New Directions in the Theory of Graphs*, pages 239–273. Academic Press, New York, NY.

Ross, S. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, NY.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, NY.

Sacha, J. P., Goodenday, L., and Cios, K. J. (2002). Bayesian learning for cardiac SPECT image interpretation. *Artif Intell Med*, 26:109–143.

Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Second International Conference on Knowledge Discovery in Databases*, pages 335–338, Portland, OR. AAAI Press.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Min Knowl Discov*, 1:317–327.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *J Am Stat Assoc*, 66:783–801.

Savický, P. and Vomlel, J. (2006). Tensor rank-one decomposition of probability tables. In Bouchon-Meunier, B. and Yager, R. R., editors, *Proceedings of the Eleventh International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems.*, pages 2292–2299.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Schadbolt, N., van de Velde, W., and Wielinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Cambridge, MA.

Schuchter, L., Schultz, D. J., Synnestvedt, M., Trock, B. J., Guerry, D., Elder, D. E., Elenitsas, R., Clark, W. H., and Halpern, A. C. (1996). A prognostic model for predicting 10-year survival in patients with primary melanoma. *Ann Intern Med*, 125(5):369–375.

Schwartz, W. B., Patil, R. S., and Szolovits, P. (1987). Artificial intelligence in medicine. Where do we stand? *New Engl J Med*, 12(11):685–688.

Shachter, R. D. and Peot, M. A. (1992). Decision making using probabilistic inference methods. In Dubois, D., Wellman, M., D'Ambrosio, B., and Smets,

P., editors, *Proceedings of the Eight Conference on Uncertainty in Artificial Intelligence*, pages 276–283, San Mateo, CA. Morgan Kaufmann.

Shenoy, P. P. (1996). Representing and solving asymmetric decision problems using valuation networks. In Fisher, D. and Lenz, H.-J., editors, *Learning from Data: Artificial Intelligence and Statistics V. Lecture Notes in Statistics*, volume 112, pages 99–108. Springer-Verlag, New York, NY.

Shenoy, P. P. (2006). Inference in hybrid Bayesian networks using mixtures of Gaussians. In *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence*, pages 428–436.

Shortliffe, E. H., Perreault, L. E., Wiederhold, G., and Fagan, L. M., editors (2001). *Medical Informatics: Computer Applications in Health Care and Biomedicine*. Springer-Verlag, New York, NY, 2nd edition.

Shortliffe, E. H., Scott, A. C., Bischoff, M. B., van Melle, W., and Jacobs, C. D. (1981). ONCOCIN: an expert system for oncology protocol management. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 876–881.

Shwe, M. A., Middleton, B., Heckerman, D. E., Henrion, M., Horvitz, E. J., Lehmann, H. P., and Cooper, G. F. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Meth Inform Med*, 30(4):241–255.

Simon, H. A. (1955). A behavioral model of rational choice. *Q J Econ*, 69(1):99–118.

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994). Learning without state estimation in partially observable Markovian decision processes. In Cohen, W. and Hirsh, H., editors, *Proceedings of the Eleventh Conference on Machine Learning*, pages 284–292, New Brunswick, NJ. Morgan Kaufmann.

Skinazi, F., Zins, M., Menu, Y., Bernades, P., and Ruszniewski, P. (1996). Liver metastases of digestive endocrine tumors. natural history and response to medical treatment. *Eur J Gastroenterol Hepatol*, 8:673–678.

Smallwood, R. D. and Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Oper Res*, 21(5):1071–1088.

Smallwood, R. D. and Sondik, E. J. (1978). The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Oper Res*, 26(2):282–304.

Sonnenberg, F. A. and Beck, J. R. (1993). Markov models in medical decision making: A practical guide. *Med Decis Making*, 13:322–338.

Sox, H. C., Blatt, M. A., Higgins, M. C., and Marton, K. I., editors (1988). *Med Decis Making*. Butterworth, Boston, MA.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian analysis in expert systems. *Stat Sci*, 8(3):219–283.

Spiegelhalter, D. J. and Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J Roy Stat Soc*, 147:35–77.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction and Search*, volume 81 of *Lecture Notes in Statistics*. Springer-Verlag, New York, NY, 2nd edition.

Srinivas, S. (1994). A probabilistic approach to hierarchical model-based diagnosis. In López de Mántaras, R. and Poole, D., editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 538–545, San Francisco, CA. Morgan Kaufmann.

Studený, M. (1989). Multiinformation and the problem of characterization of conditional independence relations. *Probl Contr Inform Theor*, 18:3–16.

Studený, M. (1992). Conditional independence relations have no finite complete characterization. In Kubík, S. and Vísek, J., editors, *Information Theory, Statistical Decision Functions and Random Processes: Transactions of Eleventh Prague Conference B*, pages 377–396, Dordrecht, The Netherlands. Kluwer.

Suermondt, H. J. and Cooper, G. F. (1993). An evaluation of explanations of probabilistic inference. *Comput Biomed Res*, 26:242–254.

Sutton, R., Doran, H. E., Williams, E. M. I., Vora, J., Vinjamuri, S., Evans, J., Campbell, F., Raraty, M. G. T., Ghaneh, P., Hartley, M., Poston, G. J., and Neoptolemos, J. P. (2003). Surgery for midgut carcinoids. *Endocr Relat Canc*, 10:469–481.

Szolovits, P. and Pauker, S. G. (1993). Categorical and probabilistic reasoning in medicine revisited. *Artif Intell*, 59:167–180.

Taal, B., Hoefnagel, C., Boot, H., Jong, D. D., and Rutgers, M. (1999). Carcinoide tumoren van de darm: ontwikkelingen binnen Nederland in diagnostiek en palliatieve behandeling. *Nederlands Tijdschrift voor de Geneeskunde*, 143(9):445–451.

Taal, B. G. and Smits, M. (2005). Developments in diagnosis and treatment of metastatic midgut carcinoid tumours. *Minerva Gastroenterol*, 51:335–344.

Taal, B. G. and Visser, O. (2004). Epidemiology of neuroendocrine tumours. *Neuro-endocrinology*, 80:3–7.

Tatman, J. A. and Shachter, R. D. (1990). Dynamic programming and influence diagrams. *IEEE Trans Syst Man Cybern*, 20(2):365–379.

Teach, R. L. and Shortliffe, E. H. (1984). An analysis of physicians' attitudes. In Buchanan, B. G. and Shortliffe, E. H., editors, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, Mass.

Tucker, L. R. (1966). Some mathematical notes of three-mode factor analysis. *Psychometrika*, 31:279–311.

Turban, E. (1992). *Expert Systems and Applied Artificial Intelligence*. Macmillan, New York, NY.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognit Psychol*, 5:207–232.

van der Gaag, L. C. and Helsper, E. M. (2002). Experiences with modelling issues in building probabilistic networks. In Gómez-Pérez, A. and Benjamins, V. R., editors, *EKAW 2002*, volume 2473 of *Lecture Notes in Computer Science*, pages 21–26, Berlin. Springer-Verlag.

van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., and Taal, B. G. (1999). How to elicit many probabilities. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 647–654.

van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P., and Taal, B. G. (2001). Probabilities for a probabilistic network: A case-study in oesophageal carcinoma. Technical Report UU-CS-2001-01, University of Utrecht, Utrecht, The Netherlands.

van Dijk, S., van der Gaag, L. C., and Thierens, D. (2003). A skeleton-based approach to learning Bayesian networks from data. In Lavrac, N., Gamberger, D., Todorovski, L., and Blockeel, H., editors, *Proceedings of the Seventh Conference on Principles and Practice of Knowledge Discovery in Databases*, volume 2838 of *Lecture Notes in Computer Science*, pages 132–143. Springer.

van Eeden, S., Quaedvlieg, P. F. H. J., Taal, B. G., Offerhaus, G. J. A., Lamers, C. B. H. W., and van Velthuysen, M. F. (2002). Classification of low-grade neuroendocrine tumors of midgut and unknown origin. *Hum Pathol*, 33(11):1126–1132.

van Gerven, M. A. J. (2006). Efficient Bayesian inference by factorizing conditional probability distributions. Technical Report ICIS-R6032, Radboud University, Nijmegen, The Netherlands.

van Gerven, M. A. J. (2007a). Approximate inference in graphical models using tensor decompositions. In *Proceedings of the Eighteenth Belgium-Netherlands Conference on Artificial Intelligence*. Submitted for publication.

van Gerven, M. A. J. (2007b). Tensor decompositions for probabilistic classification. In *Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) 2007*. In press.

van Gerven, M. A. J. and Díez, F. J. (2006). Selecting strategies for infinite-horizon dynamic LIMIDs. In Studený, M. and Vomlel, J., editors, *Proceedings of the Third European Workshop on Probabilistic Graphical Models*, pages 131–138, Prague, Czech Republic. Action M Agency.

van Gerven, M. A. J., Díez, F. J., Taal, B. G., and Lucas, P. J. F. (2006a). Prognosis of high-grade carcinoid tumor patients using dynamic limited-memory influence diagrams. In Peek, N. and Combi, C., editors, *Intelligent Data Analysis in bioMedicine and Pharmacology (IDAMAP) 2006*, pages 65–70.

van Gerven, M. A. J., Díez, F. J., Taal, B. G., and Lucas, P. J. F. (2006b). Selecting treatment strategies with dynamic limited-memory influence diagrams. *Artif Intell Med*. In press.

van Gerven, M. A. J., Jurgelenaite, R., Taal, B. G., Heskes, T., and Lucas, P. J. F. (2007a). Predicting carcinoid heart disease with the noisy-threshold classifier. *Artif Intell Med*, 40(1):45–55.

van Gerven, M. A. J. and Lucas, P. J. F. (2004a). Employing maximum mutual information for Bayesian classification. In *Biological and Medical Data Analysis*, volume 3337 of *Lecture Notes in Computer Science*, pages 188–199. Springer, Berlin, Germany.

van Gerven, M. A. J. and Lucas, P. J. F. (2004b). Using background knowledge to construct Bayesian classifiers for data-poor domains. In Bramer, M., Coenen, F., and Allen, T., editors, *Proceedings of AI-2004, the Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 269–282, London, UK. Springer-Verlag.

van Gerven, M. A. J. and Lucas, P. J. F. (2007). The role of background knowledge in Bayesian classification. In Lucas, P. J. F., Gámez, J. A., and Salmerón, A., editors, *Advanced in Probabilistic Graphical Models*, StudFuzz 213, pages 377–396. Springer-Verlag, Berlin Heidelberg.

van Gerven, M. A. J., Lucas, P. J. F., and van der Weide, T. P. (2005). A qualitative characterisation of causal independence models using Boolean polynomials. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 3571 of *Lecture Notes in Computer Science*, pages 244–256, Berlin, Germany. Springer.

van Gerven, M. A. J., Lucas, P. J. F., and van der Weide, T. P. (2006c). A qualitative characterization of causal independence. *Internat J Approx Reason*. Accepted for publication.

van Gerven, M. A. J. and Taal, B. G. (2006). Structure and parameters of a Bayesian network for carcinoid prognosis. Technical Report ICIS-R6033, Radboud University Nijmegen, Nijmegen, The Netherlands.

van Gerven, M. A. J., Taal, B. G., and Lucas, P. J. F. (2007b). A probabilistic model for carcinoid prognosis. *J Biomed Inform*. Submitted for publication.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London, UK, 2nd edition.

Vomlel, J. (2002). Exploiting functional dependence in Bayesian network inference. In Boutilier, C. and Goldszmidt, M., editors, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 528–535, San Francisco, CA. Morgan Kaufmann.

Von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley & Sons, New York, NY.

Wang, H. and Ahuja, N. (2004). Compact representation of multidimensional data using tensor rank-one decomposition. In *International Conference on Pattern Recognition*, pages 44–47. IEEE.

Warner, H. R., Toronto, A. F., Veasy, L. G., and Stephenson, R. (1961). A mathematical approach to medical diagnosis: Application to congenital heart disease. *J Am Med Informat Assoc*, 177:177–183.

Wasyluk, H., Oniśko, A., and Druzdzel, M. J. (2001). Support of diagnosis of liver disorders based on a causal Bayesian network model. *Med Sci Mon*, 7(1):327–332.

Wegener, I. (1987). *The Complexity of Boolean Functions*. John Wiley & Sons, New York, NY.

Weinstein, M. and Stason, W. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New Engl J Med*, 296(13):716–721.

Weiss, S. M. and Kulikowski, C. A. (1984). *A Practical Guide to Designing Expert Systems*. Rowman & Littlefield, Totowa, NJ.

Weiss, S. M., Kulikowski, C. A., Amarel, S., and Safir, A. (1978a). A model-based method for computer assisted medical decision making. *Artif Intell*, 11:145–172.

Weiss, S. M., Kulikowski, C. A., and Safir, A. (1978b). Glaucoma consultation by computer. *Comput Biol Med*, 8:24–40.

Wellman, M. P. (1990). Fundamental concepts of qualitative probabilistic networks. *Artif Intell*, 44:257–303.

Whitehead, S. D. and Ballard, D. H. (1991). Learning to perceive and act by trial and error. *Mach Learn*, 7:45–83.

Wiegerinck, W. (2005). Modeling Bayesian networks by learning from experts. In *Proceedings of the seventienth Belgium-Netherlands conference on artificial intelligence*, pages 305–310.

Wiegerinck, W. and Heskes, T. (2001). Probability assessment with maximum entropy in Bayesian networks. *Computing Science and Statistics*, 33.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, 2nd edition.

Woolf, S. H. (2000). Evidence-based medicine and practical guidelines: an overview. *Cancer Control*, 7(4):362–367.

Wu, X., Lucas, P. J. F., Kerr, S., and Dijkhuizen, R. (2001). Learning Bayesian-network topologies in realistic medical domains. In *Proceedings of the Second International Symposium on Medical Data Analysis*, volume 2199 of *Lecture Notes In Computer Science*, pages 302–308. Springer-Verlag, London, UK.

Wyatt, J. C. and Altman, D. G. (1995). Prognostic models: clinically useful or quickly forgotten? *Br Med J*, 311:1539–1541.

Wyatt, J. C. and Spiegelhalter, D. J. (1990). Evaluating medical expert systems: what to test and how? *Int J Med Informat*, 15(3):205–217.

Yuan, C. and Druzdzel, M. J. (2005). Importance sampling algorithms for Bayesian networks: Principles and performance. *Math Comput Model*, 43:1189–1207.

Zhang, N. and Poole, D. (1994). A simple approach to Bayesian network computations. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence*, pages 171–178.

Zhang, N. and Poole, D. (1996). Exploiting causal independence in Bayesian network inference. *J Artif Intell Res*, 5:301–328.

Zhang, T. and Golub, G. (2001). Rank-one approximation to high order tensors. *SIAM J Matrix Anal Appl*, 23:534–550.

Zuetenhorst, J., Bonfrer, J., Korse, C., Bakker, R., van Tinteren, H., and Taal, B. G. (2003). Carcinoid heart disease; the role of urinary 5-HIAA excretion and plasma levels of TGF-β and FGF. *Cancer*, 97:1609–1615.

Zuetenhorst, J. M., Korse, C. M., Bonfrer, J. M. G., Bakker, R. H., and Taal, B. G. (2004). The role of natriuretic peptides in the diagnosis and treatment of patients with carcinoid heart disease. *Br J Canc*, 90:2073–2079.

Zuetenhorst, J. M. and Taal, B. G. (2003). Carcinoid heart disease. *New Engl J Med*, 348:2359–2361.

Zuetenhorst, J. M. and Taal, B. G. (2005). Metastatic carcinoid tumors: A clinical review. *The Oncologist*, 10(2):123–131.

Zuetenhorst, J. M., Taal, B. G., Boot, H., Valdés Olmos, R., and Hoefnagel, C. (1999). Long-term palliation in metastatic carcinoid tumours with various applications of meta-iodobenzylguanidin (MIBG): Pharmacological MIBG, ^{131}I-labelled MIBG and the combination. *Eur J Gastroenterol Hepatol*, 11:1157–1164.

SIKS Dissertatiereeks

1998

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 Eduard Oskamp (RUL)
Computerondersteuning bij Straftoemeting

1999

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-6 Niek Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

2000

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-3 Carolien Metselaar (UvA)
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectie

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coupé (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

2001

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design

2001-11 Tom van Engers (VU)
Knowledge Management: The Role of Mental Models in Business Systems Design

2002

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel (KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (UvA)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UvA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UvA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politi ele gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

2005

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Reusable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

2007

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU)
Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs

2007-07 Natasa Jovanovic (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations

2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

Samenvatting

De klinische praktijk wordt gekenmerkt door complexe taken als diagnose, behandeling en prognose, waarbij de arts op ieder moment de juiste beslissing dient te nemen in onzekere situaties. Door de toenemende complexiteit van de hedendaagse geneeskunde en het streven naar doeltreffend en efficiënt medisch handelen, bestaat er behoefte aan systemen die de arts ondersteunen bij het nemen van beslissingen.

De afgelopen decennia zijn er steeds geavanceerdere technieken ontwikkeld die een basis kunnen vormen voor beslissingsondersteunende systemen. In dit proefschrift richten we ons op zogenaamde Bayesiaanse netwerken; grafische modellen die gebaseerd zijn op kansrekening en een mogelijkheid bieden om te redeneren met onzekere kennis. Het is bekend dat optimale modellen automatisch geleerd kunnen worden mits men over veel tijd en een grote hoeveelheid relevante data beschikt. De klinische praktijk wordt echter gekenmerkt door een beperkte hoeveelheid data. Dit impliceert dat het leren van optimale modellen vaak niet mogelijk is. Hier staat tegenover dat artsen beschikken over een grote hoeveelheid specialistische kennis die gebruikt kan worden om Bayesiaanse netwerken handmatig te construeren. In dit proefschrift worden verschillende technieken ontwikkeld die Bayesiaanse netwerken geschikt maken voor beslissingsondersteuning in de klinische praktijk. Met behulp van deze technieken kunnen modellen opgebouwd worden uit beschikbare medische kennis en bruikbare modellen geleerd worden uit een beperkte hoeveelheid data.

Na een beschouwing over de medische en wiskundige concepten die van belang zijn voor het onderzoek, beginnen we in hoofdstuk 3 met de beschrijving van medische beslissingsondersteuning in termen van abstracte probleemoplossing. Een duidelijke definitie van het medische probleem in combinatie met de specificatie van restricties op het te gebruiken model, geven al enig inzicht in de uiteindelijke structuur van het te bouwen Bayesiaanse netwerk. Vervolgens wordt een handleiding geboden voor het bouwen van Bayesiaanse netwerken op basis van beschikbare medische kennis welke onder andere gebaseerd is op eerder opgedane modelleerervaring.

In hoofdstuk 4 ontwikkelen we een concrete techniek die tot doel heeft om medische kennis te representeren in termen van speciale Bayesiaanse netwerk structuren. Het idee is dat causale (oorzaak-gevolg) relaties die gespecificeerd zijn op een kwalitatieve manier in combinatie met een aantal voor de hand liggende aannamen, leiden

tot een theorie die het toestaat om automatisch een model (of verzameling modellen) te identificeren die aan de kwalitatieve specificatie voldoet. Dit biedt onder andere de mogelijkheid om het ontwikkelen van Bayesiaanse netwerken op basis van expert kennis te vereenvoudigen.

Hoofdstuk 5 behandelt een ander probleem; namelijk het leren van een optimaal behandelingsmodel als we de beschikking hebben over een model van de onderliggende ziekte. Dit is een complex probleem aangezien behandeling het nemen van de juiste beslissingen op ieder moment in de tijd vereist. We beschrijven een formalisme waarin dit soort problemen gerepresenteerd kunnen worden en ontwikkelen een aantal technieken die het leren van (bij benadering) optimale behandelingsmodellen mogelijk maakt. De bruikbaarheid van de technieken wordt gedemonstreerd aan de hand van een model van hoog-gradige carcinoïde tumoren.

In hoofdstuk 6 beschrijven we de ontwikkeling van een model van laag-gradige carcinoïde tumoren waarin zowel de ziekte alsmede haar behandeling centraal staan. Met 218 variabelen en 74 342 kans-schattingen is dit zogenaamde dynamische Bayesiaanse netwerk een van de grootste in zijn soort. Het nut van dit soort modellen wordt gedemonstreerd aan de hand van een aantal casussen.

De hoofdstukken drie tot en met zes richten zich voornamelijk op behandeling en maken gebruik van aanwezige medische kennis. In hoofdstuk zeven richten we ons op diagnose en prognose, waarbij de modellen automatisch geleerd worden uit een beperkte hoeveelheid data. We demonstreren de prestaties van het maximum mutual information algoritme, decomposed tensor classifiers en noisy-threshold classifiers in de context van medische diagnose en prognose.

Hoofdstuk 8 geeft een algemene beschouwing over de ontwikkelde technieken. We concluderen dat de behandelde technieken hun nut hebben bewezen en een solide basis vormen voor beslissingsondersteuning in de klinische praktijk.

Curriculum Vitae

Marcel van Gerven was born in 's-Hertogenbosch at the fourth of September 1976 and graduated from the Jeroen Bosch College in 1995. He studied cognitive science at the Radboud University Nijmegen, obtaining a Bachelor's degree in cognitive science and a Master's degree in knowledge engineering. After his studies, Marcel worked at the Max Planck Institute for Psycholinguistics as a scientific programmer, spent some time abroad at the Insititute of Ophthalmology at University College London, and worked as a software engineer at LCN in Nijmegen, where he developed educational software.

In search of a position that combines his interests in cognitive science and machine learning, Marcel found a position as a junior researcher at the Institute for Computing and Information Sciences at the Radboud University Nijmegen. Here, he worked on the ProBayes project, which aimed to investigate whether the Bayesian network formalism offers a suitable framework for learning prognostic models from clinical datasets and for building computer-based medical decision support systems. During this project he has collaborated with physicians at the Netherlands Cancer Institute (NKI) and conducted research at the Universidad Nacional de Educación a Distancia (UNED) in Madrid. This research is presented in his dissertation "Bayesian Networks for Clinical Decision Support".

Currently, Marcel is working as a postdoctoral researcher at the Information and Knowledge Systems group at the Radboud University Nijmegen. He is working on brain-computer interfacing using MEG/EEG data in collaboration with researchers at the F. C. Donders Centre for Cognitive Neuroimaging. He enjoys spending his spare time on his many hobbies, which include sportsclimbing, photography, reading, programming, gaming, and travelling.