Notation*

List of symbols

x,$\underline{x}$,$\underline{\underline{x}}$	scalar, (column)vector or set, matrix
$X, x, p(X)$	random variable X, value x, probability mass function/density of X
$\mathrm{E}_{X,p(X)}[f(X)]$	expectation of $f(X)$ w.r.t. $p(X)$
$\mathrm{var}_{p(X)}[f(X)]$	variance of X w.r.t. $p(X)$
$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$	observational conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ w.r.t. p
$(X \perp\!\!\!\perp Y\|Z)_p$	$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$(X \not\perp\!\!\!\perp Y\|Z)_p)$	$\neg I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$CI_p(\underline{X};\underline{Y}\|\underline{Z})$	interventional conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ w.r.t. p
$\prec$	(partial) ordering
$\prec^c$	a complete reference ordering of the domain variables
$G, \underline{\theta}$	Bayesian network (BN) structure, BN parameters
$G^{\sim}$	essential graph of Directed Acyclic Graph (DAG) G
$\hat{G}_C^{\prec}(D)$	an optimal graph w.r.t. ordering $\prec$, data set D, and score/method C
$\mathcal{G}(n)/\mathcal{G}^k(n)$	set of DAGs over n nodes/with maximum k parents
$\mathcal{G}^{\prec}$	set of DAGs compatible with ordering $\prec$
$\sim$	compatibility relation
$\mathrm{pa}(X_i, G) \sim \prec$	$\mathrm{pa}(X_i, G)$ parental set is compatible with ordering $\prec$
$F, \mathcal{F}, f$	feature function, its range, a feature value
$\mathcal{F}^{\prec}$	set of values f compatible with $\prec$
$S_i(f, \prec)$	the set of valid parental sets of X_i in feature f given ordering $\prec$
$C_i(f, \prec, \mathrm{pa})$	a clause expressing $\mathrm{pa} \in S_i(f, \prec)$
$\mathrm{MB}_p(X_i)$	a Markov Blanket of X_i in p
S^{MLP}/S, $\underline{\omega}$	Multilayer perceptron (MLP) structure, MLP parameters
$\mathrm{pa}, \mathrm{pa}(X_i, G)$	set of parental variables, set of parents of X_i in G
pa_{ij}	the jth configuration of the values of the parents of X_i in an ordering
$\mathrm{bd}(X_i, G)$	set of parents, children and the children's other parents of X_i in G
$\mathrm{MBG}(X_i, G)$	the Markov Blanket/Mechanism Boundary Graph of X_i in G
$\mathrm{MB}(X_i, G)$	Markov Blanket of X_i defined by $\mathrm{bd}(X_i, G)$ in p compatible with G
$\mathrm{MBM}(X_i, X_j, G)$	the binary Markov Blanket membership
n	number of random variables
k	maximum number of parents in DAGs

*See also the remarks about style and notation in Section **??**

N	number of observed samples
$N_+/N_{\ldots,+,\ldots}$	the appropriate sum of $N_i/N_{\ldots,i,\ldots}$
D_N/D_N^L	real/literature data set with N complete observations
$D\|X$	data set D restricted to the set of variables X
$D^{\mathrm{IO}_1/\mathrm{IO}_2}$	clinical data sets
$D^{\mathrm{ME}^{HMR}_{\mathrm{O/R}}}, D^{\mathrm{PM}^{HMR}_{\mathrm{O/R}}}$	literature data sets based on a Medline (ME) and Pubmed (PM) corpus with $H/M/R$ filters binarized with Occurence/Relevance
D^*/D'	artifical data set by bootstrap or Monte Carlo methods
$\|\|$	cardinality
$1()$	indicator function
$S_i^{h/m/r/n}$	set of undirected edges with node i with high, medium, reasonable, and negligible pairwise relevance
$G^{H/M/R}$	the three prior DAG structures with high, medium, and reasonable relevance
$S_i^{H/M/R}$	the set of parents of node i in DAGs $G^{H/M/R}$
f', f''	first and second derivatives of function f
A^T	transpose of the matrix A
$\mathcal{A}()$	free-text annotation for an object
ξ^+/ξ^-	informative/noninformative background knowledge
KB	knowledge base (axioms)
$KB \models \alpha$	the entailment of sentence α w.r.t. knowledge base KB
$\mathcal{M}(KB)$	the set of models of a knowledge base KB
$\neg, \wedge, \vee, \not\equiv, \rightarrow$	operations of negation, and, or, exclusive or, implication
$\cap, \cup, \setminus, \Delta$	the intersection, union, difference, and symmetric difference
$KB \vdash_i \alpha$	the provability of sentence α by method $\vdash_i$ from axioms KB
Γ	the Gamma function
$\mathrm{Beta}(x\|\alpha,\beta)$	the probability density function (pdf) of the Beta distribution
$\mathrm{Dir}(x\|\underline{\alpha})$	the pdf of the Dirichlet distribution
$\mathrm{N}(x\|\mu,\sigma)$, $\mathrm{N}(x\|\underline{\mu},\underline{\underline{\Sigma}})$	the pdf of the normal distribution
BD,BD_e	Bayesian Dirichlet prior, observationally equivalent BD prior
BD_{CH}	a Bayesian Dirichlet (BD) prior with hyperparameters 1
$\mathrm{BD}_{\mathrm{eu}}$	a BD prior, where the hyperparameters are the converse of the number of parameters in the local dependency model
$L(\underline{\theta}; D_N)$	the likelihood function $p(D_N\|\underline{\theta})$
$H(X,Y)$, $I(X;Y)$	the entropy and the mutual information of X and Y
$\mathrm{KL}(X\|Y)$	the Kullback-Leibler divergence of X and Y
$H(X\|Y)$	the cross-entropy of X and Y
$L_1(,)$, $L_2(,)$	the Manhattan and the Euclidean distances the absolute and the quadratic losses
$L_0(,)$	the 0-1 1oss
$\mathcal{O}()/\Theta()$	asymptotic, proportional upper/upper and lower bound
$\mathrm{maxK}^{\mathrm{th}}(s)$	the Kth value in decreasing ordering in the set of scalars s

Acronyms

ABN	Annotated Bayesian Network
AUC	Area Under the ROC curve
BAN-BN/BAN	Bayesian Network Augmented Naive Bayesian Network
BMA	Bayesian Model Averaging
BN	Bayesian Network
BNC	Bayesian Network Classifier
DAG	Directed Acyclic Graph
FSS	Feature Subset Selection (problem)
FGS	Feature (sub)Graph Selection (problem)
HPD	High Probability Density (region)
IDO	IDO/99/03 project (K.U.Leuven) entitled "Predictive computer models for medical classification problems using patient data and expert knowledge"
IOTA	a study by the "International Ovarian Tumor Analysis" consortium
IR	Information Retrieval
LR	Logistic Regression
KE	Knowledge Engineering
KB	Knowledge Base
MAP	Maximum A Posteriori
MD	MEDLINE
MI	mutual information
ML	Maximum Likelihood
MLP	Multilayer perceptron
MBG	Markov Blanket/Mechanism Boundary Graph (a.k.a. classification or feature subgraph)
MB	Markov Blanket/Boundary set
MBM	Markov Blanket/Boundary Membership
(MC)MC	(Markov Chain) Monte Carlo
MPFs	Most Probable Features (problem)
Naive-BN/N-BN	Naive Bayesian network
OC	Ovarian Cancer
pABN-KB	Probabilistic Annotated Bayesian Network Knowledge Base
PM	PUBMED
ROC	Receiver Operating Characteristic (ROC) Curve
TAN-BN/TAN	Tree Augmented Naive Bayesian Network

List of Figures

List of Tables

Contents

Chapter 1

Bayesian networks primer

RELEVANT>

We summarize the Bayesian network model class, its probabilistic and causal interpretations and its Bayesian application. Then we overview the main issues of knowledge engineering, model evaluation and finally the learning of Bayesian networks.

The Bayesian framework overviewed in Chapter **??** leaves open the question of the model class, it is equally applicable with domain models discussed in this chapter or with conditional models discussed in Chapter **??**. In this chapter we investigate a domain model class called *Bayesian network*s, conditional models are discussed in Chapter **??**. <RELEVANT Bayesian networks form a subclass of graphical models that is using directed acyclic graphs (DAGs) instead of more general graphs to represent a probability distribution and optionally the causal structure of the domain. In an intuitive causal interpretation, the nodes represent the uncertain quantities, the edges denote direct causal influences, defining the model structure. A local probabilistic model is attached to each node to quantify the stochastic effect of its parents (causes). The descriptors of the local models give the model parameters.

The widespread popularity of this representation is probably the consequence of its applicability in multiple disciplines. FULLVERSION>These includes the research of causality investigating conditions for experimental and observational identification of causal effect, knowledge engineering (knowledge acquisition, formalization, verification, refinement and maintenance), probability theory (axiomatization of independencies of a distribution and decomposition, low-order approximation of a distribution), graph theory (graph representation of independencies of a distribution and decomposition of a distribution using graphs), and in (Bayesian) statistics/machine learning application (practical methods for prior incorporation and for performing Bayesian inference over observables and the model itself). <FULLVERSION The multifaceted nature of Bayesian networks follows from the fact that this representation addresses jointly three autonomous levels of the domain: the causal model, the probabilistic dependency-

independency structure, and the distribution over the uncertain quantities. FULLVERSION> These levels are generally connected with one-to-many relations, for example equivalence classes can be defined for distributions and causal models w.r.t. their conditional independence structure. Respectively, the questions of knowledge engineering (prior acquisition and formalization), inference and learning (i.e., Bayesian inference) is immediately tripled, additionally becomes highly entangled on these three levels. That is how can we represent, learn and perform inference with probabilistic causal relations. How can we represent, learn and perform inference with conditional independencies. How can we represent, learn and perform inference with complex, high-dimensional distributions. <FULLVERSION Additionally, the Bayesian network, as a complete probabilistic domain model, can be applied as an input-output model, for example as a classifier, so it can be investigated in the conditional framework as well (see Chapter **??** and **??**).

FULLVERSION>

The investigation of graphical models for probabilistic causal models goes back to 1920 in the work of Wright on path diagrams [149]. The first (medical) applications of a special class of Bayesian networks as a probabilistic expert system, including knowledge elicitation and learning appeared in 1970 [40], large scale applications were reported from the late 1980's. The axiomatic investigation of the structure of independencies in a probability distribution was reported in 1979 [38] and complemented with the issue of representability with DAGs in 1988 [114]. The decomposition of a probability distribution using annotated DAGs was reported in 1982 (for a general treatment of graph based decomposition see [98]). The causal interpretation of Bayesian networks and the related causal research is present from the proposal of the representation [114, 142, 134], though first seen as auxiliary human constructs and in the probabilistic research of causation the goals were to understand the limits of learnability from observational data and the identifyability of causal effect [114, 107, 115]. Later the role of the causal structure behind the independence structure and distribution became central and a model-based semantics for counterfactuals and the "probability of causation" has been formalized by using structural equations [58, 116]. An efficient inference method for a restricted class of Bayesian networks (for polytrees) has appeared in 1983 (for a detalied treatment see [114]) and a generally applicable inference method (the so-called join tree algorithm) in 1988 [130]. The Bayesian approach to the parameters using Dirichlet priors was reported in 1990 [132], a related evaluation methodology based on the prequential framework in 1993 [131] and the Bayesian approach to parameters was axiomatized in 1995 [74]. The Bayesian approach to the structure of the model was proposed in 1991 for models that compatible with a fixed causal ordering of the domain variables [17], the general treatment and practical learning was reported in 1992 [29]. A full-fledged Bayesian approach to perform Bayesian inference over structural properties was reported in 1995 [102] and a large-scale application in 2000 [55, 56]. A decomposed representation of Bayesian networks has appeared in 1989 [61], though first related to representing contextual inde-

pendencies. Later extensions related to knowledge engineering and attempts to first-order probabilistic logical extension were reported in [68, 81, 88, 89].

Our selection of the Bayesian network model class as a representation for the domain is mainly explained by our goal to analyze the compatibility between observations and heterogeneous, voluminous prior domain knowledge in a biomedical field, in which prior knowledge is mostly uncertain and includes the causal level, the associative level and the parametric level. The causal interpretation was frequently used in practice in the knowledge engineering phases and it was important in the development of knowledge discovery and information extraction methods. Consequently, we follow the recent trend (actually the revival of the original interpretation from 1920), which accepts the primacy, though not exclusivity of the causal interpretation.

<FULLVERSION

First we summarize the probabilistic interpretation of Bayesian networks, which is based on a DAG representation of an independence model of a distribution and on a decomposed representation of a distribution by DAGs annotated with local probabilistic models. Then we introduce the causal interpretation of Bayesian networks. Next we discuss the Bayesian approach to the parameters and to the structure. Then we discuss the knowledge acquisition methods and model (prior) evaluation methodologies. Finally we discuss fundamental results for model identification.

1.1 Representational issues

1.1.1 Three aspects: belief, relevance and causation

Suppose that our goal is to model uncertain events, furthermore we assume that the number of events and the corresponding outcomes (observables) are finite. According to the discussion in Chapter **??**, it corresponds to modeling a subjective joint distribution over the event space with elementary events defined by the Cartesian product of the possible outcomes. We denote the joint set of random events with $\underline{V}$, $p(\underline{V})$ denotes the joint (mass) probability distribution representing the personal belief over events. If it is necessary to differentiate, capitals with underline such as $\underline{X}$, $\underline{Y}$, $\underline{Z}$ denotes subsets and capitals such as X, Y, Z single random events, lowercase letters denotes values (outcomes) such as $X = x$. To simplify terminology we call each discrete random event a random variable (i.e., as if their outcomes would be always in $\mathcal{R}$). FULLVERSION> Note that because of the coherence argument in Section **??** a representation for $p(\underline{V})$ defines each of the respective marginals and conditionals corresponding to passive (i.e., non-interventionist) observations. <FULLVERSION

1.1.1.1 The model of observational independencies

We introduce now the notation for the independencies of random events.

Definition 1.1.1. *Let $p(\underline{V})$ be a joint distribution over $\underline{V}$ and $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ are disjoint subsets. Then denote the conditional independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ with $I_p(\underline{X}|\underline{Z}|\underline{Y})$, that is*

$$I_p(\underline{X}|\underline{Z}|\underline{Y}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\ p(\underline{y}|\underline{z}, \underline{x}) = p(\underline{y}|\underline{z}) \text{ whenever } p(\underline{z}, \underline{x}) > 0). \quad (1.1)$$

Note that conditional independence is required for all the relevant values of $\underline{Z}$. A weakened form of independence is the contextual independence, if conditional independence is valid only for a certain value $\underline{c}$ of another disjoint set $\underline{C}$. Then denote the contextual independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ and context c with $I_p(\underline{X}|\underline{Z}, \underline{c}|\underline{Y})$, that is

$$I_p(\underline{X}|\underline{Z}, \underline{c}|\underline{Y}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\ p(\underline{y}|\underline{z}, \underline{c}, \underline{x}) = p(\underline{y}|\underline{z}, \underline{c}) \text{ whenever } p(\underline{z}, \underline{c}, \underline{x}) > 0). \quad (1.2)$$

Another notation for $I_p(X|Z|Y)$ is $(X \perp\!\!\!\perp Y|Z)_p$. If it is nonambiguous, the subscript from $I_p(.)$ is omitted as well as the empty condition part. The negated independence proposition (i.e., dependency) is denoted with $(X \not\perp\!\!\!\perp Y|Z)_p$. It is a *direct dependency*, if for any disjoint $X, Y, Z \subseteq V$ $(X \not\perp\!\!\!\perp Y|Z)$ holds. A set of independence statements is called *independence model* (note that this is always a finite set in our case). We use the terms (probabilistic) independence and (information) irrelevance interchangeably.

FULLVERSION>

A standard measure for the strength of the dependence (association) between X and Y is the (conditional) *mutual information*

$$MI_p(X; Y|Z) = \mathrm{KL}(p(X, Y|Z)|p(X|Z)p(Y|Z)). \quad (1.3)$$

<FULLVERSION

Whereas the independencies or the complete independence model is an ideal candidate to represent qualitatively the target distribution, the autonomous, local mechanisms (rules) composing modularly the domain are the basis of both common sense and scientific understanding and explanation. FULLVERSION> The primary reason of it is their autonomy, which allows (1) the prediction of the effect of intervention (control) in the domain, (2) potential reuse of mechanisms with slight changes and (3) its use in other domains (which can be conceived as complex imaginary interventions). <FULLVERSION The autonomous relations are asymmetric w.r.t. time and interventions suggesting a causal interpretation.

1.1.1.2 The model of causal (in)dependencies

For the discussion of causality, we need a concept and notation for intervention.

Definition 1.1.2. *Let $do(x)$ denote the intervention of setting variable(s) X to value x and $p(Y|do(x))$ the corresponding interventional distribution [115].*

Note that despite the symmetry of the probabilistic dependence relation, the causal dependence relation is asymmetric. For example in a hypothetical world

with two variables X, Y and a single causal relation $X \rightarrow Y$ inducing $p(X, Y)$, the intervention on X and the observation of X are identical operations, but the intervention on Y will not influence the cause X (i.e., $p(Y|do(x)) = p(Y|x)$, but $p(X|do(y))$ is equal to $p(X)$ and not to $p(X|y)$). Now we introduce a notation for the *causal irrelevance* (independency) [116, 58]. FULLVERSION> Note that the definition does no meant to be an exhaustive definition of causation (e.g., the counterfactual aspects remains outside this definition), and it formalizes the concept of randomized clinical trials [148, 66]. <FULLVERSION

Definition 1.1.3. *Let $p(.|do(.))$ denote the appropriate interventional distributions over $\underline{V}$ and $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ are disjoint subsets. Then denote the causal independence of $\underline{X}$ and $\underline{Y}$ given $\underline{Z}$ with $CI_p(\underline{X}; \underline{Y}|\underline{Z})$, that is*

$$CI_p(\underline{X}; \underline{Y}|\underline{Z}) \text{ iff } (\forall \underline{x}, \underline{y}, \underline{z}\, p(\underline{y}|do(\underline{z}), do(\underline{x})) = p(\underline{y}|do(\underline{z}))) \tag{1.4}$$

The negated independence proposition (i.e., causal relevance or dependency) is denoted by $CD_p(X; Y|Z)$. If $\underline{Z} = \{\underline{V} \backslash X, Y\}\}$, then the causal relevancy/dependency relation is called *direct causal dependency* and denoted by $DCD_p(X; Y|Z)$ (or $(X \rightarrow Y|Z)_p$). A set of causal (in)dependence statements is called *causal model*.

FULLVERSION>

Measures for the strength of a causal relation are usually defined for binary X and Y and corresponds to standard measures in epidemiology for the strength of a (putatively) causal relation between a binary X (i.e., exposure) and Y (i.e., disease), such as the risk difference (or causal effect) (δ), the excess risk ratio or attributable risk (θ) and the *odds ratio* (Ψ) (see [148] p.133 and [116] p.292).

$$\delta = p(y|do(x)) - p(y|do(\neg x)), \tag{1.5}$$

$$\theta = \frac{p(y|do(x)) - p(y|do(\neg x))}{p(y|do(x))}, \tag{1.6}$$

$$\Psi = \frac{p(y|do(x))/p(\neg y|do(x))}{p(y|do(\neg x))/p(\neg y|do(\neg x)).} \tag{1.7}$$

In epidemiology these measures are usually defined using a non-interventional terminology, using "adjusted" estimates of observational probabilities ($\tilde{p}(y|x)$) instead of their interventionist counterparts $p(y|do(x))$. The operation of adjusting (or " controlling"), refers to the elimination of the effect of "confounders" $\underline{Z}$, which are common causes of X and Y, by evaluating the effect of change of X under the same values of the potential confounders, that is by conditioning and "holding" them fixed:

$$\tilde{p}(y|x) = \sum_z p(y|x, z)p(z). \tag{1.8}$$

For an epidemilogical overview of confounder selection and techniques see [148]. Beside this interventional definition of "X is a cause of Y" based on the $P(.|do())$ semantics other standard conditions for causation are the following (adapted from the list of "*principles of causality*" suggested within epidemiology [148]:

(1) strong dependency between X, Y (e.g., see Def. 1.3), (2) X precedes temporally Y, (3) plausible explanation of the mechanism between X, Y without alternative explanations based on confounding, (4) necessity (i.e., if the cause is removed, effect is decreased), (5) sufficiency (if exposure to cause is increased, effect is increased).

The probabilistic definition of causation in Def. 1.4 formalizes many, but for example not the counterfactual aspects. The last two conditions can be reformalized using counterfactuals as (4') y would not have been occurred with that much probability if x had not been present and (5') y would have been occurred with larger probability if x had been present. Furthermore, in biomedical domains an equally important condition for the establishment of a causal relation is the existence of a scientific explanation for the relation between X and Y, usually based on a hypothesized autonomous, local rule or mechanism, that is the concept of causation and intervention is deeply connected with the scientific understanding of "stable and transportable" mechanism as indicated above, what makes the causal interpretation highly relevant for knowledge discovery from scientific publications and for prior incorporation in Chapters **??**, **??**. For the general treatment of causation in scientific explanations and in philosophy of science see [123, 147].

<FULLVERSION

FULLVERSION>

1.1.1.3 Summary

Now the multiple goals of the Bayesian network representation can be more specifically circumscribed:

- P representation for the joint distribution (e.g., to support knowledge acquisition, learning and inference),
- M sound and complete representation for the independency model,
- P-M understanding relation between P and M (i.e., the use of a representation of independence model for a compact representation of the joint),
- C sound and complete representation for the causal model with a causal interpretation,
- M-C understanding the relation between M and C (i.e., the relation between the observationally defined, symmetric (in)dependence relations and the interventionally defined asymmetric causal relation),
- P-C understanding the relation between P and C (i.e., the conversion of causally defined quantities $P(y|do(x), z)$ into "do()"-free observational quantities $P(y|w)$ or to more appropriate causal quantities $P(y|do(x'), z')$),
- C' definition of counterfactuals with a (logical) model-based probabilistic semantics and respectively a probabilistic account of (actual/individualistic)

causation using structural equations (e.g., the probability of y would not have occurred if x had not been present conditioned on that x was present and y occurred).

Results necessary for the development of the thesis are related to the issues of $P, M, C, P-M, M-C$ in case of full observation (i.e., no hidden variables). For the general treatment of these issues and for references, see [114, 66, 134].

<FULLVERSION

1.1.2 Probabilistic Bayesian networks

Before investigating the role of directed acyclic graphs (DAGs) in representing causal relations, we have to clarify their purely probabilistic role in representing a joint distribution numerically and its (in)dependence model.

1.1.2.1 Markov conditions

Assume that each vertice (node) in DAG G corresponds to a random variable. We need the following concepts (cited from [114, 98, 33, 116]).

Definition 1.1.4. *A distribution $p(X_1, \ldots, X_n)$ is Markov relative to DAG G or factorizes w.r.t G, if*

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | \operatorname{Pa}(X_i)), \tag{1.9}$$

where $\operatorname{Pa}(X_i)$ denotes the parents of X_i in G.

Definition 1.1.5. *A distribution $p(X_1, \ldots, X_n)$ obeys the ordered Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_{\prec(i)} \perp\!\!\!\perp \{\{X_{\prec(1)}, \ldots\, X_{\prec(i-1)}\} \setminus \operatorname{Pa}(X_{\prec(i)})\} | \operatorname{Pa}(X_{\prec(i)}))_p, \tag{1.10}$$

where $\prec$ is some ancestral ordering w.r.t. G (i.e., compatible with arrows in G) and $\{X_{\prec(1)}, \ldots\, X_{\prec(i-1)}\} \setminus \operatorname{Pa}(X_{\prec(i)})$ denotes all the predecessors of $X_{\prec(i)}$ except its parents.

Definition 1.1.6. *A distribution $p(X_1, \ldots, X_n)$ obeys the local (or parental) Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \operatorname{Nondescendants}(X_i) | \operatorname{Pa}(X_i))_p, \tag{1.11}$$

where $\operatorname{Nondescendants}(X_i)$ denotes the nondescendants of X_i in G (i.e., without directed path from X_i).

Definition 1.1.7. *A distribution $p(X_1, \ldots, X_n)$ obeys the global Markov condition w.r.t. DAG G, if*

$$\forall\, X, Y, Z \subseteq V :\ (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_p, \tag{1.12}$$

where $(X \perp\!\!\!\perp Y | Z)_G$ denotes that X and Y are d-separated by Z, that is if every path p between a node in X and a node in Y is blocked by Z as follows

1. *Either path p contains a node n in Z with non-converging arrows (i.e.,* $\rightarrow n \rightarrow$ *or* $\leftarrow n \rightarrow$*),*

2. *Or path p contains a node n not in Z with converging arrows (i.e.,* $\rightarrow n \leftarrow$*) and none of the descendants of n is in Z.*

FULLVERSION> For an equivalent definition of a global $(X \perp\!\!\!\perp Y|Z)_G$ based on "m-separation" in the moralized graph of G, see [98]. <FULLVERSION

Now we can state a central result connecting the DAG representation of the joint distribution and the DAG representation of the independence model [98].

Theorem 1.1.1 ([98])**.** *Let $p(V)$ be a probability distribution and G a DAG, then the conditions in Def. 1.1.4, 1.1.5, 1.1.6, and 1.1.7 are equivalent:*

(F) p is Markov relative G or p factorizes w.r.t G,

(O) p obeys the ordered Markov condition w.r.t. G,

(L) p obeys the local Markov condition w.r.t. G,

(G) p obeys the global Markov condition w.r.t. G.

Because of their equivalence, we can refer to these as the (directed) Markov condition for the pair (p, G). To show the necessity and sufficiency of these conditions, we refer to a result that a sound and complete, computationally efficient algorithm exists to read off exactly (!) the independencies that are valid in all distributions that are Markov relative to a given DAG G [114].

Theorem 1.1.2 ([114])**.**

$$\forall\, X, Y, Z \subseteq V:\ (X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_p \text{ in all } p \text{ Markov relative to } G.$$

Two further properties are implied by any of the (FOLG) conditions: the pairwise Markov condition [98] and the boundary Markov conditions [114].

Definition 1.1.8. *A distribution $p(X_1, \ldots, X_n)$ obeys the pairwise Markov condition w.r.t. DAG G, if for any pair of variables X_i, X_j nonadjacent in G and $X_j \in \text{Nondescendants}(X_i)$, $(X_i \perp\!\!\!\perp X_j|\text{Nondescendants}(X_i) \setminus \{X_j\})_p$ holds [98].*

To state the other implication, we need the following concepts.

Definition 1.1.9. *A set of variables $MB_p(X_i)$ is called a Markov blanket of X_i w.r.t. the distribution $p(X_1, \ldots, X_n)$, if $(X_i \perp\!\!\!\perp V \setminus \text{MB}(X_i)|\,\text{MB}(X_i))_p$ (see Fig. 1.1). A minimal Markov blanket is called Markov boundary [114].*

Definition 1.1.10. *A distribution $p(X_1, \ldots, X_n)$ obeys the boundary Markov condition w.r.t. DAG G, if the boundary* $\text{bd}(X_i, G)$ *is a Markov blanket of X_i, where* $\text{bd}(X_i, G)$ *denotes the set of parents, children and the children's other parents for X_i (i.e., parents with common child with X_i, see Fig.* **??** *and Fig. 1.1):*

$$\text{bd}(X_i, G) = \{\text{Pa}(X_i, G) \cup \text{Ch}(X_i, G) \cup \text{Pa}(\text{Ch}(X_i, G), G)\}. \tag{1.13}$$

The boundary $\mathrm{bd}(X_i, G)$ *coincides with the standard graph-theoretic notion of boundary (i.e., set of neighbours) of* X_i *in the moral graph of* G*, which is the graph where edges are added between parents with a common child and the orientation is dropped [33].*

Theorem 1.1.3 ([114])**.** *The (FOLG) Markov condition for* (p, G) *implies that the set* $\mathrm{bd}(X_i, G)$ *is a Markov blanket (*$\mathrm{MB}_p(X_i)$*) for* X_i*.*

Note that the set $\mathrm{bd}(X_i, G)$ is not necessarily Markov boundary as it may not be minimal (because of the non-optimality of G to p). In the Bayesian context this problem is negligible as Th. 2.1.2 and the discussion in Section 1.1.2.3 show, so we will also refer to $\mathrm{bd}(X_i, G)$ as the Markov blanket for X_i in G using the notation $\mathrm{MB}(X_i, G)$ by the implicit assumption that p is Markov compatible with G and stable. The induced (symmetric) pairwise relation $\mathrm{MBM}(X_i, X_j, G)$ w.r.t. G between X_i and X_j

$$\mathrm{MBM}(X_i, X_j, G) \Leftrightarrow X_j \in \mathrm{bd}(X_i, G) \tag{1.14}$$

is called *Markov blanket membership* [54]. In short, the set $\{\mathrm{MBM}(X_i, G)\}$ includes the variables with non-blockable pairwise (observational) dependencies 1.1 to X_i including the unconditionally related variables (parents and children) and the purely conditionally related ones (the rest).

Finally, we introduce here the definition of the Markov Blanket (sub)Graph (MBG) (for a discussion of the MBG feature, see Section 2.2).

Definition 1.1.11. *A subgraph of* G *is called the Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph* $\mathrm{MBG}(X_i, G)$ *of variable* X_i *if it includes the nodes in the Markov blanket defined by* $\mathrm{bd}(X_i, G)$ *and the incoming edges into* X_i *and into its children* $\mathrm{Ch}(X_i, G)$ *(see Fig.* **??** *and Fig. 1.1).*

Fig. 1.1 shows an example for a Markov Blanket set and the Markov Blanket graph in a Markov chain.

Figure 1.1: A Bayesian network structure G defining a Markov chain $p(X_1, X_2, Y, X_4, X_5)$. Underscore denotes the members of a Markov Blanket set of variable Y $\mathrm{MB}_p(Y)$, which is the unique Markov Boundary $\mathrm{MB}(Y, G)$ as well (defined by the boundary $\mathrm{bd}(X_i, G)$). Solid lines denote the edges of the Markov Blanket Graph $\mathrm{MBG}(Y, G)$.

1.1.2.2 Definitions of Bayesian networks

The equivalence of the conditions *FOLG* in Th. 1.1.1 allows versatile definitions of Bayesian networks. FULLVERSION> Each condition has its own appeal to base the definition on it and leaves the others as derived theorems. The condition (O) is useful for constructing a DAG to factorize P. The condition (L)

is attractive in the causal interpretation as fixing the direct causes renders the nondescendant other variables independent. The condition (G) is important because it emphasize that a DAG G offers a logical representation for the qualitative description of the joint distribution. <FULLVERSION A neutral definition is as follows.

Definition 1.1.12. *A directed acyclic graph (DAG) G is a Bayesian network of distribution $p(V)$, if the variables are represented with nodes in G and (G, p) satisfies any of the conditions F, O, L, G such that G is minimal (i.e., no edge(s) can be omitted without violating a condition F, O, L, G).*

If the distribution P is strictly positive, then the Bayesian network compatible with a given ordering $\prec$ is unique (i.e., composed of the unique minimal parental sets that makes the variable independent of the variables before w.r.t $\prec$) [114]. Note that depending on the ordering different Bayesian networks can be gained, representing more or fewer independencies of P.

In engineering practice Bayesian networks are frequently informally defined as a DAG annotated with local probabilistic models for each node.

Definition 1.1.13. *A Bayesian network model M of a domain with variables V consists of a structure G and parameters $\underline{\theta}$. The structure G is a directed acyclic graph (DAG) such that each node represents a variable and local probabilistic models $p(X_i|\operatorname{pa}(X_i))$ are attached to each node w.r.t. the structure G, that is they describe the stochastic dependency of variable X_i on its parents $\operatorname{pa}(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\underline{\theta}_i$, and $\underline{\theta}$ denotes all the parameters of the model.*

When the conditionals are combined together as in Eq. 1.9, they define an overall joint distribution p. It trivially satisfies Markov relativity to G and the structure satisfies the conditions O, L, G. The lack of minimality requirement causes only potential redundancy (parameters) and fewer implied independencies. In most cases, we use the term Bayesian network to refer to both structure and parameters.

1.1.2.3 Stability

A limitation of DAGs to represent a given (in)dependency model is that (1) probabilistic dependencies are not necessarily transitive and (2) lower order (e.g., pairwise) probabilistic independencies does not imply higher order (e.g., multivariate) independencies. RELEVANT> These are illustrated with the following examples.

Example 1.1.1. *Consider $p(X, Y, Z)$ with binary X, Z and ternary Y in a Markov chain ($X \to Y \to Z$). The intransitivity condition — $(X \not\perp\!\!\!\perp Y)$, $(Y \not\perp\!\!\!\perp Z)$, and $(X \perp\!\!\!\perp Z)$ — can be rewritten as an equation system with the probabilities. Its solvability demonstrates that the "naturally" expected transitivity of dependency can be destroyed by properly selected values. For the other case, consider $p(X, Y, Z)$ with binary variables, where $p(x) = p(y) = 0.5$ and $p(Z|X, Y)$*

is defined by the logical function $Z = XOR(X, Y)$*. In this case* $(X \perp\!\!\!\perp Z)$ *and* $(Y \perp\!\!\!\perp Z)$*, but* $(\{X, Y\} \not\perp\!\!\!\perp Z)$*, which demonstrates that pairwise independence does not imply total independence.*

<RELEVANT

However, such numerically encoded independencies correspond to solutions of systems of equations describing these constraints, which are not stable for numerical perturbations. This leads to the following definition.

Definition 1.1.14. *The distribution* p *is stable** *(or faithfull), if there exists a DAG called perfect map exactly representing its (in)dependencies (i.e.,* $(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_p$*,* $\forall\ X, Y, Z \subseteq V$ *). The distribution* p *is stable w.r.t. a DAG* G*, if* G *exactly represents its (in)dependencies.*

FULLVERSION> Note that because there are multiple Bayesian network models for P not necessarily representing the same independencies, P can be stable (i.e., there is a perfect map for it), whereas unstable w.r.t. other DAGs not representing all the independencies. The corresponding factorized representations in such cases are sensitive to perturbations. It is an open question whether the independence models with perfect DAG representation can be characterized. However not only DAGs has limited capacity to represent (in)dependency models, a result shows that in general there is no finite characterization of the independence models corresponding to probability distributions (for special cases, discussion and references see [142, 143, 19]). Because certain distributions cannot be faithfully represented by any DAG, they pose a problem for knowledge representation and learning. <FULLVERSION

Whereas in many domains the possibility of an unstable distributions is a real cause for concern, particularly containing deterministic relations, the following result shows that it is reasonable to expect that in a natural, "noisy" domain almost all the distributions are stable in a strict sense, which is also relevant for the applied Bayesian framework. If a "smooth" distribution is defined over the distributions Markov relative to G (such as in Section 1.1.5.1 in the Bayesian framework), it can be shown that the measure of unstable distributions is zero (as being a solution of a system of equations) [107]. It allows to sharpen Th. 1.1.2 that the DAG-based relation $(X \perp\!\!\!\perp Y|Z)_G$ offers a computationally efficient algorithm to read off exactly the independencies that are valid in a distribution Markov relative to G in case of "almost all" such distributions.

1.1.2.4 Equivalence classes of Bayesian networks

The assumption of stability and strict positivity does not exclude the possibility of having multiple perfect maps encoding the same independencies in p. RELEVANT>

Example 1.1.2. *Consider a Markov chain* $\mathcal{X} = \{X_1, \ldots, X_n\}$ *with a stable distribution. Its independence model includes* i=1, …, n:$(X_i \perp\!\!\!\perp \{X_1, \ldots X_{i-2}\}|X_{i-1})$,

*For a different interpretation of this term in probability theory, see [122].

and also the implied $(X_i \perp\!\!\!\perp \{X_1, \dots X_{i-2}, X_{i+2}, \dots, X_n\} | \{X_{i-1}, X_{i+1}\})$. *This independence model can be exactly represented by n equivalent linear Bayesian networks without introducing convergent arrows, including the two special cases of the "forward" and the "backward" network (see Fig. 1.2).*

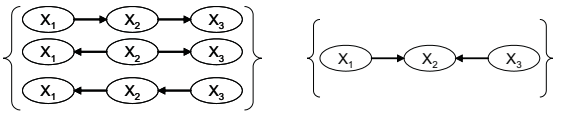

Figure 1.2: The equivalence classes of Bayesian network structures over three variables with direct dependences between X_1, X_2 and X_2, X_3, but not between X_1, X_3.

<RELEVANT

The induced independence models allow the definition of an equivalence relation between DAGs [114, 143, 107].

Definition 1.1.15. *Two DAGs* G_1, G_2 *are observationally equivalent, if they imply the same set of independence relations (i.e.,* $(X \perp\!\!\!\perp Y|Z)_{G_1} \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$*).*

RELEVANT>The implied equivalence classes may contain $n!$ number of DAGs (e.g., all the full networks representing no independencies) or just 1 (e.g., the empty DAG representing total independence of the variables)FULLVERSION>, however experimental results indicate that in practical domains on average the number of DAGs in an equivalent class are around 3 [85]<FULLVERSION. The characterization of the DAGs within the same equivalence class relies on two observations. First, the undirected skeleton of the observationally equivalent DAGs are the same, because an edge in a DAG denotes a direct dependency, which has to appear in any Markov compatible DAG [114]. Second, the direct dependencies between X, Y and Y, Z without direct dependence between X, Z and without independence such that $(X \perp\!\!\!\perp Z|\{Y, S\})$ has to be expressed with a unique converging orientation $X \rightarrow Y \leftarrow Z$ creating a so-called *v-structure* according to the global semantics. <RELEVANTThe theorem characterizing the DAGs within the same observational (and distributional) equivalence class is as follows.

Theorem 1.1.4 ([114, 24])**.** *Two DAGs* G_1, G_2 *are observationally equivalent, iff they have the same skeleton (i.e., the same edges without directions) and the same set of v-structures (i.e., two converging arrows without an arrow between their tails) [114]. If in the Bayesian networks* $(G_1, \underline{\theta}_1)$ *and* $(G_2, \underline{\theta}_2)$ *the variables are discrete and the local conditional probabilistic models are multinomial distributions, then the observational equivalence of* G_1, G_2 *implies equal dimensionality and bijective relation between the parameterizations* $\underline{\theta}_1$ *and* $\underline{\theta}_2$ *called distributional equivalence [24].*

The limitation of DAGs to represent uniquely a given (in)dependency model poses a problem for the interpretation of the direction of the edges. It also

poses the question of representing the identically oriented edges in observationally equivalent DAGs. As the definition of the observational equivalence class suggests the common v-structures identify the starting common edges and further identical orientations are the consequences of the constraint that no new v-structures can be created. This leads to the following definition (for an efficient, sound, and complete algorithm, see [107]).

Definition 1.1.16. *The essential graph representing DAGs in a given observational equivalence class is a partially oriented DAG (PDAG) that represents the edges that are identically oriented among all DAGs from the equivalence class (called compelled edges) in such a way that exactly the compelled edges are directed in the common skeleton, the others are undirected representing inconclusiveness.*

FULLVERSION> Note that the definition satisfies the Ockham principle in that the essential graph of a stable distribution encompasses only the observationally equivalent DAGs, that is only the "minimal" models (inducing minimal set of dependencies or in case of multinomial local models only models having minimum dimensionality). Another feature is that the edges in an essential graph of a stable distribution (irrespective of their status of orientation) exactly represent the direct dependencies. <FULLVERSION

1.1.3 Causal Bayesian networks

Now we continue with the causal interpretation of Bayesian networks, because of its relevance for prior acquisition and incorporation (i.e., knowledge acquisition from experts, for the discovery from scientific publications and for prior incorporation in Chapters **??**, **??**).

1.1.3.1 On the possibility of causal interpretation

The classical problem of "from (observational) correlation to causation", that is the question of determining causal status of a passively observed dependency between X and Y can be decomposed using the concepts introduced earlier to the question about the DAG-based representation of independencies (i.e., probabilistic Bayesian network), the existence of exact representation (i.e., stability) and the existence of unambiguous representation (i.e., essential graph). First, we have to consider whether all direct dependencies among the constructed domain variables are causal. This assumption is highly questionable and is discussed in detail below. Second, we have to consider stability that would guarantee that a corresponding Bayesian network exactly represents the independencies. Third, we have to adopt the "Boolean" Ockham principle, namely that only the minimal, consistent models are relevant (see Section 2.4, for the "soft" Ockham principle in the Bayesian approach to causal discovery). The essential graph resulting from the joint analysis of the observational conditional independencies (i.e., "correlations) indicates causal relations under these conditions. In short, under the condition of stability the essential graph represents the direct causal

dependencies and the orientations that are dictated by (in)dependencies in the domain through the minimal models (DAGs) compatible with them. Furthermore, the direction of the edges corresponds to the intuitive expectation as the intransitive dependency triplets are represented as v-structures.

FULLVERSION> For example in an (unconfounded) v-structure $X \to Y \leftarrow Z$ with direct dependencies between X, Y and Y, Z and with the only independence $(X \perp\!\!\!\perp Z)$ the direction of the arrows are compatible with the expectation that X and Z being independent events and both of them are dependent with Y, then X and Z are independent causes preceding temporarily Y (for a discussion of the validity of this principle on macroscopic level and the statistical approach to time, see [116]). <FULLVERSION

Correspondingly we can define a causal model as a Bayesian network according to Definition 1.1.13 with the causal interpretation that edges denote direct influences.

Definition 1.1.17. *A DAG is called a causal structure over a set of variables V, if each node represents a variable and edges direct influences. A causal model is a causal structure extended with local probabilistic models $p(X_i|\operatorname{pa}(X_i))$ for each node w.r.t. the structure G describing the causal stochastic dependency of variable X_i on its parents $\operatorname{pa}(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\underline{\theta}_i$, and $\underline{\theta}$ denotes all the parameters, so a causal model consists of a structure G and parameters $\underline{\theta}$.*

With further assumption of stability, the essential graph shows exactly the independency relations and exhaustively the identifiable causal relations, which suggests that whereas the question of causation is underconstrained for a pair of variables (restricted to "no dependency-no causation), the joint analysis of the system of independencies allows partial identification.

1.1.3.2 The Causal Markov Condition

The following condition ensures the validity and sufficiency of a causal structure.

Definition 1.1.18. *A causal structure G and distribution p satisfies the Causal Markov Condition, if p obeys the local Markov condition w.r.t. G.*

The Causal Markov condition relies on Reichenbach's "common cause principle" that dependency between events X and Y occurs either because X causes Y, or Y causes X or there is a common cause of X and Y (it is possibly an aggregate of multiple events) [116, 66]. Consequently, the precondition of the Causal Markov condition for (p, G) is that the set of variables V is *causally sufficient* for P, that is all the common causes for the pairs $X, Y \in V$ are inside V. Note that hidden variables are allowed fitting to the usually high level of abstraction of the model, only variables that influence two or more variables in V are necessary for causal sufficiency. Interestingly, in the presence of potential hidden common causes (*confounders*), that is if the Causal Markov Condition is violated, certain causal dependencies can still be identified [116].

FULLVERSION>

Example 1.1.3. *The Causal Markov Condition (i.e., the assumption of no hidden common causes) guarantees that from the observation of no more than three variables we can infer causal relation as follows. The direct dependencies between X, Y and Y, Z without direct dependence between X, Z and without conditional independence such that $(X \perp\!\!\!\perp Z|\{Y, S\})$ (i.e., with conditional dependence) should be expressed with a unique converging orientation $X \rightarrow Y \leftarrow Z$ according to the global semantics (i.e., DAG-based relation $(X \perp\!\!\!\perp Y|Z)_G$ from Def. 1.1.7) resulting in a v-structure. If potential confounders are not excluded a priori, we have to observe at least four variables to possibly exclude that direct dependency is caused by a confounder. Continuing the example, assume furthermore that we observe a forth variable W with the direct dependence Y, W and conditional independence $(W \perp\!\!\!\perp \{X, Z\}|Y)$ (because of stability W depends on X and Z). As Y induces independence the global semantics dictates an $Y \rightarrow W$ (note the earlier v-structure) and it cannot be mediated by a confounder $* \; Y \rightarrow * \rightarrow W$ (Y as an effect would not block).*

<FULLVERSION

The causal Markov condition links the causal relations to dependencies and states sufficiency to model the observed probabilistic dependencies. On the other hand, the condition of stability of P w.r.t. a causal structure G states the necessity of G.

These two assumptions guarantee that observational (in)dependence (1.1) is exactly represented by the DAG-based relation (Def. 1.1.7) in a Markov compatible graph G and that causal (in)dependence (Def. 1.4) is exactly represented by standard separation in the causal structure G [58]. Furthermore, the Causal Markov condition allows the computation of interventional distributions corresponding to the $do()$ operation (1.1.2) according to the "*Manipulation theorem*" ([134]) or "graph surgery" ([116]). It is performed simply by deleting the incoming edges for the intervened variables in the $do()$ operator and omitting these factors from the factorization in Eq. 1.9 FULLVERSION> resulting in the truncated factorization as explained below <FULLVERSION.

1.1.3.3 The interventionist and mechanistic views

In general, the causal structure G satisfying the Causal Markov Condition for a domain with (observational) distribution P can encode all the interventional distributions in a single causal model, which is formalized in the interventionist definition of "causal Bayesian networks" [116].

FULLVERSION>

Definition 1.1.19. *Let $p(\underline{V}|do(\underline{x}))$ denote an interventional distribution corresponding to setting variable(s) $\underline{X} \subseteq \underline{V}$ to value $\underline{x}$ and P_* the set of all interventional distributions (including $p(\underline{V}|do(0))$ the observational target distribution without intervention). A DAG G is said to be a causal Bayesian network compatible with P_* iff for each $p(\underline{V}|do(\underline{x})) \in P_*$ the following three conditions hold*

1. $p(\underline{V}|do(\underline{x})$ *is Markov relative to* G,

2. $\forall\ X_i \in \underline{X}\ p(x_i|do(\underline{x}) = 1$ *if value* x_i *and* $\underline{x}$ *is compatible,*

3. $\forall\ X_i \notin \underline{X}\ p(x_i|pa_i, do(\underline{x}) = p(x_i|pa_i)$ *if value(s)* pa_i *and* $\underline{x}$ *is compatible.*

<FULLVERSION

This definition of causal Bayesian networks explicitly shows that the concept of causation is based on the concept of intervention, more exactly on the systematic ability to intervene. This boils down to the assumption of autonomous, local "mechanisms" composing the domain, which can be triggered by interventions independently and can be understood independently. A formalization of this "mechanism-based interpretation" of DAG representations is offered by the so-called "*functional Bayesian networks*" using a formalism of mechanisms as deterministic functions with disturbances (cf. with structural equation) [44, 116]. Whereas the functional Bayesian network formalism allows the probabilistic modeling of counterfactuals, in the thesis we adopt a more modest causal interpretation termed "mechanism-based interpretation" meaning that under the Causal Markov condition the local probabilistic dependency models correspond to the autonomous, local mechanisms in the causal model.

FULLVERSION>

1.1.3.4 The ubiquity of mechanism-based interpretation

Because of this intermingled nature of observational and interventionist interpretations, the causal interpretation of Bayesian networks, particularly the corresponding mechanism-based interpretation of Bayesian networks is equally relevant in an observational data analysis task, whenever the incorporation of prior knowledge or the evaluation of the probabilistic Bayesian network model is relevant. This is exemplified by a wide range of studies and concepts, such as the following: the estimation of parameters in local, conditional models is preferred and better solvable in "causal" directions [83], the intuitive causal interpretation of the DAG structure in knowledge elicitation (w.r.t the independency-based interpretation), the special "causal" conditional models for the "independent" combination of the effect of causes such as the noisy-OR and logistic regression models [114, 78, 72], the assumption of decomposability of parameter priors for a Bayesian network w.r.t. mechanisms (for local and global parameter independence see Section 1.1.5.1) [132], the assumption of decomposability of structure prior for a Bayesian network (for structure independence see Section 1.1.5.2) [17, 29, 55], special conditional models for modeling the ensemble of alternatives of (sub) conditional models

(inducing contextual (in)dependencies) [16, 114]. These show that the causal (e.g., mechanism-based interpretation) is especially relevant in the Bayesian framework.

<FULLVERSION

1.1.3.5 Pairwise causal relations

The causal interpretation of Bayesian networks allows the definition of the following logical pairwise relations in a causal structure (recall that in stable causal models the dependency relations always represent exactly the probabilistic dependency relations):

1. *Causal path (P, $CaP(X_i, X_j|G)$):* There is a directed path from node X_i to node X_j in DAG G (also denoted by $X_i \prec_G X_j$).
2. *Causal edge (E, $CaE(X_i, X_j|G)$):* There is an edge from node X_i to node X_j in DAG G (also denoted by $(X_i \rightarrow_G X_j)$).
3. *Compelled edge (CompE, $CompE(X_i, X_j|G)$):* There is a compelled edge from node X_i to node X_j in the essential graph for DAG G.
4. *(Pure) Confounded (Conf, $Conf(X_i, X_j|G)$):* The two nodes X_i and X_j have a common ancestor in DAG G. The confounded relation is said to be pure, if there is no edge or path between the nodes.
5. *Independent (I, $Ind(X_i, X_j|G)$):* None of the previous.

Note that these pairwise relations can be also used in an acausal context using the differences w.r.t. the independence relation.

1.1.4 On the relativity of the interpretations

The causal interpretation has been challenged from many points of view. The Causal Markov assumption can be questioned as the presence of unobserved (hidden) variables as potential confounders seriously constrains the causal interpretation and automated causal discovery (for the Bayesian analysis of potentially infinite number of confounders, see [66]). Another violation called *selection bias* can occur if the observations depend on the joint combination of otherwise independent events, which induces non-causal dependencies between them. The next difficulty is related to the mixture of causal models, if conditionally both X causes Y and vice versa. A similar problem is the presence of feedback and indirectly temporality. Finally, the causal nature of the relations can be questioned because of global physical and semantic constraints between the variables [146]. It can occur if there is a global constraint on the joint set of the variables, outside the scope of the modeled domain or if the definitions of the variables are overlapping (i.e., there are logical dependencies).

In both the causal and probabilistic interpretations, the assumption of stability can be also questioned, for example because of deterministic dependencies, resulting in the lack of guarantee for the uniqueness and exactness of the representation.

Finally, obviously the (in)dependencies are relative to the set of variables and specifically, also to the values of the variables (consider the conversion of a nth order Markov chain into a first-order by augmenting the state space), so

both the probabilistic and causal interpretation has to be conditional on the set of variables and values [66].

These considerations are free of any data size issue and they are free of the question of the subjectivity of the prior in the Bayesian analysis of causation. The data set and the subjective prior information are further essential factors in the relativity of the causal and probabilistic inferences.

1.1.5 Bayesian networks in the Bayesian framework

FULLVERSION> The high-dimensional, small sample data in a prior rich domain motivates a full Bayesian approach to the learning of Bayesian networks. In fact, the uncertainty over the causal diagram itself, that is the uncertainty over the causal theory itself, is of primary importance, as all the results relying on it are and should be conditional or averaged as noted in [39]. It needs the Bayesian generalization of the results related to the causal Bayesian networks, such as the identification of genuine causes (for the Bayesian view of compelled edges, see [76, 54]) or the identifiability and computation of the strength of a causal effect from observational data (for the Bayesian inference of causal effect, see [116], p.xx) or the design of an optimal interventionist data collection [150].

An important factor for the popularity of Bayesian networks is the possibility to incorporate many kinds of prior knowledge into learning—ranging from logical constraints on the model structure [135, 29, 94, 21] or qualitative monotonicity relations between the variables [137, 69] to prior distributions for network structures and parameterizations of local dependencies [131, 17, 29, 74, 18]. <FULLVERSION

In the Bayesian framework the prior probabilities over the Bayesian network model is represented by a joint distribution $p(G, \underline{\theta}))$ over the DAG structures G and corresponding parameters $\underline{\theta}$. Because of the generality of the Bayesian network representation this distribution itself can be represented by a Bayesian network as we shall see below. However the specification of the joint or the conditionals $p(G)$ and $p(\underline{\theta}|G)$ requires practical simplifications and careful theoretical considerations, because of the huge size of the space and because of the observational equivalence of the structures. As in the thesis in general, in this section we also assume that the variables $V = \{X_1, \ldots, X_n\}$ are discrete with r_i number of values. We start with the parameter prior and then discuss the structure prior.

1.1.5.1 Parameter priors for Bayesian network models

The specification of parameter prior $p(\underline{\theta}|G)$ for Bayesian networks poses the following questions: the parametric form of the prior, the relation of the decomposition of the prior to the decomposition of P, the consistent confidence of the decomposed priors for the parts of a single structure, the consistency of the priors for observationally equivalent structures (recall that observational equivalence implies distributional equivalence in the discrete, multinomial case, see Th. 1.1.4). There is a remarkable result to clarify these problems. First, if the

parameter prior decomposes w.r.t. the structure and the parameter priors are equivalent for observationally equivalent structures, then the parameter prior is a Dirichlet distribution. Furthermore, if the parts of the decomposed parameter prior are invariant w.r.t. the structure, then for any structure G $p(\underline{\theta}|G)$ can be derived from a point-specification θ_0 of a complete model and from the number of a priori seen complete cases. To state this formally, we need the following concepts. The concept of parameter independence ([132, 33]) is as follows:

Definition 1.1.20. *For a Bayesian network structure G, the global parameter independence assumption means that*

$$p(\underline{\theta}|G) = \prod_{i=1}^{n} p(\underline{\theta}_i|G), \tag{1.15}$$

where $\underline{\theta}_i$ denotes the parameters corresponding to the conditional $p(X_i|\operatorname{Pa}(X_i))$ in G. The local parameter independence assumption means that

$$p(\underline{\theta}_i|G) = \prod_{j=1}^{q_i} p(\underline{\theta}_{ij}|G), \tag{1.16}$$

where q_i denotes the number of parental configurations ($\operatorname{pa}(X_i)$) for X_i in G and $\underline{\theta}_{ij}$ denotes the parameters corresponding to the conditional $p(X_i|\operatorname{pa}(X_i)_j)$ in some fixed ordering of the $\operatorname{pa}(X_i)$ configurations. The parameter independence assumption means global and local parameter independence.

The concept of likelihood equivalence extends observational equivalence of the structure coherently to the parameters ([74, 59]).

Definition 1.1.21. *The likelihood equivalence assumption means that for two observationally equivalent Bayesian network structures G_1, G_2,*

$$p(\underline{\theta}_V|G_1) = p(\underline{\theta}_V|G_2), \tag{1.17}$$

where $\underline{\theta}_V$ denotes a non-redundant set of the multinomial parameters for the joint distribution over V. (The multinomiality of local models ensures distributional equivalence and that the Jacobian for parameter transformation exists.)

Now the following theorem can be stated [59, 74].

Theorem 1.1.5 ([59, 74])**.** *The assumption of positive densities, likelihood equivalence and parameter independence for complete structures G_c implies that $p(\underline{\theta}_V)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n}$.*

The $p(\underline{\theta}_i|G_i) = J_{G_i} p(\underline{\theta}_V)$, where J_{G_i} is the Jacobian of the transformation from $\underline{\theta}_V$ to $\underline{\theta}_{G_i}$. It is remarkable that a structure level acausal constraint (i.e., likelihood equivalence of structures with multinomial local dependency models) implies a strong parameter-level constraint (i.e., Dirichlet parameter priors). To state the following theorem it is convenient to rewrite the hyperparameters as $N' = \sum_{x_1,\ldots,x_n} N_{x_1,\ldots,x_n}$ called *prior or virtual sample size* and $p(x_1, \ldots, x_n|\xi^+) = N_{x_1,\ldots,x_n}/N'$. Furthermore, we need the following concept:

Definition 1.1.22. *The parameter modularity assumption means that if* $\mathrm{pa}(X_i)$ *are identical in two Bayesian network structures* G_1, G_2*, then*

$$p(\underline{\theta}_{ij}|G_1) = p(\underline{\theta}_{ij}|G_2), \tag{1.18}$$

where $\underline{\theta}_{ij}$ *denotes the parameters corresponding to the conditional* $p(X_i|\mathrm{pa}(X_i)_j)$ *in some fixed ordering of the* $\mathrm{pa}(X_i)$ *configurations.*

The assumption of parameter modularity allows to induce parameter distributions for incomplete models from the parameter prior of a complete model.

Theorem 1.1.6 ([59, 74])**.** *If* N' *is the global prior sample size and* $p(\underline{\theta}_V)$ *is a Dirichlet distribution with hyperparameters* $N_{x_1,\ldots,x_n} = N'p(x_1,\ldots,x_n)$ *and the parameter modularity assumption holds and for all complete networks* G_c*,* $p(G_c) > 0$*, then for any network structure* G *the parameter independence and the likelihood equivalence holds and the decomposed distribution of the parameters is the product of Dirichlet distributions*

$$p(\underline{\theta}|G,\xi^+) \quad = \quad \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N'p(X_i=k,\mathrm{pa}(X_i,G|\xi^+)=\mathrm{pa}_{ij})-1}, \tag{1.19}$$

where r_i *denotes the number of values of* X_i*,* q_i *denotes the number of parental configurations* $\mathrm{pa}(X_i,G)$ *and* pa_{ij} *denotes the values of the parents for the* j*th parental configuration in some fixed ordering of the* $\mathrm{pa}(X_i)$ *configurations.*

Th. 1.1.6 offers a practical method to specify (likelihood equivalent) parameter priors for all the structures: by specifying point parameters for a complete or for a maximally detailed model $p(\underline{V}|G_c,\xi^+)$ and expressing confidence by specifying a prior sample size N' representing the complete cases underlying the point estimates (see Section **??** and Section **??** for its application). Then for any other model G we can compute hyperparameters according to the theorem. Note that these are determined by the marginalization of the distribution $p(\underline{V}|G_c,\xi^+)$ into $p(\underline{V}|G,\xi^+)$ (i.e., by averaging over potentially missing parents) and it can be shown that this is the best approximation w.r.t. the KL distance of $p(\underline{V}|G_c,\xi^+)$ by a product of lower order distributions compatible with G [127].

However, Th. 1.1.6 also indicates that incomplete prior observations inducing different confidence for various parts of the network cannot be incorporated without violating these assumptions. For example, specifying a parameter prior as product of Dirichlets according to a structure with hyperparameters incompatible w.r.t. the theorem cannot be transformed to a product of Dirichlets for another observationally equivalent structure (i.e., the parameter prior will be different for observationally equivalent structures). In this case, the prior knowledge can be represented by a collection of incomplete cases called *prior database* instead of a *prior data set* with complete cases [66].

In case of a fixed structure G, the usage of Dirichlets with parameter independence can be attractive on its own right to specify a parameter distribution $p(\underline{\theta}|G,\xi^+)$ as follows

$$p(\underline{\theta}|G,\xi^+) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \mathrm{Dir}(\underline{\theta}_{ij}|N_{ij}) \propto \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}-1}. \tag{1.20}$$

FULLVERSION>

Note that the overall distribution $p(X_1,\ldots,X_n,\underline{\theta}_1,\ldots,\underline{\theta}_n)$ can be represented in an augmented Bayesian network by introducing extra root nodes for the continuous vector-valued parameters $\underline{\theta}_{ij}$ for $i = 1,\ldots,n$ and $j = 1,\ldots,q_i$ with Dirichlet distributions $\mathrm{Dir}(\underline{\theta}_{ij}|\underline{N}_{ij})$ and special conditional distributions $p(X_i = k|pa_{ij},\underline{\theta}_i) = \theta_{ijk}$ for the variables X_i [132].

<FULLVERSION

1.1.5.2 Structure priors for Bayesian network models

The Bayesian approach to the parameters of Bayesian network models (reported from the end of the eighties [130, 131, 33]) provided answers for many long-standing objections against the elicitation and usage of complex probabilistic models ([20]). The Bayesian approach to the structure of Bayesian networks was similarly proposed from the beginning of the nineties, but was hindered by the high computational demand. An ordering-specific, analytic approach was reported in [17], general analytic results and methodology were reported in [29], and the application of MCMC methods to perform Bayesian inference over structural features in [100]). With the increase in computational resources it became possible to investigate structural properties of Bayesian networks. Consequently, recently there is much emphasis on the automated or manual construction of the structure prior $p(G)$ for incorporation and for evaluation against a reference as well (see Section **??**, **??**, **??** and **??**). Note the structure prior $p(G)$ complements the earlier investigated parameter prior $p(\underline{\theta}|G)$.

1.1.5.2.1 Using a prior data set Whereas the parameter prior and the structure prior can be specified independently, the structure prior can be induced from the *prior data set* $D^+_{N'}$ using Eq. 1.51 [101]. FULLVERSION>That is by denoting the informative background belief with ξ^+ and the noninformative with ξ^-,

$$p(\underline{\theta}|G,\xi^+) = p(\underline{\theta}|G,D^+_{N'},\xi^-), \text{implied by the assumptions} \tag{1.21}$$
$$p(G|\xi^+) = p(G|D^+_{N'},\xi^-), \text{informally.} \tag{1.22}$$

<FULLVERSION

1.1.5.2.2 Using reference structure and substructures Other suggestions for the structure prior include the use of *deviation priors* (penalizing the deviations from a prior "reference" structure) and the *feature priors* (penalizing the presence and absence of various independent or dependent structural features).

The deviation prior [74] is defined by a "reference" network structure G_0 and a probability κ penalizing each missing or extra edge e_{ij} independently:

$$p(G) \propto \kappa^{\delta}, \text{where } \delta = \sum_{1 \leq i < j \leq n} 1(1(e_{ij} \in G) \neq 1(e_{ij} \in G_0)).$$

The *feature priors* are defined proportionally by the product of priors for the individual features (as they were totally independent). By denoting the value of feature F_i in G with $F_i(G) = f_i$, $i = 1, \ldots K$, we have

$$p(G) = c \prod_{i=1}^{K} p(F_i(G)), \tag{1.23}$$

where the c normalizing constant deals with the probability of inconsistent feature combinations $f_1, \ldots, f_K$. The possible structural features include the undirected edges or compelled edges (as direct relations or direct causal relations under the causal Markov Assumption), pairwise or partial ancestral ordering (related to causal ordering), relevance relations (Markov blankets) and even arbitrary subgraphs. However, these features are dependent in general, because of the global DAG constraint, so either the feature set should be selected carefully, or preprocessing can be applied to increase its approximation or the strength of the attached prior should reflect its approximative nature.

1.1.5.2.3 Modular priors It is particularly useful in the Bayesian analysis, if the features are "modular" in the following sense [17, 29, 74, 55]

Definition 1.1.23. *The structure modularity holds, if each feature function $F_i(G)$ depends only on the parents of X_i for $i = 1, \ldots n$, defining the modular prior*

$$p(G) \propto \prod_{i=1}^{n} p(\mathrm{pa}(X_i, G)). \tag{1.24}$$

Because the DAG constraint creates dependencies, the modular features are not independent (i.e., $(F_i(G) \not\perp\!\!\!\perp F_j(G) | DAG(G))$, see Section 2.1.6), but it provides an efficient approach to define a decomposable ratio for the priors of valid structures (for certain automated corrections of the distortion because of the DAG constraint, see [18]).

A generalization of the modular prior is the *ordering-modular prior*, when modularity holds only conditionally on the orderings.

1.1.5.2.4 Edge priors With further assumption about the a priori independence of membership of edges in parental sets, we get the *directed pairwise prior* that defines the probability of each parental set as a product of individual arc probabilities. In general, the prior is defined only proportionally as follows by denoting the parents of X_i with $\mathrm{pa}(X_i) = \{\mathrm{pa}(X_i)_1, \ldots, \mathrm{pa}(X_i)_{L_i}\}$:

$$p(\mathrm{pa}(X_i)) \quad \propto \quad \prod_{k=1}^{L_i} p(\mathrm{pa}(X_i)_k \in \mathrm{Pa}(X_i)) \prod_{Y \notin \mathrm{pa}(X_i)} (1 - p(Y \in \mathrm{Pa}(X_i))).$$

Originally, modular priors and directed pairwise priors were suggested conditional on a fixed ordering $\prec_0$ of the variables [17],

$$p(\mathrm{pa}(X_i)) = \prod_{\substack{X_j \prec_0 X_i \\ X_j \in \mathrm{pa}(X_i)}} p(X_j \in \mathrm{Pa}(X_i) | \prec_0) \prod_{\substack{X_j \prec_0 X_i \\ X_j \notin \mathrm{pa}(X_i)}} (1 - p(X_j \in \mathrm{Pa}(X_i) | \prec_0)), \quad (1.25)$$

in which case these features remain independent in the joint distribution over DAGs compatible with the ordering $\prec_0$. In fact, the assumption of "edge independence" first appeared implicitly in the noisy-OR canonical local dependency model, because its parameterization can be interpreted as encoding the probability of the edges [114].

To reach independent pairwise features for DAGs without constraining the ordering, we have to further simplify the features to avoid global constraints due to their interactions. Note that with independence, the marginals are not distorted and the prior is normalized, which allows the introduction of hyperparameters for modifying the prior to satisfy higher-order constraints as follows. By defining the prior over the skeleton in a pairwise manner (i.e., by retaining only the directness and omitting directionality), we get the *undirected pairwise prior* $p_{ij} \triangleq p(X_j \in \mathrm{Pa}(X_i) \vee X_i \in \mathrm{Pa}(X_j))$ represents the beliefs in direct influence between X_i and X_j [8]. The edge probabilities define the following prior probability for a structure G:

$$P(G|\xi) \propto \prod_{i=1}^{n} \prod_{j=1}^{i-1} p_{ij}^{1(e_{ij} \in G)} (1 - p_{ij})^{1(e_{ij} \notin G)}. \quad (1.26)$$

The expectation of the number of edges L is given by $\sum_{0<i<j<n} p_{ij}$. Assuming that there is an a priori estimate for the number of direct influences in the overall model or related to a single variable, the prior p_{ij} can be scaled by an exponent ν to approximate this edge density in the prior Bayesian network (see [8]). By denoting the value that scales the expectation of the number of parental edges to L_0 with $\nu(L_0)$ we define the following scaling (it is always possible if we apply a lower limit $\epsilon < p_{ij}$ for the edge probabilities):

$$q_{ij} \triangleq p_{ij}^{\nu(L_0)}, \quad \text{with } \nu(L_0) \text{ so that } \sum_{0<i<j<n} q_{ij} = L_0. \quad (1.27)$$

The deviation prior and the feature prior can be combined into a *feature-deviation prior* to utilize the prior information about a global model and about local features by penalizing the differences from a global model w.r.t. selected features. For example, by replacing the uniform κ with an edge-specific pairwise

prior p_{ij}:

$$P(G|G_0, p_{ij}) \quad \propto \quad \prod_{1 \leq i < j \leq n} p_{ij}^{1(\{(e_{ij} \in G) \wedge (e_{ij} \notin G_0)\})} (1 - p_{ij})^{1(\{(e_{ij} \notin G) \wedge (e_{ij} \in G_0)\})}.$$

Note that the scaling of p_{ij} provides an option to control the penalization (i.e., to express the prior beliefs in the prior structure). These priors except the undirected pairwise prior assign potentially different values for observationally equivalent structures (i.e., violates the structural *prior equivalence* property [74]). Because they are closely related to the causal, mechanism-based interpretation of Bayesian networks, they offer the possibility of representing a priori beliefs about the individual mechanisms in the domain and we call them *causal (structure) priors* vs *acausal (structure) priors*. FULLVERSION> Note that if the hypotheses are the observational equivalence classes in an acausal approach, then a prior $p(G^{\sim})$ over their representant PDAGs $G^{\sim}$ is either specified directly (i.e., independently of member DAGs) or by inducing it from a possibly causal structure prior for the DAGs $p(G)$ as

$$p(G^{\sim}) = p(G : G \in G^{\sim}). \tag{1.28}$$

<FULLVERSION
FULLVERSION>

1.1.5.2.5 Augmented representation of structure priors Finally, we mention an explicit representation of structure priors in a restricted case. It applies the same technique as the explicit representation of Dirichlet parameter priors for the multinomial case. Assuming furthermore conditional independence of the parental sets given a fixed ordering $\prec_0$ of the variables, the overall distribution $p(X_1, \ldots, X_n, \underline{\theta}_1, \ldots, \underline{\theta}_n, \mathrm{pa}(X_1), \ldots, \mathrm{pa}(X_n)| \prec_0)$ can be represented in an augmented Bayesian network by introducing an additional extra root nodes for the parental sets at each variable in a complete network or in a "covering" DAG structure that includes all parental sets with nonzero probability. For simplicity, we assume independence between the structure prior and the parameter prior, so we assume that the parameter priors for various parental sets are the appropriate marginals (fitting to the property of the Dirichlet distribution that the marginal $\underline{\theta}' \subseteq \underline{\theta}$ for $\underline{\theta} \sim \mathrm{Dir}(\underline{\theta}|\underline{\alpha})$ is $\mathrm{Dir}(\underline{\theta}'|\underline{\alpha}')$ with hyperparameters $\underline{\alpha}' \subseteq \underline{\alpha}$ corresponding to the not marginalized variables, see [12]). The corresponding conditional distributions for each X_i are $p(X_i = k| \operatorname{Pa}(X_i) = \mathrm{pa}(X_i), \underline{\Theta}_i = \underline{\theta}_i, \mathrm{pa}(X_i) = \mathrm{pa}_{ij},) = \theta_{ijk}$.

<FULLVERSION
FULLVERSION>

1.1.6 Extensions of the Bayesian network representation

We refer the reader to the following sources regarding the Bayesian multinets [61, 19] and qualitative Bayesian networks [145, 121].

<FULLVERSION

1.2 Inference methods

The Bayesian network model makes possible various types of inferences thanks to the possibility of

1. the multiple interpretation, such as causal vs. probabilistic,

2. the multilevel interpretation, such as at the level of domain values, independence relations or causal relations,

3. the adoption of the Bayesian framework at the parameter or the structure level,

4. embedding the Bayesian network model into a larger knowledge base to formulate more complex propositions (see Chapter **??**).

Next we catalogue these inferences, summarize results and techniques used in the thesis.

1.2.1 Inference over values with observations

The goal in the following cases is to compute the value of marginal or conditional probabilities over domain values $P(\underline{y}|\underline{x})$ and possibly related quantities.

1.2.1.1 Fixed parameter and fixed structure

In the simplest case the structure and the parameters of a Bayesian network model are fixed. The computation of $p(\underline{y}|\underline{x})$ is NP-complete in general in the number of variables [27]. However in practice, an exact inference method has demonstrated its applicability, the *clique-tree* or *join-tree* algorithm [130]. We used this exact algorithm following the recommendations for implementation from [79]. The algorithm is exponential in the largest clique size of an intermediate Markov network and our experience similarly shows that the networks arisen in knowledge engineering and learning can be efficiently managed with this algorithm. OPTIONAL>For estimating a marginal $p(y)$ with i.i.d. Monte Carlo sampling $p(V)$ the Hoeffding-inequality in Eq. **??** gives the sample complexity $N(\epsilon, \delta) = 1/\epsilon^2 \log 1/delta)$, however with the increasing number of variables in the condition $p(y|\underline{x})$ the number of useful samples drops exponentially. <OPTIONAL A general result shows that the Monte Carlo approximation is hard as well: if $NP \nsubseteq RP$, then there is no random algorithm with polynomial time-complexity, whose estimate $\hat{p}$ is $|p(y|\underline{x}) - \hat{p}| < \epsilon$ accurate with δ confidence for all $\epsilon, \delta < 1/2$ [34].

1.2.1.2 Bayesian parameter and fixed structure

In case of a Bayesian approach to parameters with a fixed structure G, a parameter distribution $p(\underline{\theta}|G)$ is specified. The conditional probability over the domain values $p(\underline{y}|\underline{x},\underline{\Theta})$ is a random variable and its mean, variance, credible regions are the target.

If the parameter distribution $p(\underline{\theta}|G)$ is specified according to the conditions of Th. 1.1.6, then it guarantees that $p(\underline{Y}|\underline{x},\underline{\Theta})$ has a Dirichlet distribution with hyperparameters $Np_0(\underline{Y},\underline{x})$, so the mean and credible regions can be efficiently computed.

If the parameter distribution $p(\underline{\theta}|G)$ is specified by using Dirichlet distributions and assuming parameter independence, but with arbitrary hyperparameters according to Eq. 1.20, then the marginal distribution $\bar{p}(X_1,\ldots,X_n)$ over the domain values is given by

$$\begin{aligned}\bar{p}(x_1,\ldots,x_n) &= \int p(x_1,\ldots,x_n,\underline{\theta}_1,\ldots,\underline{\theta}_n)\prod_{i=1}^{n} p(\underline{\theta}_i)\,\mathrm{d}\underline{\theta} &(1.29)\\ &= \prod_{i=1}^{n}\int p(x_i|\,\mathrm{pa}(x_i),\underline{\theta}_i)p(\underline{\theta}_i)\,\mathrm{d}\underline{\theta}_i &(1.30)\\ &= \prod_{i=1}^{n}\bar{p}(x_i|\,\mathrm{pa}(x_i)), &(1.31)\end{aligned}$$

where the $\bar{p}(x_i|\,\mathrm{pa}(x_i))$ are the local mean probabilities [132, 131, 33]. The expectations of the parameters at each node for each parental configuration (i.e., the integration of the Dirichlets) have a closed form solution (see Eq. **??**)

$$\bar{p}(X_i = k|\,\mathrm{pa}(X_i) = \mathrm{pa}_{ij}) = \mathrm{E}_{\underline{\Theta}_i}[p(X_i = k|pa_{ij},\underline{\Theta}_i)] = \mathrm{E}_{\underline{\Theta}_{ij}}[\Theta_{ijk}] = N_{ijk}/N_{ij}.$$

The closed solution for $\bar{p}(X_1,\ldots,X_n)$ ensures that any Bayesian inference over the domain values can be equivalently performed using this mean-valued point parameters, instead of Bayesian averaging over the parameter space [132, 29], that is

$$\mathrm{E}_{\underline{\Theta}}[p(\underline{y}|\underline{x},\underline{\Theta})] = \bar{p}(\underline{y}|\underline{x}). \tag{1.32}$$

For the computation of variance and credible regions in this case we used a Monte Carlo sampling algorithm, an efficient method for the approximation is reported in [6]. FULLVERSION> For the representation of the overall distribution $p(X_1,\ldots,X_n,\underline{\theta}_1,\ldots,\underline{\theta}_n)$ by an augmented Bayesian network see Section 1.1.5.2.5. The parameter independence assumption guarantees that the independencies of the domain variables are the same in the marginalized distribution $\bar{p}(X_1,\ldots,X_n)$ and in the original, so it can be represented by the original DAG G. <FULLVERSION

1.2.1.3 Bayesian parameter and structure

In the general case there is a distribution over the structures $p(G)$ and over the corresponding parameters $p(\underline{\theta}|G)$. The conditional probability over the domain values $p(\underline{y}|\underline{x}, \underline{\Theta}, G)$ is a random variable itself and its mean, variance, credible regions are the target. The computation of these quantities, for example of the mean involves both a summation over the space of DAGs and the integration over the parameters.

$$\bar{p}(\underline{y}|\underline{x}) = \mathrm{E}_{p(G)}[\mathrm{E}_{p(\underline{\theta}|G)}[p(\underline{y}|\underline{x}, \underline{\theta}, G)]]. \tag{1.33}$$

In general there is no closed formula for this quantity and for the marginal distribution ("superparameters") $\bar{\bar{p}}(X_1, \ldots, X_n)$, as in the case of fixed structure and Bayesian parameter.

FULLVERSION>

However if the structure prior satisfies conditional independence of the parental sets for each variable given an ordering $\prec_0$, then a covering Bayesian network G^c can be defined that $\forall i: \; p(\mathrm{pa}(X_i)) > 0 \Rightarrow \mathrm{pa}(X_i) \subseteq \mathrm{pa}(X_i, G^c)$. If parameter independence holds for the parameter prior, then the overall distribution is decomposed as

$$p(X_1, \ldots, X_n, \underline{\Theta}_1, \ldots, \underline{\Theta}_n, \mathrm{Pa}(X_1), \ldots, \mathrm{Pa}(X_n)| \prec_0) \tag{1.34}$$

$$= \prod_{i=1}^{n} p(X_i|X_1, \ldots, X_{i-1}, \underline{\Theta}_i, \mathrm{Pa}(X_i)) p(\underline{\Theta}_i| \mathrm{Pa}(X_i)) p(\mathrm{Pa}(X_i)) \tag{1.35}$$

This distribution can be represented in an augmented Bayesian network by introducing an additional extra root node for the parental sets at each variable in a DAG G^c that covers any parental set with nonzero probability and special conditional distributions for each X_i that $p(X_i = k|pa_{ij}, \underline{\theta}_i, \mathrm{pa}(X_i)) = \theta_{ijk}$ (allowing dependence of the parameter prior on the structure prior at each node, $p(\underline{\theta}_i| \mathrm{pa}(X_i))$). The marginal distribution $\bar{\bar{p}}(X_1, \ldots, X_n)$ is marginalized in a decomposed fashion

$$\bar{\bar{p}}(x_1, \ldots, x_n) = \mathrm{E}_{G|\prec_0}[\mathrm{E}_{\underline{\Theta}|G}[\bar{p}(x_1, \ldots, x_n|\underline{\Theta}, G)]] \tag{1.36}$$

$$= \mathrm{E}_{G|\prec_0}[\prod_{i=1}^{n} \bar{p}(x_i| \mathrm{pa}(X_i, G))] \tag{1.37}$$

$$= \sum_{P(G|\prec_0)>0} \prod_{i=1}^{n} \bar{p}(x_i| \mathrm{pa}(X_i, G)) p(\mathrm{pa}(X_i, G)| \prec_0) \tag{1.38}$$

$$= \prod_{i=1}^{n} \sum_{\mathrm{pa}(X_i|\prec_0)>0} \bar{p}(x_i| \mathrm{pa}(x_i)) p(\mathrm{pa}(X_i)| \prec_0) \tag{1.39}$$

$$= \prod_{i=1}^{n} \bar{\bar{p}}(x_i| \mathrm{pa}(x_i)). \tag{1.40}$$

That is the marginal distribution is the product of expectations of the parameters at each node, which have the following closed form (i.e., the summation over the parental sets and integration of the Dirichlets)

$$\bar{\bar{p}}(X_i = k|\,\mathrm{pa}(X_i, G^c) = \mathrm{pa}_{ij}) = \sum_{p(\mathrm{pa}(X_i)|\prec_0)>0} p(\mathrm{pa}(X_i)|\prec_0)\,\mathrm{E}_{p(\underline{\theta}_{ij}|\,\mathrm{pa}(X_i))}[\theta_{ijk}], \tag{1.41}$$

where $\mathrm{pa}(X_i, G^c)$ denotes a "covering" parental set. The existence of a closed solution $\bar{\bar{p}}(X_1, \ldots, X_n)$ for G^c ensures that in general any Bayesian inference over the domain values can be equivalently performed using this point parameters, instead of the Bayesian summation over the DAGs compatible with the ordering $\prec_0$ and averaging over the parameter space:

$$\mathrm{E}_{G|\prec_0}[\mathrm{E}_{\underline{\Theta}|G}[p(\underline{y}|\underline{x}, \underline{\Theta}, G)]] = \bar{\bar{p}}(\underline{y}|\underline{x}). \tag{1.42}$$

where $\bar{\bar{p}}$ denotes the parameters in the covering DAG G^c. This "smoothed" parameters and their computation was suggested in [17]. On the same bases, a similar result for the posterior distributions in the special case of a naive Bayesian network were derived in [36]. The replacement of summation and production in step 1.39 conditional on an ordering were similarly used in deriving closed forms for the conditional probabilities of structural features for a given ordering in [54] (see Section 2.5.2.1).

<FULLVERSION

1.2.2 Inference over domain values with interventions

In the thesis the analyzed data set is observational. The interventional "do" semantics was necessary only for the causal interpretation, which is used in developing models for the analysis of domain literature with Bayesian networks. For the conversion of causally defined quantities $P(y|do(x), z)$ into "do"-free observational quantities $P(y|w)$ (question of identifiability) or to more appropriate causal quantities $P(y|do(x'), z')$ see [115, 58, 116].

1.2.3 Inference over model parameters

After the inference over the domain values we summarize now a basic result about the inductive Bayesian inference over the parameters. Let us assume the observation of a complete case x, parameter independence, and Dirichlet priors $\underline{\theta}_{ij} \sim \mathrm{Dir}(\alpha_{ij1}, \ldots, \alpha_{ijr_i})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, q_i$ (where r_i is the number of values of variable X_i, q_i are the number of parental configurations $\mathrm{pa}(X_i, G)_j = \mathrm{pa}_{ij}$ for variable X_i w.r.t. the Bayesian network G). Then the a posteriori distribution for an "observed" parameter family $\underline{\theta}_{ij_0}$ where j_0 is the

index of $pa_i(x)$ is given by

$$p(\underline{\theta}|x) = \frac{\prod_{i=1}^{n} p(x_i|pa_i(x), \underline{\theta}_{ij_0})p(\theta_{ij_0})}{p(x)} \prod_{i=1}^{n} \prod_{j \neq j_0} p(\theta_{ij}) \tag{1.43}$$

$$\propto \prod_{i=1}^{n} \theta_{ij_0 x_i} \mathrm{Dir}(\theta_{ij_0}|\underline{\alpha}_{ij_0}) \tag{1.44}$$

$$\propto \prod_{i=1}^{n} \mathrm{Dir}(\theta_{ij_0}|\alpha_{ij_0 1}, \ldots, \alpha_{ij_0 x_i} + 1, \ldots, \alpha_{ij_0 r_i}), \tag{1.45}$$

which shows that the parameter posterior preserves the parameter independence property and that local standard Bayesian updating is performed on the hyperparameters of the "observed" Dirichlets (the hyperparameters for the other parameter families $\underline{\theta}_{i_0 j}$ with $j \neq j_0$ are unchanged).

1.2.4 Inference over model structures

The posterior of the Bayesian network (structure) is the product of the model likelihood and the structure prior.

$$p(G|D_N) \propto p(G) \int p(D_N|\underline{\theta}, G)p(\underline{\theta}|G)\, \mathrm{d}\underline{\theta} = p(G)p(D_N|G). \tag{1.46}$$

To reach a closed form for the likelihood term we continue with the assumption of the previous paragraph: N complete observations, i.i.d. multinomial sampling, Bayesian network model with parameter independence and Dirichlet parameter priors following [29, 131, 74]. Under these assumptions the observation of a complete case results in a local standard Bayesian updating of the hyperparameters of the "observed" Dirichlets retaining the parameter independence (see Eq. 1.43). The maintained parameter independence allows a standard parental decomposition w.r.t. the Bayesian network G for each observation (see Eq. 1.29), which allows the following rearrangement:

$$p(x^{(1)}, \ldots, x^{(N)}|G) = \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)}|pa_i^{(l)}) \tag{1.47}$$

$$= \prod_{i=1}^{n} \prod_{l=1}^{N} p(x_i^{(l)}|pa_i^{(l)}) \tag{1.48}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{l=1}^{N} p(x_i^{(l)}|pa_{ij})^{1(\mathrm{pa}_{ij} = \mathrm{pa}_i^{(l)})}, \tag{1.49}$$

where $pa_i^{(l)}$ denotes the value(s) of parental set of X_i in case l. The marginal probability of the data for a single Dirichlet prior and multinomial sampling was derived in Eq. **??** and Eq. **??**, **??**. Now if r_i denotes the cardinality of the discrete values of variable X_i, α_{ijk} the initial Dirichlet hyperparameters, and

n_{ijk} the number of occurrences for the variable X_i, its parental configuration pa_{ij} and its value r_k, then for each variable X_i and parental configurations j independently

$$\begin{aligned}\prod_{l=1}^{N} p(x_i^{(l)}|pa_{ij},G)^{1(\mathrm{pa}_{ij}=\mathrm{pa}_i^{(l)})} &= \frac{\prod_{k=1}^{r_i}(\alpha_{ijk}\ldots(\alpha_{ijk}+n_k))}{\alpha_{ij+}\ldots(\alpha_{ij+}+n)} \qquad (1.50)\\ &= \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+}+n_{ij+})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})},\end{aligned}$$

Putting everything together, if the prior satisfies the structure modularity, then the posterior of the Bayesian network structure has the following product form

$$p(G|D_N) \propto \prod_{i=1}^{n} p(\mathrm{Pa}(X_i,G))S(X_i,\mathrm{Pa}(X_i,G),D_N) \text{ where} \qquad (1.51)$$

$$S(X_i,\mathrm{Pa}(X_i,G),D_N) = \prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+}+n_{ij+})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+n_{ijk})}{\Gamma(\alpha_{ijk})}.$$

FULLVERSION> For a condition when the posterior probability of a Bayesian network structure decomposes into a product of terms expressing the (unconditional) posterior probability of the parental set $\mathrm{Pa}(X_i,G)$ for variable X_i given the data, see Section 1.4.1.
<FULLVERSION

1.3 Knowledge engineering

As discussed in Section 1.1 and enumerated in List 1.2, the Bayesian network can serve as a multilevel (structural or parametric), multiple-point-of-view (probabilistic or causal) representation of the domain. Besides being a model ("*surrogate*"), it fulfills other important roles of a knowledge representation (following the proposed roles from [37]): *ontological* (what kind of objects and relations exists in the domain), *inferential* (what kind of inference is possible in the domain), *computational* (what kinds of embedding of the model and real-world applications are possible), *communicational* (what kind of understanding and communication is supported by the model between domain experts, knowledge engineers, and users).

Because of the versatility of the Bayesian network representation as a knowledge representation, knowledge engineering methodologies are necessary for proper and efficient real-world applications. Particularly, if a Bayesian network model serves as a probabilistic expert system or as the engine of a decision support system, its construction should be subject to engineering standards, which include specifications with quantitative quality measures for the process and

the product and complexity measures related to budgetary, personal and time limits, etc. However, these issues are still largely unexplored and the knowledge engineering of Bayesian networks is still in its early stage (described for example in [1]). The main reasons are the versatility of the representation mentioned above, the continuing extensions of the representation and the newly evolved knowledge engineering context of the "e-science" era.

FULLVERSION>

1.3.1 The "classical" knowledge engineering

First, we summarize the general steps of the "classical" knowledge engineering of Bayesian networks, which are synchronous with the general knowledge engineering of logical knowledge bases [126]. This list also incorporates our experience of model construction in the ovarian cancer domain [11].

1. *I*dentification of purpose, scopes and levels. The major factors underlying the purpose of Bayesian network modeling are the following: probabilistic or causal interpretation, structural or parametric level, decision-support or explanation, domain-wide or classification (i.e., are there any specifically interesting variable(s) or model feature). Next a reasonable scope and level have to be identified including variables and a level of granularity.

2. *C*ollection of informal knowledge. Make a list of all prior knowledge about variables, discretizations, existing dependency models. Classify different types of priors that exist (from exactly specified prior sub-models to high level guesses about qualitative dependencies). Conversion formulas can be constructed to compile the raw prior knowledge into compatible with the conditions of the task and the format of the Bayesian network.

3. *A*doption of terminology and ontology Adopt a terminology hopefully from an existing domain ontology and select a "coverable variable set" that seems to be quantifiable from the prior background knowledge.

4. *S*tructure elicitation Specify a complete domain model by following standard construction mechanism for Bayesian networks based on either the Markov conditions (see Def. 1.1.5, 1.1.6, 1.1.7 or on the causal Markov assumption (Def. 1.1.18). Consider also the existing prior sub-models.

5. *P*arameter and hyperparameter elicitation. Perform parameter and hyperparameter elicitation [141, 60, 73, 62, 125, 109, 45]. Construct secondary conversion models and formulas to quantify the final model, considering consistency issues [112].

6. *S*ensitivity analysis, refinement, verification and validation Perform sensitivity analysis, possibly refining the model [31]. Evaluate the performance of the system on test cases or possibly on benchmark cases and in real-world circumstances.

<FULLVERSION

The "classical" knowledge engineering of Bayesian networks in complex domains was criticized as aiming at a "one-shot" and "monolithic" Bayesian network. Its extension led to new representational methods, especially to modularized representations [119, 104, 43, 111, 97]. The object-oriented and frame-based approaches were partly responses to problems of modularization, validation, verification, maintenance and reuse [96, 88, 89]. Other approaches extended the Bayesian network representation itself. The multi-net representation was partly a response to a problem related to the elicitation and representation of contextual independencies [61]. The qualitative Bayesian networks and other semantic extension of the represented relations were partly a response to the problem of the elicitation and refinement of parameters [145, 99, 121], similarly to the investigation of special local dependency models [72, 52].

FULLVERSION>

The semi-final result of Bayesian knowledge engineering is a prior distribution over the model space and parameters, so its evaluation in Bayesian data analysis corresponds to the general issue of model evaluation and comparison. A speciality of this mixture of knowledge engineering and data analysis is that because of the modularized description of the model, the evaluation can support the detailed compatibility of the data and the parts, modules of the model, possibly identified by the attached semantic context.

<FULLVERSION

FULLVERSION>

First we discuss the application of the predictive sequential ("prequential") analysis for Bayesian networks. Second we discuss the relation of the sequential analysis of the posteriors for properties of the model. Third, we discuss methods to define general informative utility for the network structures based on an ABN-KB, which allows a full-scale decision theoretic evaluation of the posteriors. Next, we summarize an evaluation of structure posteriors using a utility-free measure and reference structure. Finally, heuristic methods are summarized to evaluate and compare a (structure) posterior against another (structure) posterior.

<FULLVERSION

1.4 Prequential analysis by Bayesian networks

The Bayes factor in Eq. **??** is typically used in a non-sequential setup. In Section **??** we summarized the prequential framework, which evaluates the model from a forecasting point of view by scoring its sequential predictions based on the actual observations [131, 33]. Because of its sequentiality, it also offers a sample-by-sample evaluation of the compatibility of the data and the model (see Section **??**). For us, the case of a (discrete and finite) probabilistic forecasting system (PFS) is relevant predicting a distribution $p(X_i|x_1, \ldots, x_{i-1})$ for the observation at step i. FULLVERSION> We also reviewed the advan-

tages of a logarithmic scoring function (see Th. **??** and comments afterwards). <FULLVERSION For the application of the prequential evaluation for Bayesian networks and parts of the model we have to interpret them as PFSs and compare them using the logarithmic score (see Eq. **??**).

The PFS shall be defined as a Bayesian forecasting system (see Section **??**) using a fixed Bayesian network structure with Dirichlet parameter priors under the condition of parameter independence.

The *global monitor* tracks the overall performance of the Bayesian network model $M = (G, \underline{\theta})$ over a data set D_N:

$$S(M; D_N) = \sum_{l=1}^{N} -\log p(x^{(l)}|x^{(1)}, \ldots, x^{(l-1)}, M) \tag{1.52}$$

$$= -\log p(x^{(1)}, \ldots, x^{(N)}|M). \tag{1.53}$$

The equation shows the ordering-insensitivity and batch-sequential equivalence of the log-score for PFSs. By noting that this is the model likelihood derived in Eq. 1.47, 1.50, the score is given by

$$S(M; D_N) = -\log \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \frac{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk} + n_{ijk})}{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk})}. \tag{1.54}$$

In line with the decomposition w.r.t. the structure (see Eq. 1.51) various monitors were suggested for the parts of the Bayesian network model.

The (unconditional) *node monitor* tracks the performance of the Bayesian network model M w.r.t. a given variable X_i:

$$S(X_i; D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)}|x^{(1)}, \ldots, x^{(l-1)}, M). \tag{1.55}$$

Two variants of the node monitor are the conditional node monitors, because the target variable is predicted conditioned on all the other variables or only on the parental set in the actual case. This monitor was called a "conditional node monitor" [131]), but in the case of complete data assumption this is equivalent with scoring the predictive performance of the Markov blanket subgraph $\mathrm{MBG}(X_i, G)$. So we will adopt the term *Markov blanket subgraph monitor*.

$$S(\mathrm{MBG}(X_i, G); D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)}|x^{(l)} \setminus \{X_i\}, x^{(1)}, \ldots, x^{(l-1)}). \tag{1.56}$$

Conditioning only on the parental set in a causal approach, we get the *mechanism monitor* that tracks the performance of the parental set $\mathrm{Pa}(X_i, G)$ in forecasting a variable:

$$S(\mathrm{Pa}(X_i, G); D_N) = -\log \prod_{l=1}^{N} p(x_i^{(l)}|\,\mathrm{pa}(X_i) = \mathrm{pa}_i^{(l)}, x^{(1)}, \ldots, x^{(l-1)}). \tag{1.57}$$

The specialization of the mechanism monitor is the *configuration monitor* that tracks the performance of a parental set in case of a specific parental configuration pa_{ij}:

$$S(\mathrm{pa}_{ij}; D_N) \quad = \quad -\log \prod_{l=1}^{N} p(x_i^{(l)} | pa_{ij}, x^{(1)}, \ldots, x^{(l-1)})^{1(\mathrm{pa}_i^{(l)} = \mathrm{pa}_{ij})}. \quad (1.58)$$

By these definitions we can rewrite the model score as the sum of the mechanism monitors or the total sum of all of the configuration monitors in M.

$$S(M; D_N) = \sum_{i=1}^{n} S(\mathrm{Pa}(X_i, G); D_N) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} S(\mathrm{pa}_{ij}; D_N). \quad (1.59)$$

The application of the model monitor, mechanism monitor and parent-child monitor in the ovarian cancer domain are reported in Section **??**.

FULLVERSION>

1.4.1 Sequential evaluation of posteriors for structural features

The prequential analysis of the Bayesian network structure is closely related to (Bayesian) sequential analysis of various posteriors related to the structure. Indeed, the batch score corresponding to a model monitor in Eq. 1.52 defined over a Bayesian network model with parameter prior is the model log-likelihood (see Eq. 1.47, 1.50). The posterior of the Bayesian network (structure) is the product of the likelihood from the model monitor and a structure prior. Especially, in case of uniform priors these are identical:

$$p(G|D_N) \propto \exp(S(G; D_N))p(G). \quad (1.60)$$

However, the relation is not that direct in general between a node monitor and the posterior of the structural features determining the node monitor. So, the sequential analysis of the posterior of a model or the posterior of structural features or ABN-propositions is a distinct evaluation methodology, which is particularly important in the case of a rich semantic context such as an ABN-KB. In general the computation of the sequence of posteriors of a structural BN feature $F(G)$ for $l = 1, \ldots, N$ requires computationally intensive methods described in Chapter 2.7 and corresponds to the task of computing the sequential posteriors for an arbitrary ABN-proposition:

$$p(\alpha(G|\mathcal{K})|D_l) = \sum_{\alpha(G|\mathcal{K}) \; is \; true} p(G|D_l) \; for \; l = 1, \ldots, N. \quad (1.61)$$

Recall, that an ABN-proposition can express structural features such as parental sets π_k and Markov blanket graphs mbg ($P(\pi|D_l)$ or $P(\mathrm{mbg}\,|D_l)$). For example, the sequence of posteriors of the following ABN-proposition is presented in Section **??**: $|\,\mathrm{Pa}(X_k, G) - \mathrm{Pa}(X_k, G^{ref_0})| < 3$.

Now we continue with the investigation of special cases, when the posteriors such as $P(\pi_k|D_l)$ has simple analytic forms directly related to the prequential score corresponding to the underlying structure. As a counter-example, in the case of mechanism monitor, the corresponding score is a partial data log-likelihood related to the parental substructure of the Bayesian network $\pi_k = \mathrm{Pa}(X_k, G_0)$, but the posterior of the Bayesian network is not decomposable to a product of the posterior of this parental substructure and the rest of the structure. So in general the mechanism monitor score is not directly related to the posterior of the parental substructure:

$$p(\pi_k|D_N) = \sum_{G\sim\pi_k} p(G|D_N) \propto \sum_{G\sim\pi_k} p(D_N|G)p(G) \tag{1.62}$$

$$= \sum_{G\sim\pi_k} p(G) \prod_{i=1}^{n}\prod_{l=1}^{N} p(x_i^{(l)} \,|\, \mathrm{pa}(X_i, G) = \mathrm{pa}_i^{(l)}, \underline{C_1}, \ldots, \underline{C_{l-1}}) \tag{1.63}$$

$$= \sum_{G\sim\pi_k} p(G) \prod_{i=1}^{n} \exp(S(\mathrm{Pa}(X_i, G); D_N)). \tag{1.64}$$

If $p(G)$ is an unnormalized modular prior (see Def. 1.1.23) over the structures compatible with a fixed ordering of the variables $\prec_0$ (as in Eq. 1.34), then the posterior of the Bayesian network can be factorized to a product of the posteriors of parental substructures and

$$p(\pi_k|D_N) \propto \sum_{G\sim\pi_k} \prod_{i=1}^{n} p(\mathrm{Pa}(X_i, G)) \exp(S(\mathrm{Pa}(X_i, G); D_N)) \tag{1.65}$$

$$= p(\pi_k) \exp(S(\pi_k, G); D_N)c, \text{where} \tag{1.66}$$

$$1/c = \sum_{P(\mathrm{Pa}(X_k,G))>0} p(\mathrm{Pa}(X_k, G)) p(D_N^{|X_k} \,|\, \mathrm{Pa}(X_k, G), D_N^{|\mathrm{Pa}(X_k,G)}).$$

The $D_N^{|bsX}$ denotes the partial data set including only the values for the variables $\underline{X}$. A special case is the set of Naive Bayesian networks with a fixed root variable and leaves with a structure prior being a product of edge probabilities (see Section **??**). Another special case when the posterior of a parental substructure $\pi_k = \mathrm{Pa}(X_k, G_0)$ is independently present in a decomposed form of the posterior of the structure, if (1) $p(G)$ is positive only for structures without outgoing edges from X_k (i.e., over structures for which there exists an ancestral/topological ordering of the variables $\prec$ with X_k is the last) and (2) $p(G)$ is decomposed as $p(G) = p(G \setminus X_k)p(\mathrm{Pa}(X_k))$ (i.e., in the Bayesian regression/classification context: the belief in the factors for the dependency variable is independent of the belief for the interactions between the factors, see Section **??**). Then

$$p(\pi_k|D_N) \propto \sum_{G\sim\pi_k} p(D_N|G)p(G) \quad (1.67)$$

$$= p(\pi_k)\exp(S(\mathrm{Pa}(X_i,G);D_N))c, \text{where} \quad (1.68)$$

$$c = \sum G \setminus X_k p(G \setminus X_k)p(D_N^{|V\setminus X_k}|G \setminus X_k). \quad (1.69)$$

Under this condition the prequential results of mechanism monitors are the log-posteriors of the corresponding parental sets with uniform prior.

1.4.2 Evaluation using informative utilities

After discussing the sequential evaluation of the logarithmic losses and the posteriors for a given model and for its structural features, now we investigate the specification of general "informative" utilities for Bayesian network structures G (see Def. **??**), which is necessary for a decision theoretic evaluation of model structures. As earlier (see Def. **??**), we consider the model selection as an action with this outcome in case of "true" G (e.g., the report of $\hat{G}$). So the goal is to specify an "informative" loss function $L(G,\hat{G})$. The term "informative" indicates the potential semantic background knowledge from the ABN-KB possibly involved in the definition, so if it is necessary to emphasize this aspect we use the term ABN-utility or ABN-loss function. We assume that a same ABN-knowledge base and the standard textual and numeric functions are used in the composition (see Def. **??**).

$$U(G,\hat{G}|ABN-KB) : \{\mathcal{G}\times\mathcal{G}\} \to \mathcal{R} \quad (1.70)$$

In fact the ABN-utility function $L(G,\hat{G}|ABN-KB)$ can be seen as the generalization of ABN sentences $\alpha(G) : \mathcal{G} \to true/false$).

The simplest use of the loss function is if we have a gold standard G_0 and a goodness/penalty score can be computed for a particular model $\hat{G}$ as $L(G_0,\hat{G})$. A related application in evaluating and comparing priors is to qualify the posterior (so the priors) by their corresponding expected losses $\mathrm{E}_{p(G|D_l)}[L(G_0,G)]$.

Without a gold standard model the loss function $L(G,\hat{G})$ can be used in a standard decision theoretic framework to evaluate a particular model $\hat{G}$ by its expected loss $\mathrm{E}_{p(G|D_l)}[L(G,\hat{G})]$, possibly sequentially for $l = 1,\ldots,N$. If only the evolution of the consequence of the best action (i.e., the minimal expected loss) is interesting, a learning curve can be constructed by performing optimal model selection sequentially for the data set D_l and indicate

$$L^*_{D_l} = min_{\hat{G}}\,\mathrm{E}_{p(G|D_l)}[L(G,\hat{G})] \propto \sum_G \underbrace{L(G,\hat{G})p(G)}\, p(D_l|G). \quad (1.71)$$

We indicated again the complementarity of the prior and the utility. Now we discuss forms and ingredients of the loss function $L(G,\hat{G}|ABN-KB)$. A

general form is similar to the combination of deviation-based and feature-based structure priors (see Section 1.1.5.2).

$$L(G, \hat{G}) = \sum_{i=1}^{K} w_i \alpha_i(\hat{G}|G, ABN - KB), \tag{1.72}$$

where $\alpha_i(\hat{G}|G, ABN-KB)$ for $i = 1, \ldots, K$ denote arbitrary ABN-propositions over the knowledge base $\{ABN - KB, G\}$ including the structure G as a reference structure and w_i are their weights. This form satisfies a property called preferential independence between the propositions (features), that is independent penalties for the propositions in the loss function. Because these features are usually logically dependent, the effect of this possible inconsistency should be considered in evaluating such loss functions.

In its most general form an ABN-proposition can express arbitrary well-defined semantic property for $\hat{G}$ w.r.t. G and at the other extreme it can represents an elementary structural difference between G and $\hat{G}$. For example the ABN-KB in the ovarian cancer domain includes four-graded rating for the pairwise dependency relations and for the causal mechanisms (i.e., for the parental sets), see Chapter **??** for details. The represented structural differences can be local features such as the different status of an edge between X_i and X_j in G and in $\hat{G}$ or global such as the different pairwise causal relation between X_i and X_j induced by G and by $\hat{G}$. The combination of these leads to features, such as the following:

$$\begin{aligned} \alpha(\hat{G}|G_0, ABN - KB) = \quad & CaE(X_i, X_j|G_0) \\ & \wedge(Rate(UndirectedEdge(X_i, X_j) >=' medium')) \\ & \wedge Independent(X_j, X_i|\hat{G}) \end{aligned}$$

The application of the loss function with a reference (gold standard) is frequently complicated by the following two issues, which also present in the thesis. First, there can be multiple references as in our case we have three embedded reference structures (see Section **??**). Second, if the propositions do not include explicitly a reference structure and they express only desirable properties of $\hat{G}$, they can be inconsistent in the sense that there is no reference structure satisfying all the desiderata.

1.4.3 Evaluation using reference structure and structural features

As mentioned above, the loss function $L(G, \hat{G})$ with a reference structure G_0 can be used to evaluate and compare posterior (so the priors) based on their expected losses $\mathrm{E}_{p(\hat{G}|D_l)}[L(G_0, \hat{G})]$. This corresponds to the situation of assuming random model selection according to the posterior over G given a data set D_l (in a similar vein as in the Gibbs algorithm in [71]). If the loss function is specified in a decomposed form of Eq. 1.72 by the propositions $\alpha_i(\hat{G}|G_0)$ and

their weightw_i, then a reported $\hat{G}$ determines jointly the truth-values of the propositions denoted by $\alpha_i(\hat{G}|G_0)$.

Another approach can be gained for the numeric qualification of the posterior using a reference structure G_0 if we consider $p(\hat{G})$ as a probabilistic knowledge base, which induces distributions for a given set of ABN-propositions $\alpha_i(\hat{G}|G_0)$ with 2×2 loss matrices W_i. These should not be independent or completely determine a structure. In this decision theoretic situation for each proposition $\alpha_i(\hat{G}|G_0)$ an optimal, minimal loss decision (truth-value) $\alpha_i(.|G_0)$ can be determined independently according to its induced probability $p(\hat{G} : \alpha_i(\hat{G}|G_0))$ and the corresponding loss matrix W_i. Note the difference between the earlier situation of complete model selection and this situation of using $p(G)$ in the Bayesian framework for predicting binary functions. For example as neither independence of the features nor the consistency of the values (i.e., DAG property) is required it is possible that there is no $\hat{G}$ that the jointly generated $\alpha(\hat{G}|G_0)$ is equal to the individually generated $\alpha(|G_0)$.Now the total, minimal loss can be used to numerically qualify the posterior belief $p(\hat{G})$, which is determined by the reported values $\alpha_i(.|G_0)$, the reference values $\alpha_i(G_0|G_0)$ and the loss matrices W_i.

Further generalization can be reached if $p(\hat{G})$ is evaluated with a "loss-free" method without the loss matrices W_i using only the induced probabilities $p(\hat{G} : \alpha_i(\hat{G}|G_0))$ and the reference values $\alpha_i(G_0|G_0)$. The method is based on the idea of interpreting $p(\hat{G})$ as a binary classifier and evaluate its performance on the fixed set of (binary) propositions. First, predicted binary values are determined with a threshold τ from the probabilities for the propositions

$$\alpha_i(\tau) = 1(\tau < p(\hat{G} : \alpha_i(\hat{G}|G_0))) \tag{1.73}$$

Then the quality of $p(\hat{G})$ can be characterized w.r.t. the given propositions and their reference values by the sensitivity and 1-specificity pairs for well-selected values of τ (see Section **??** for the definition of sensitivity and specificity). Another option is to use the so-called Recevier Operating Characteristic curve (ROC curve) that plots these pairs for a changing τ in the $[0, 1]$ interval. Finally a single scalar the so-called Area Under the Curve (AUC) value can be gained by integrating the ROC curve.

The Chapter 2 discusses the computation of the probabilities of ABN-propositions. In Chapter **??** sensitivity and specificity pairs, ROC curves and AUC values are reported in the ovarian cancer domain for various propositions such edge presence and Markov blanket membership in the model and related to a central node.

1.4.4 Evaluation using reference posterior and ranks for structural features

Finally, we discuss a method to evaluate a structure posterior against reference ranks and scalar scores for structural features. That is we assume that for one or more multivalued structural features $F_i(G)$ with values f_{ij} there exist referential

ranks and scalar scores $Rank_0(fi)$ and $Score_0(fi)$. Such features for a variable X_i are the parental relation ($CaE(X_i, X_j, G)$), the parental set ($\mathrm{Pa}(X_i, G)$), the Markov blanket membership ($\mathrm{MBM}(X_i, X_j, G)$) and the Markov blanket set ($\mathrm{MB}(X_i, G)$). Because the posterior probability $p(G)$ induces a distribution for the structural features $P(F_i(G))$ and a corresponding $Rank(F_i)$, the reference can defined by a posterior with a reference (possibly uniform) prior, so this technique as the previous can be used to compare directly posteriors (so priors), not only a data based posterior against expert's reference.

If the reference scores are probabilities, then standard distance measures such as L_1 or the Kullback-Leibler semi-distance can be used to quantify the closeness of the posterior probability $p(G)$ to the reference in terms of the marginals $P(F_i(G))$.

If only scalar scores and ranks are available, then scatter plots can be used for the manual comparison (see Section **??** for such results). Because the assumption of linear relation is usually inadequate (i.e., tests on the Pearson correlation coefficient), we investigated the more robust hypothesis of the existence of a monotonic relation using the Spearman rank correlation coefficient r_S. It can be defined equivalently with the following more succinct form

$$r_S(F_i) \quad = \quad 1 - 6\frac{\sum_{j=1}^{\#(F_i)} (Rank_0(f_{ij}) - Rank(f_{ij}))^2}{K(K^2-1)}. \tag{1.74}$$

Additionally, we report a special rank-correlation measure penalizing only the 25% differences of ranks defined as follows. Define a matrix R_k in which the a_{ij} element is the number of times the feature values f_k have reference rank i and data rank j. Now define a matrix R'_k which is the $4 - by - 4$ partitioning of R with the following intuitive interpretation for the four partitions: highly relevant, moderately relevant, less relevant, and not relevant. We report the normalized trace of R', that is the correspondence between the ranks using this 4-graded granularity.

<FULLVERSION

1.5 Learning Bayesian networks

By now we summarized a framework for general, normative, inductive inferences using probabilistic domain models: the Bayesian decision-theoretic framework with Bayesian networks. FULLVERSION> This includes the analysis of (1) parameter and structure posteriors, (2) posteriors of structural features, (3) posteriors of ABN-propositions or (4) the expected loss of related actions. <FULLVERSION Frequently, it is restricted to optimization, particularly over structures, which is termed the "standard" Bayesian network (structure) learning, not necessarily within the Bayesian decision theoretic framework. This mode of operation is particularly relevant if a large amount of data is available w.r.t. the complexity of the model. So, in this section we finish our overview

with the summary of the score-based learning of Bayesian networks, including Bayesian and non-Bayesian inductive scores and search algorithms.

Another large family of methods for finding complete models best fitting the observations are the constraint-based algorithms. These construct a network by performing independence tests with certain prespecified significance level, which is an NP-hard task (see Th. 1.5.4). Assuming no hidden variables, a stable distribution and correct hypothesis tests, the Inductive Causation (IC) algorithm correctly identifies a Bayesian network that exactly represents the independencies (see [116, 66, 134]). It means that the score-based and the constraint-based learning algorithms behave identically for stable distributions in the limit w.r.t. the sample size(see Th. 1.5.3). However, there is no generally recommendable prespecified significance level and final significance level for the identified model. Furthermore, because of the frequentist approach, there is no principled way to incorporate uncertain prior information. On the other hand, efficient constraint-based algorithms exist that work in the presence of hidden variables, which is currently not tractable with Bayesian methods.

Our assumption of complete, observational and discrete data modeled with a fixed set of discrete variables is a serious restriction, but it provides a sufficient conceptual framework to develop the main topics in the thesis such as the (automated) construction of priors, the computation of posteriors of complex structural features and their role in classification. We direct the reader to the following sources regarding the treatment of mixture of discrete and continuous variables [98, 33, 74]; the mixture of observational and interventional data [66]; the issue of incomplete data[64, 50]; the issue of special local probabilistic dependency models [52] and the issue of temporal data and variables [126].

1.5.1 Score functions and their properties

The score-based learning of Bayesian networks best fitting to the data D_N consists of the definition of a score function $S(G, D_N) : \{G \times D_N\} \to \mathcal{R}$ and a search method in the space of DAGs. In a Bayesian decision theoretic framework the score function is specified as the expected loss $\mathrm{E}_{P(\hat{G}|D)}[L(G, \hat{G})]$ of selecting (i.e., reporting) the structure $\hat{G}$. Whereas the advantages of knowledge rich utility functions are apparent, standard score functions lack domain knowledge. For example, in case of 0-1 utility function the model with maximum expected utility corresponds to the structure with *maximum a posteriori* probability or in case of uniform prior to finding the *maximum likelihood* structure.:

$$G^{\mathrm{MAP}} = \arg\max_{\hat{G}} \mathrm{E}_{p(G|D)}[L(G, \hat{G})] = \arg\max_{\hat{G}} p(\hat{G}|D), \text{ if } L(G, \hat{G}) = 1(G = \hat{G}). \tag{1.75}$$

In Eq. 1.51 we derived a closed form for the posterior of a structure G,

$$p(G, D_N) = p(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})}, \tag{1.76}$$

termed *Bayesian Dirichlet* metric [74]. If the initial hyperparameters $\underline{\alpha}$ satisfy the conditions of Th. 1.1.6 (ensuring indistinguishability within an equivalence class), then it is denoted as BD_e. If the initial hyperparameters $\underline{\alpha}$ are constant 1 then it is denoted by $\mathrm{BD}_{\mathrm{CH}}$ [29]. If the initial hyperparameters are the converse of the number of parameters corresponding to the local, overall multinomial models of the variables then it is denoted by $\mathrm{BD}_{\mathrm{eu}}$ [17, 74]. The corresponding score functions are defined as $\mathrm{BD}(G; D_N) = \log(p(G, D_N))$.

Another family of non-Bayesian score functions can be derived within the likelihood framework. The maximum likelihood score is defined as follows

$$\mathrm{ML}(G; D_N) = \max_{\underline{\theta}} p(D_N | G, \underline{\theta}). \tag{1.77}$$

Assuming a complete, discrete value, i.i.d. data set, it can be shown that this is maximized by the selection of $\theta^*_{ijk} = N_{ijk}/N_{ij+}$, where N_{ijk} are the occurrences of value x_k and parental configuration q_j for variable X_i and its parental set $\mathrm{pa}(X_i)$ (N_{ij+} is the appropriate sum) [57, 127]. By substituting this maximum likelihood parameter selection, we get

$$\begin{aligned} \mathrm{ML}(G; D_N) &= p(D_N | G, \underline{\theta}^*) = \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)} | pa_i^{(l)}) && (1.78) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \left(\frac{N_{ijk}}{N_{ij+}} \right)^{N_{ijk}}, && (1.79) \end{aligned}$$

by taking logarithm, rearranging and expanding with N

$$\log(\mathrm{ML}(G; D_N)) = N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \left(\frac{N_{ijk}}{N_{ij+}} \right). \tag{1.80}$$

Using the definition of conditional entropy $H(Y|X) = \sum_x p(x) \sum_y p(y|x) \log(p(y|x))$, the chain rule $H(X, Y) = H(Y|X) + H(X)$ and the definition of mutual information $I(Y; X) = H(Y) - H(Y|X)$ [32], it can be rewritten as [127]

$$\begin{aligned} \log(\mathrm{ML}(G; D_N)) &= -N \sum_{i=1}^{n} H(X_i | \mathrm{Pa}(X_i, G)) && (1.81) \\ &= -N H(X_1, \ldots, X_n) && (1.82) \\ &= N \sum_{i=1}^{n} I(X_i; \mathrm{Pa}(X_i, G)) - N \sum_{i=1}^{n} H(X_i) && (1.83) \end{aligned}$$

This shows that the maximization of the maximum likelihood score is equivalent to finding a Bayesian network parameterized with the observed frequencies that has maximum mutual information between its children and their parents (the terms not depending on the structure can be neglected in Eq. 1.83). Note the close connection of this interpretation to the concept that causal ordering is related to the determination of each variable by the earlier variables [44].

Because of the monotonicity of mutual information — if $\mathrm{Pa}(X_i) \subset \mathrm{Pa}(X_i)'$, then $I(X_i; \mathrm{Pa}(X_i)) \leq I(X_i; Pa'(X_i))$ [32] — the complete network maximizes the maximum likelihood score. However, score functions such as the MDL-score derived from the minimum description length (MDL) principle or the Bayesian information criterion (BIC)-score derived with a non-informative Bayesian approach contains various complexity penalty terms. We used only the following MDL/BIC-score defined as follows

$$\mathrm{BIC}(G; D_N) = \log(\mathrm{ML}(G; D_N)) - \frac{1}{2}\dim(G)\log(N), \tag{1.84}$$

where $dim(G)$ denotes the number of free parameters. For overviews of other score functions and for the derivation of the BIC-score, see [94, 15, 25, 57, 75]. We discuss now the properties scoring metrics w.r.t. observational equivalence and sample size.

Definition 1.5.1. *A score function $S(G; D_N)$ is called score equivalent, if for each pair of observationally equivalent Bayesian network structure G_1, G_2 the scores are equal $S(G_1; D_N) = S(G_2; D_N)$ for all D_N [74].*

Theorem 1.5.1 ([74])**.** *The $BD_e(G; D_N)$ scoring metric is likelihood equivalent, that is if G_1, G_2 are observational equivalent, then $p(D_N|G_1) = p(D_N|G_2)$. Furthermore, if the structure prior is acausal (i.e., equal for such G_1, G_2), then the BD_e scoring metric is score equivalent [74].*

Consequently, the score can be used directly in an acausal approach if the hypotheses are the observational equivalence classes. In a causal approach to Bayesian network structure learning with the BD metrics the structure prior can incorporate information differentiating observationally equivalent structures, which means an asymptotically vanishing term w.r.t. the likelihood term. The differentiation within an equivalence class by a non-likelihood equivalent BD score (i.e., by a non-likelihood equivalent parameter prior such as the $\mathrm{BD_{CH}}$) is similarly vanishing.

The score equivalence of the BIC score is the direct consequence of the result that the number of free parameters (i.e., the term $dim(G)$) are equal in observationally equivalent Bayesian networks (here again as throughout the thesis, we assume discrete variables and multinomial local dependency models) [15, 25, 24].

Theorem 1.5.2 ([24])**.** *The $BIC(G; D_N)$ scoring metric is score equivalent.*

The next theorem, due to Bouckaert [15], ensures the asymptotic consistency of an idealized (algorithm-free) score based Bayesian network learning that always returns $\arg\max_G S(G, D_N)$.

Theorem 1.5.3 ([15])**.** *Let $\underline{V}$ be a set of variables. Let the prior distribution $p(G)$ over Bayesian network structures be strictly positive. Let $p(\underline{V})$ be a positive and stable distribution and G_0 is a corresponding perfect map (i.e., a Bayesian network representing exactly the independencies). Now, let D_N be an i.i.d. data*

set generated from $p(\underline{V})$*. Then, for any network structure* G *over* $\underline{V}$ *that is not a perfect map of* $p(\underline{V})$ *we have that*

$$\lim_{N\to\infty} \mathrm{BD}_e(G_0; D_N) - \mathrm{BD}_e(G; D_N) = -\infty \text{ and also} \quad (1.85)$$

$$\lim_{N\to\infty} \mathrm{BIC}_e(G_0; D_N) - \mathrm{BIC}_e(G; D_N) = -\infty. \quad (1.86)$$

For further results about the asymptotic optimality of the scores for not stable distributions, see [15]. Results about asymptotic consistency and rate of convergence results for maximum likelihood scores are derived in [15, 57]. Furthermore, a rate of convergence is also derived and a corresponding *sample complexity* $N(\epsilon, \delta)$ to select an appropriate sample size for a given accuracy between the target distribution p_0 and the distribution p_{BN} represented by the learned Bayesian network with a given confidence

$$p(D_N : \mathrm{KL}(p_0|p_{BN}) > \epsilon) < \delta. \quad (1.87)$$

Unfortunately, the tightness of this bound depends strongly on the properties of the target distribution and it scores the parametric closeness and not the structural one. For the sample complexity of parameter learning, see [35].

Another bad news is that in general different structures maximize the BD scores and BIC scores for a finite sample size, despite the asymptotic equivalence of the scores (for a related result about their certain equivalence under specific circumstances see [15]). For example it can be shown that the BIC metric is never maximal for structures with larger parental set size than $\log(\frac{2N}{\log N} + 1)$, on the contrary a data set D_N can be constructed for which the $\mathrm{BD}_{\mathrm{CH}}$ metric is maximal for a structure containing a parental set with size $N/2$ [15]. As discussed by Bouckaert, this difference is partly the consequence of the fact that the BD scores penalize only the instantiated parental configurations, whereas the BIC score has a fixed complexity penalty term, preferring sparser networks. A related anomaly of the BD_e score is that it includes mostly constant parents (usually occurring in case of small data sets).

1.5.2 Search algorithms for finding high-scoring BNs

As discussed in the beginning of this section, the recently used loss functions or more generally the score functions $S(G, D_N)$ are usually efficiently computable in $\mathcal{O}(nN)$. It is partly the consequence of the decomposability of the score, which allows even further computational speed-ups as discussed later on. However, the global DAG constraint does not allow the decomposition, so we have to perform a combinatorial optimization in the space of DAGs over n nodes (variables). The cardinality of the space of DAGs is given by a recursion [29].

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} f(n-i) \text{ with } f(0) = 1. \quad (1.88)$$

By neglecting the DAG-constraint, this can be bounded by the number of the combinations of the edges between different nodes ($2^{n(n-1)}$). By limiting the

maximum number of parents to k it is still super-exponential (consider that the number of parental sets for a given ordering of the variables is in the order of n^{kn}, so $2^{\mathcal{O}(kn \log n)}$ [54]).

The computational complexity of learning BNs in the constraint-based and in the score-based framework is bounded by the following two theorems (assuming $P \neq NP$). The first states the NP-hardness of finding a Bayesian network for the observations (as minimal representation of the observed independencies, see Def. 1.1.12) [15].

Theorem 1.5.4 ([15]). *Let $\underline{V}$ be a set of variables with joint distribution $p(\underline{V})$. Assume that an oracle is available that reveals in $\mathcal{O}(1)$ time whether an independence statement holds in p. Let $0 < k \leq |\underline{V}|$ and $s = \frac{1}{2}n(n-1) - \frac{1}{2}k(k-1)$. Then, the problem of deciding whether or not there is a (non-minimal) Bayesian network that represents p with at most s edges by consulting the oracle is NP-complete.*

The second theorem states the NP-hardness of finding a best scoring Bayesian network (i.e., the NP-hardness of optimization over DAGs) [25].

Theorem 1.5.5 ([25]). *Let $\underline{V}$ be a set of variables, D_N is a complete data set, and $S(G, D_N)$ is a score function. Then, it is NP-complete to decide whether or not there exists such a Bayesian network structure G_0 defined over the variables $\underline{V}$ that each node in G_0 has at most $1 < k$ parents and $c \leq S(G_0, D_N)$, where $c \in \mathcal{R}$.*

In the special case of $k = 1$ (that is for trees and polytrees) standard maximum weight spanning tree (MWST) construction algorithms can be applied, which has polynomial time complexity, see [114, 25]. The NP-hard nature of the problem remains if the learning takes place over the smaller space of equivalence classes [25, 85].

Consequently, a frequently used suboptimal approach is to use iterative improvement algorithms with local search. These start from a good or at least a neutral candidate satisfying the prior knowledge and the DAG constraint. In each step i a structure with an improved score is selected from the prespecified neighborhood $Nb(G_i)$ of G_i, otherwise the algorithm is stopped. Usually this neighborhood is defined as structures with 1 edge difference. However, the result of the iterative improvement algorithms with local search is probably a local optimum, so frequently the algorithms are restarted with a random initial candidate. This problem can be avoided by replacing the greedy element of the algorithm with a stochastic scheme allowing selections of structures with worse score, as in the simulated annealing algorithm. A greedy algorithm called $K2$ can be applied if the score is decomposed and the ordering of the variables are well-restricted, because for each ordering the parental sets can be optimized independently with a greedy algorithm [29]. We shall use two minor extensions of the $K2$ algorithm that arise if only a partial ordering is available. The $K2cyc$ denotes the systematic application of the $K2$ for each $1 \leq nth$ allowed permutation using an alphabetic ordering of the permutations. The $K2rnd$ denotes

the application of $K2$ for randomly drawn admissible orderings. Studies of the performance of various iterative improvement algorithms using local search and simulated annealing are reported in [25, 15], which indicate a robustly good performance with relatively low computational complexity for the $K2$ algorithm without tuning to the domain, data set, etc. Our experiments in the ovarian cancer domain with various iterative improvement algorithms with local search and simulated annealing algorithm similarly strengthened this result. In the thesis the reported results are usually computed with a $K2$ variant algorithm using the implementational tricks of the sample tree to compute the score for a parental set in $\mathcal{O}(N)$ as proposed in [29] and storing the parental scores as also proposed in [17].

Chapter 2

Inference over BN features

RELEVANT>

First we categorize structural properties (i.e., features) of Bayesian networks, and introduce a feature called Markov blanket graph. Second we summarize the advantages of the Bayesian approach to BN features, and formalize the applicability and the statistical advantages of the ordering-based MCMC estimation method. Third we discuss the consequences of the exponential cardinality of feature values for decisions based on their MC estimates. Finally, the integration of estimation and search of high-scoring MBG feature values is analyzed.

<RELEVANT

The increasing complexity of the models, the incorporated prior knowledge and the queries leads to the issue of Bayesian inference over general properties of Bayesian networks (i.e., to estimation of the expectation of binary random variables). Although we discuss this problem from the point of inference over structural features, note that the expectation of functions over the space of DAGs w.r.t. a posterior appears in a wide range of problems, such as in the posterior of a feature (i.e., structural model property) F_c, in the posterior of an ABN sentence (see Def. **??**), in the expected loss of the selection of a given model and in the full-scale Bayesian inference over domain values (see Eq. 1.33):

$$p(F_c = f_c|D_N) = \sum_G 1(F_c(G) = f_c)p(G|D_N) \tag{2.1}$$

$$p(\alpha(G)|\mathcal{K}, D_N) = \sum_{M(G)\in\mathcal{M}(\mathcal{K})} \alpha(M(G))p(G|D_N) \tag{2.2}$$

$$L_{\hat{G}|D_N} = \mathrm{E}_{p(G|D_N)}[L(G,\hat{G})] = \sum_G L(G,\hat{G})p(G|D_N), \tag{2.3}$$

$$p(\underline{y}|\underline{x}, D_N) = \mathrm{E}_{p(G|D_N)}[\mathrm{E}_{p(\underline{\Theta}|G,D_N)}[p(\underline{y}|\underline{x}, \underline{\Theta}, G)]]. \tag{2.4}$$

First, we overview Bayesian network features in Section 2.1 and introduce the Markov blanket subgraph feature in Section 2.2. In Section 2.3 and 2.4 we discuss the advantages of feature posteriors as confidence measures w.r.t. the

bootstrap probabilities. In Section 2.5 we will concentrate on the approximation of Eq. 2.1, when the feature is a standard graph-theoretic property of DAG G with values $F(G) = f_i$, $i = 1, \ldots K$. The growing importance of such model-based, feature-oriented statistical inferences is the result of (1) frequent high sample complexity for the identification of the complete model, (2) the lack of prior for the complete model, (3) the high computational complexity for the identification of the complete model, (4) the availability of computational resources and stochastic methods for estimation, and (5) the availability of complex semantic propositions with statistical semantics as the ABN sentences in Eq. 2.2.

The most important factor is the relatively small amount of data. A general expectation is that, in case of small amount of data, at least certain properties with high significance of a complex model can be inferred and perhaps with lower computational cost. So the goal is the automated learning of what is learnable with high confidence in the considered model space given the data and to support the interpretation of statistical inference by indicating confidence measures for such properties. Furthermore, the model properties with high significance can be used heuristically as "hard" constraints or "soft" bias to support the inference of the complete model, either by influencing it through priors in learning from heterogeneous sources or in the case of using the same data set by influencing the optimization process itself (see Chapter **??**). Note the similarity of this approach to the frequentist constraint-based Bayesian network learning methods, which perform hypothesis tests on local model properties (on features) and integrate them into a consistent domain model. In a potential Bayesian analog the hypothesis tests are replaced by the model-based feature posteriors instead of the significance levels and p-values of hypothesis tests, enhancing their integration in subsequent phases of learning a complete domain model.

However, the Bayesian approach to feature learning has many additional aspects beside the estimation of the posterior. Such related issues are the effect of the cardinality of feature values on the selection of optimal value(s) and the integration of estimation and search processes in case of high numbers of features, which are discussed in Section 2.6 and 2.7. Additional issues related to classification in our case are the support of full scale Bayesian inference over domain values (i.e., the use of the estimated posterior distribution over the features as a probabilistic knowledge base) and the transformation or inducement of priors for a subsequent learning phase either using Bayesian networks or using other more specialized representations, for example logistic regression or multilayer perceptrons. These are discussed in Chapter **??**.

Whereas these inferences are investigated mainly in fundamental research, they may soon appear in standard statistical data analysis software and in decision support systems as they can offer a more personalized and knowledge intensive environment for inductive inferences. For example, the combination of the electronic clinical and genomic patient data, the semantic web and evidence-based medicine can be driving force for such complex probabilistic queries over standardized knowledge bases and data-bases. A special case is the area of

statistical analysis of biomedical literature, where we can treat the domain literature as a special data set and formulate queries against this voluminous knowledge base (see Chapter **??**). In general, it means that the knowledge intensive Bayesian approach over large, distributed knowledge and data-bases will get more and more emphasis within the area of knowledge and data analysis.

FULLVERSION>

The central topic of this chapter is the computation or approximation of the posterior of such feature oriented ABN-proposition in Eq. 2.2 or the induced posterior distribution of the feature $p(F(G))$ or its maximum a posteriori feature value

$$f^{\mathrm{MAP}} = \arg\max_f p(G : F(G) = f | D_N). \tag{2.5}$$

Note that this corresponds to the 0-1 utility function over the feature values and that in general the value f^{MAP} is not equal to the value of the feature in the maximum a posteriori model $F(G^{\mathrm{MAP}})$ (see Eq. 1.75).

<FULLVERSION

2.1 Bayesian network features

Before considering the induction of confidence measures over a Bayesian network feature F, first we overview standard Bayesian network features, together with proposed identification methods and the corresponding Bayesian tasks.

There is a large variety of features (i.e., model properties) to provide an overall or specialized characterization of the underlying model, such as the undirected edges or compelled edges (as direct relations or direct causal relations under CMA), pairwise or partial ancestral ordering (related to causal ordering), the parental sets, the pairwise relevance relations, the subset relevance relations (Markov blankets) or the partially parametric features such as the pairwise qualitative features. Despite this variety and the presence of the parental set features, which are the ultimate building blocks of Bayesian networks, the usefulness of these features are still seriously restricted by their unexplored dependency in all application areas, such as in data analysis, in probabilistic knowledge bases, in prior acquisition and in posterior-to-prior inducement for later phases of Bayesian learning. This seems to be unavoidable because even small sets of simple local features quickly become dependent, because of the DAG constraint, what biases this model-based approach with hardly estimatable effects.

A possible solution is the definition of complex features (subtheories) that are *sufficient* features for a given aspect of the domain theory and still more efficiently learnable than the complete domain model. So, it is an open issue to define complex features that on the one hand exactly model a semantically interesting fragment (subtheory) of the domain and on the other hand they are still considerable simpler than the complete domain model. Such a feature would

exactly represent the interesting dependencies between the relevant simpler features and the statistical and computational complexity of the estimation of its distribution over the feature space would be lower and better interpretable.

In fact, we can define two approaches to Bayesian network features. The first approach relies on the assumption that the feature set is fixed, the features are significantly simpler than the complete domain model, though they provide an overall characterization as a fragmentary representation, and the number of features and feature values are tractable (not exponential, but linear or quadratic in the number of variables). Such features are the pairwise edge or relevance features (i.e., the compelled edges and Markov blanket relations). These simple features are easily interpretable or can be used to support a subsequent learning phase of a complete Bayesian network model. The main challenge in this approach is the computation of the corresponding expectations.

At the other extreme of feature learning we find the identification of arbitrary subgraphs with statistical significance, which is an idealistically autonomous approach to feature learning consisting of a mixture of search and the computation of the achieved significance. This is close to our approach to Bayesian network features investigated in the thesis, but we restrict the subgraphs to Markov blanket subgraphs to have a focused representation from a single, but complex point of view (i.e., from conditional modeling) and we use the Bayesian framework instead of the frequentist framework.

2.1.1 Edges: direct pairwise dependencies

The first family of frequentist algorithms for learning a Bayesian network feature targets the identification of "direct" (unconditional) causal pairwise relations ("direct" in the sense discussed in Section 1.1.3.2). If the hypotheses are the DAGs as causal models, then this feature corresponds to the edges. If the hypotheses are the observational equivalence classes as independence models, then such relations are exactly identified by the compelled edges assuming no hidden variables, the causal Markov condition and stability. The corresponding posteriors in the Bayesian context are the following

$$p(X_i \to_G X_j | D_N) = \sum_G 1(X_i \to_G X_j) p(G|D_N) \tag{2.6}$$

$$p(CompE(X_i, X_j|G)|D_N) = \sum_G CompE(X_i, X_j|G) p(G|D_N). \tag{2.7}$$

In the presence of possible hidden variables there are more advanced constraint-based algorithms for identifying relations with various causal interpretations, though not in the Bayesian framework (see [116, 66], [26, 129]). FULLVERSION> Interestingly, the starting point for these algorithms shown in Example 1.1.3 can be used autonomously for the identification of "direct" causal pairwise relations requiring only limited background knowledge (exogenous variables) and four local independency tests [26]. Despite its incompleteness, its low computational complexity ($\mathcal{O}(n^2??)$) and asymptotic correctness makes this method attractive,

particularly for large data sets such in case of text-mining, [133, 105, 106]. The data-mining application of a related local algorithm for identifying potential v-structures is reported [129]. <FULLVERSION For the application of bootstrap and Bayesian method over edge features, see Section 2.3 and 2.5.2.3.

2.1.2 Ordering of the variables

FULLVERSION> In a non-interventionist approach the causal ordering was defined as an ordering that allows incremental determination of the variables (i.e., the incremental solvability of system equations), see [44]. In its stochastic counterpart a suggestion for "statistical time", that is temporal ordering, was a partial ordering compatible with the essential graph (but see also the principles of causality on p. 5).

<FULLVERSION

Whereas the identification of the ordering of the variables rarely appears as a direct target, indirectly it is usually present in BN learning. In the acausal approach the identification of an acausal Bayesian network heavily influenced by the identification of a good ordering of the variables, because the learning of an acausal Bayesian network structure for a given ordering is computationally efficiently doable (both in the frequentist or Bayesian framework). In the causal approach when the hypotheses are the DAGs, the causal structures directly define causal orderings as ancestral orderings. Consequently a score for a Bayesian network G can be interpreted as an approximate scores for the underlying partial orderings. Recall that the ML structure score can be interpreted as the summed mutual information between the parents-child pairs and that the BD and the BIC scores are asymptotically equivalent (see Section 1.5.1). So, in a broad sense, any structure learning can be interpreted as an indirect learning of orderings, but certain algorithms explicitly use orderings as a central representation. For example, the use of genetic algorithms has been reported to find the best ordering for the learning of Bayesian network structures [95]. The corresponding posterior over the complete orderings $\prec$ in the Bayesian context is the following

$$p(\prec |D_N) \quad = \quad \sum_G 1(G \in \mathcal{G}^{\prec}) p(G|D_N). \tag{2.8}$$

2.1.3 Relevant variables

The concept of relevance is a fundamental concept in the definitions of the Bayesian network representation (see Def. 1.1 and 1.4 for the observational and causal relevance), but it is also central to AI, to decision theory (e.g., the value of further information) and to statistics (for an overview, see [136]). An important special case is the relevance of explanatory variables to predict a target variable given a data set, hopefully with a domain-specific interpretation. The selection of the relevant variables in this context is called the *feature subset*

selection (FSS) problem, which is part of the broader problem of input preprocessing, construction of variables (e.g., interaction terms) and dimensionality reduction. We will discuss only the relation of the FSS problem to BN feature learning. Note that even in the conditional approach in general the features are not independent, so the concept of relevance corresponds to the subsets and not to the individual features.

To explain the generality of the Bayesian approach to relevance using Bayesian network features, we summarize the most widespread conditional approaches to FSS in sequence (see Section **??** for the conditional Bayesian modeling). We start with the concept of relevance and with a non-Bayesian approach specific to the applied optimization algorithm, the data set, the model class, and the loss function. Then we generalize these specifics step by step, which leads to a standard conditional probabilistic concept of relevance in the end. Finally, we relate the Bayesian conditional approach to the general Bayesian approach, particularly to the Bayesian inference over Bayesian network features. In short, we show that the Bayesian inference over Bayesian network features offers an algorithm-free, model-free*, loss-free and non-conditional (i.e., domain model based) solution for the feature subset selection problem.

RELEVANT>

The *conditional approach* to FSS relies on the separate modeling of the dependence of a target variable Y on $\underline{X}'$ (i.e., without modeling the overall domain). It has been investigated using various conditional model classes M, such as linear regression, decision trees, logistic regression, multilayer perceptrons or support vector machines [78, 14, 70, 42]. It defines a score function $S^S(\underline{X}', D_N, M, L)$ for the subsets $\underline{X}' \subseteq \underline{X}$ and performs a search in the space of subsets of the features.

The wrapper approach to feature selection uses an optimization algorithm $\hat{f}_{\mathcal{C}}(\underline{X}') = \mathcal{C}(\underline{X}', D_N, M, L)$ [82, 86]. It defines the score function as

$$S^S(\underline{X}', D_N, M, L) = S^F(\hat{f}_{\mathcal{C}}(\underline{X}'), D_N, M, L).$$

The conditional model score $S^F(\hat{f}_{\mathcal{C}}(\underline{X}'), D_N, M, L)$ may incorporate factors for the interpretability or complexity of the conditional models $f(\underline{X}') \in M^{\underline{X}'}$ and their estimated expected predictive loss (risk).

In an algorithm-free and asymptotic case the subset score $S^S(\underline{X}', M, L)$ can be defined as the best expected predictive loss in a conditional model class $M^{\underline{X}'}$ with features $\underline{X}'$

$$S^S(\underline{X}', M, L) = \arg \min_{f(\underline{X}') \in M^{\underline{X}'}} \int L(y, f(\underline{x}')) p(y|\underline{x}') \, \mathrm{d}y p(\underline{x}') \, \mathrm{d}\underline{x}'. \tag{2.9}$$

However, this asymptotic and algorithm-free optimality of a subset for a given model class is not appropriate to define the relevance of a feature, as it was demonstrated in [82, 86].

*In the assumed case of discrete variables with multinomial conditionals.

The model-free subset score $S^S(\underline{X}', L)$ can be defined as the best achievable risk with subset $\underline{X}'$ for a given loss L, called Bayes risk

$$R_L^* = \int L(y, g^*(\underline{x}'))p(y|\underline{x}')\,\mathrm{d}yp(\underline{x}')\,\mathrm{d}\underline{x}', \tag{2.10}$$

where g^* is the Bayes decision, which minimizes the expected loss of prediction for each x (see Section **??**).

Because of the specific choice of the loss function $L(Y, \hat{Y})$, it is still possible that the minimal subset would miss certain features relevant for another loss. The following theorem for the case of binary output Y shows that the final loss-free generalization of the concept of relevance necessarily leads to the standard conditional probabilistic definition of relevance [42].

Theorem 2.1.1 ([42]). *A transformation $T(\underline{X}')$ is a mapping from the feature space $\mathcal{R}^n$ to $\mathcal{R}^{n'}$ and its Bayes risk with loss L is denoted with $R^*_{L,T}$. It is called admissible if for any loss function L, $R^*_{L,T} = R^*_L$, where R^*_L is the original Bayes risk. A transformation is admissible, if $T(\underline{X}')$ is a sufficient statistics (i.e., $p(Y|T(\underline{X}'), \underline{X}') = p(Y|T(\underline{X}'))$).*

<RELEVANT

The *relevance* of a feature can be defined in an algorithm-free, asymptotic, model-free and loss-free way as follows.

Definition 2.1.1. *A feature X_i is strongly relevant, if there exists some x_i, y and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exists some x_i, y and s'_i for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant [82, 86].*

The model-free, algorithm-free and loss-free conditional approach is called *filter approach* (for references, see [82, 86]). In the filter approach to feature selection we have to select a minimal subset $\underline{X}'$ that fully determines the conditional distribution of the target ($p(Y|\underline{X}') = p(Y|\underline{X}')$) without modeling the complete domain $p(Y, \underline{X}')$ or the explanatory variables $p(\underline{X}')$.

The Bayesian networks as representation of the independencies in the domain motivated a series of methods for identifying such a subset for the variable Y, particularly using the boundary of Y in DAG G in a distribution compatible with G (see Def. 1.1.9). However this set is not necessarily unique and not even minimal. The following theorem gives a sufficient condition for both [138].

Theorem 2.1.2 ([138]). *If distribution P is stable w.r.t. the DAG G, then the variables corresponding to the nodes in the boundary of Y, $\mathrm{bd}(Y, G)$ (the parents and children of Y and other parents of its children) forms a unique and minimal Markov blanket of Y, $\mathrm{MB}_P(Y)$ (the Markov boundary). Furthermore, $X_i \in \mathrm{MB}_P(Y)$, if X_i is strongly relevant.*

The Markov Blanket Approximating Algorithm assumes that the number of relevant variables is usually much larger for the target variable then for the explanatory variables, so it iteratively omits features for which there is a subset of features forming a Markov blanket without the target variable, consequently not influencing the conditional distribution of the target variable [90]. It uses pairwise correlation for finding a Markov blanket for the features and the KL distance to test the change of the conditional distribution. Recent extension of the algorithm and its applicaton to microarray data are reported in [151]. The Incremential Association Markov Blanket algorithm and its variants similarly use correlation measures and independence tests in forward-backward phases for identifying Markov Blankets, with asymptotic correctness and low computational complexity [138, 139]. In a recent extension a further wrapper phase were incorporated to filter the features that are irrelevant w.r.t. a specific classification method [5]. Other filter methods directly use Bayesian networks for a preliminary feature selection, which provides usually a restricted set of variables for a computationally more intensive classifier learning in the next phase. The $K2MB$ method first identifies a parental set for the target variables from all the explanatory variables using the $K2$ greedy method (see Section 1.5.2), then it applies the $K2$ algorithm for random orderings of this subset [28]. The learning of a GBN classifier similarly first applies a Bayesian network learning method [21], then it selects the boundary of the target node $\mathrm{MB}(Y, g)$ as a Markov blanket from the resulting Bayesian network G and applies a general Bayesian network learning algorithm or the learning of Bayesian multinets representing also contextual independencies [22, 23, 61].

The *wrapper approach* to feature selection similarly can apply the Bayesian networks as classifiers, in this case jointly in the feature selection phase and the phase of classifier learning [120, 80]. These filter methods indicate that the feature subset selection problem can be approached in a conditional and a model-based way. In the first case, to avoid the statistical (sample) and computational complexity corresponding to complete domain models, the Markov blanket is inferred independently of any other aspect of the domain model (i.e., without evaluating the implications of the identified features for the domain model). Thus these Bayesian network methods have still conditional and frequentist foundation, beside being model free and loss free. So on the one hand, unavoidably the scores for the subsets in these model-free methods has a vague relation to the performance of a given loss function and algorithm over specific model class restricted to the subsets [82, 86]. But on the other hand, (1) the scores do not utilize the potential of Bayesian networks as domain models (i.e., conditional scores), (2) they have hidden biases, and (3) they have no confidence measures with clear interpretation, partly because of the sequential application of statistical tests on a finite, frequently rather small amount of data. These can be answered in a domain model-based, Bayesian approach to the feature subset selection problem using Bayesian networks.

In the Bayesian conditional approach to feature selection θ encodes the presence of the explanatory variables, so $p(\theta|D)$ induce a (conditional) posterior distribution over the subsets (for an overview of using MCMC methods in a

conditional model space over structures with varying input features, see [113], for applications [41, 124]). A hierarchical conditional approach is the Automatic Relevance Determination (ARD) method [110], in which certain parameters represent the weights W_i(relevance) of the inputs (features) X_i, so the parameter posterior for the inputs $p(W_1, \ldots, W_n|D_N)$ can be used for the evaluation of a feature subset.

In the Bayesian domain-based (non-conditional) approach a conditional model of the target variable cannot be separated from the overall domain model or at least the conditional model and the model over the potential explanatory variables are dependent. For example, it is generally so for Bayesian network structure priors, so, as we shall see, we have to average over the model space to derive posterior for the part of the model relevant conditionally (see Eq. **??**).

As we saw in Th. 2.1.2, the boundary of the variable Y in the Bayesian network G identifies a minimal and unique Markov blanket $\mathrm{MB}(Y, G)$ for variable Y in any stable distribution w.r.t. the DAG G. Using Bayesian network with multinomial local dependency models as unconstrained domain models for discrete values and with Dirichlet parameter priors, the posterior probability of the Markov blanket expresses exactly the belief in the (observational) probabilistic relevance of the subset $\underline{X}'$:

$$p(\mathrm{MB}(Y) = \underline{X}'|D_N) = \sum_G 1(\mathrm{MB}(Y, G) = \underline{X}')p(G|D_N). \tag{2.11}$$

Recall that the structure posterior $p(G|D_N)$ represents the posterior belief in stable distributions w.r.t. G (the non-stables have measure zero see Section 1.1.2.3) and that DAGs in a equivalence class $G \in G^{\sim}$ represent the same set of independencies, so imply the same Markov blanket.

Though the concept of relevance corresponds to subsets, a corresponding pairwise measure can be introduced that defines individual "feature relevance"

$$p(\mathrm{MBM}(Y, X_i)|D_N) = \sum_G 1(X_i \in \mathrm{MB}(Y, G))p(G|D_N). \tag{2.12}$$

Because of model averaging it is still model-based (!), consequently biased towards "domain consistency", contrary to standard pairwise correlation and association measures. Note that only the Bayes risk based subset score is monotone, similarly to a mutual information based subset score, which makes the search in the space of subsets harder. For the application of bootstrap and Bayesian method over MBM features, see Section 2.3 and 2.5.2.3.

2.1.4 MBG subnetworks

The feature subset selection problem does not include explicitly the issue of dependencies between the features, though the interaction between the selected features is important for their interpretation. A generalization of the FSS problem includes the construction of a model containing the variables $\underline{X}'$ relevant to

a target variable Y and their observational dependency and causal dependency relations w.r.t. Y.

As shown in Eq. 2.15, the classification performance of a Bayesian network in case of complete data is fully determined by the Markov blanket spanning subgraph $\mathrm{MBG}(Y, G)$ and its parameters (the local models for Y and its children). Another interpretation of the MBG feature is that it encompasses all the causal mechanisms directly related to a given variable Y. Because of the generality of the MBG feature discussed in Section 2.2, we call such model a Markov Blanket Graph or Mechanism Boundary Graph (a.k.a. classification subgraph, feature subgraph).

In the conditional approach, the importance of the MBG feature was already identified, because early methods used the score of a complete Bayesian network G to score the classification performance of the model and to score the Markov blanket of the target variable. As noted in [51] and discussed in Section **??**, this is incorrect from the point of prediction of the target Y, particularly in the case of complete data, because this score includes (direct or indirect) complexity penalization w.r.t. the complete domain model that is not relevant for the MBG submodel relevant for classification. It is more appropriate to use special scores for the classification relevance of the MBG subnetwork and possibly even for scoring the feature subset. Such a classification oriented score is the conditional node monitor (or MBG monitor), its use was reported in [91, 92, 93, 3].

In conditional approaches using other models, the dependency models may contain such additional information about the conditional dependence structure. In Chapter **??** we discuss the logistic regression model, the tree augmented Bayesian network classifiers [51] and the augmented Bayesian classifier [84], which explicitly contain interactions and the MLP model, in which such information is rather implicit.

In the Bayesian framework using Bayesian networks, the corresponding score for the MBG feature is the posterior

$$p(\mathrm{MBG}(Y, G) = \mathrm{mbg}\,|D_N) = \sum_G 1(\mathrm{MBG}(Y, G) = \mathrm{mbg}) p(G|D_N). \tag{2.13}$$

FULLVERSION> Note that as for other features in general

$$\arg\max_{G^{\subseteq}} p(\mathrm{MBG}(Y) = G^{\subseteq}|D_N) \neq \mathrm{MBG}(Y, G^{\mathrm{MAP}}) \tag{2.14}$$

<FULLVERSION

2.1.5 Learning of subnetworks

The most general structural feature is a general subgraph of a Bayesian network. The identification of subgraphs with statistical significance was reported in [117]. In the first phase, this method generates confidence measure for the pairwise Markov blanket memberships $\mathrm{MBM}(X_i, X_j)$ using the bootstrap OPTIONAL> (for a discussion of the interpretations of bootstrap probabilities, see Section 2.3) <OPTIONAL.

Because it used interventionist data, the modified form of the closed expression of the posterior score of a Bayesian network was applied (see [30]). Next, using a heuristic threshold on the bootstrap probabilities for the pairs, it identifies components as starting seeds for a bottom-up expansion to generate multivariable features from the pairwise features. Finally in a greedy hill-climbing search it collects subnetworks in a pool using a statistical test with the null hypothesis that the (rank of) bootstrap probabilities of the pairwise features are independent in a subnetwork. The attractive assumption behind this approach is that pairwise features corresponding to the same or dependent causal mechanisms are dependent, so they can be identified jointly with higher significance. The evaluation indicated the advantage of this model-based (called "context specific" in their terminology) approach for detecting "correlation" compared to the investigation of direct associations of features with Pearson correlation. The continuation of this work similarly indicated the advantage of learning parts and modules using a special decomposed representation for the Bayesian network [128, 118]. This study also investigated the learning of global pairwise features, such as the existence of a directed path, causal effect between two variables and the learning of parametric features, such as the qualitative type of the local dependency models.

2.1.6 The properties and taxonomy of features

We introduce a terminology to analyze Bayesian network features, particularly the properties of a new BN feature we propose later. The concept of feature over DAGs (Bayesian networks) has a broad usage, it is used for random variables (i.e., a mapping from DAGs G to the real line), for their values, and even for mappings from DAGs G to a set of complete and mutually exclusive composite events. OPTIONAL> In its full generality, the event space of Bayesian networks over discrete variables V assuming multinomial local dependency models and parameter independence contains the DAGs with their parameters $p(G, \underline{\theta})$ (we do not consider further hyperparameters). Additionally, we allowed corresponding textual annotations in the form of an ABN-KB. The introduced probabilistic ABN-KB allows the definition of wide range of random variables and composite events representing parametric, qualitative monotonicity, structural and textual model properties. <OPTIONAL From another point of view, there are simple quantitative random graph properties such as mean in-degrees, out-degrees, clique sizes or lengths of directed paths, and there are complex indicators such as the ABN sentences or complex mappings to subgraphs such as the essential graphs. We use the term feature in a broad sense to denote any function over DAGs G or BNs (G, θ) (e.g., $F(G) : \mathcal{G} \to \mathcal{F}$). If the context allows, e.g. in case of binary features, we use the term feature to refer to the feature function, feature value, and also to the denoted graph property. Frequently a set of feature functions can be indexed by the variables $X_i \in V$ (i.e., $\{F_{X_i}(G)\}$) or pairs of the variables, etc., as it would be another argument of the feature function, so we can talk about univariate features $F(X_i, G)$ or pairwise features $F(X_i, X_j, G)$,

instead of referring to the corresponding sets of features. OPTIONAL> This includes random variables, the definitions of composite events by partitioning of the event space and the transformation of the event space into a space with less cardinality or dimensionality. For example, the number of parental edges to variable Y ($\#\,\mathrm{pa}(Y,G) : G \to \mathcal{N}$), the parental sets of variable Y ($\mathrm{pa}(Y,G) : G \to S, where S \subseteq \{V \setminus Y\}$) and the conditional distribution of variable Y given its parents in G ($\mathrm{pa}(Y,G,\underline{\theta}) : (G,\underline{\theta}) \to (S,\underline{\theta}_{Y|S}), where S \subseteq \{V \setminus Y\}$). In the subsequent part we concentrate mainly on the structural aspects (i.e., we assume that the parameters are averaged out giving $p(G)$ and we neglect the annotations as well). <OPTIONAL

FULLVERSION>

A feature F is called a *subset feature*, if it maps DAGs G to the subsets of $V' \subseteq V$. A feature F is called a *subgraph feature*, if it maps DAGs G to partially directed (sub)graphs (PDAG) over $V' \subseteq V$ (so it is a subset feature as well). As this is not restrictive per se, we use the term *structural feature* for "simple", graphically interpretable mapping functions, such as compelled edge selection or Markov blanket subgraph selection (we expect that the descriptive complexity of the function does not allow the encoding and using the PDAG (sub)graphs as mere indices, see [144]). Note that the partial ordering feature can be conceived of mapping each the DAG G to a representant DAG with minimal number of edges in the partition of the DAG space that partition exactly represents the topological orderings of G.

<FULLVERSION

A feature F is a *local feature* $F(V',G)$, if its value depends only on the subgraph of G spanned by the argument variables $V' \subseteq V$ denoted with $G^{|V'}$ (i.e., $(G_1^{|V'} = G_2^{|V'}) \Rightarrow (F(G_1) = F(G_2))$, where $G^{|V'}$ contains nodes $V' \subseteq V$ and edges of G from V' to V'). A non-local called *global feature* indicates a potential relation to other features and increased computational complexity.

FULLVERSION> The *parental power set feature* $Pa_n(G)$ is the mapping of DAGs over n nodes to power sets $Pa : G \to \{\mathrm{Pa}(X_1,G), \ldots \mathrm{Pa}(X_n,G)\}$, which allows the investigation of regularities of the parental sets such as common regulators. <FULLVERSION A feature F is a *modular*, if it depends only on the parental sets in DAG G (i.e., $(\mathrm{Pa}(G_1) = \mathrm{Pa}(G_2)) \Rightarrow (F(G_1) = F(G_2))$). A feature is *ordering-modular*, if for all except at most one feature value f and for each complete ordering $\prec$ there is a conjunctive normal form $C_1 \wedge \cdots \wedge C_n$ such that each clause $C_i(f,\prec),G$ for $i = 1, \ldots n$ depends only on $\mathrm{Pa}(X_i,G)$ for all $G^{\prec}$ (i.e., $C_i(f,\prec,\mathrm{pa}(X_i,G))$). Note that the compelled edge relation and the pairwise MBM relevance relation between X_i, X_j are not local, but the MBM relation (through its false value) is modular [55].

Another general type is the *observationally equivalent feature* F, if the mapped subgraph $F(G)$ over the variables $V' \subseteq V$ depends on only the essential graph of G, $G^{\sim}$ (i.e., $(G_1^{\sim} = G_2^{\sim}) \Rightarrow (F(G_1) = F(G_2))$).

A feature F is called a *complex feature*, if the number of values of the feature is exponential in the number of domain variables. FULLVERSION> Note that

the number of values of a complex structural feature (i.e., number of partitions) can be in the order of the cardinality of PDAGs over maximum n variables, $\#PDAG(\leq n) = \sum_1^n \#PDAG(i)$, which is super-exponential in n $2^{\mathcal{O}(n^2 \log n)}$.
<FULLVERSION FULLVERSION>

Now we discuss the properties of sets of features. The number of structural features (mappings $DAG(n) \rightarrow PDAG(\leq n)$) is $\mathcal{O}(\#PDAG(\leq n)^{\#DAG(n)})$.
<FULLVERSION

A set of features $\{F_1, \ldots, F_L\}$ called *DAG-independent feature set*, if the values of the features can be selected arbitrarily without violating the DAG-constraint (i.e., for each L-tuples of feature values, there is one or more DAG G with these feature values: $\forall\{f_1, \ldots, f_L\}\exists G : (F_1(G) = f_1) \wedge \cdots \wedge (F_L(G) = f_L)$). Because the L-tuples of feature values partition the DAG space, it also means that for each distribution over the feature set $p(F_1, \ldots, F_L)$ there exists a distribution over DAGs $p(G)$ such that $p(F_1, \ldots, F_L) = p(F_1(G), \ldots, F_L(G))$. Consequently, we can treat the features as totally independent and specify a distribution $p(G)$ by using the form $\prod_{i=1}^L p(F_i)$ and for example spreading uniformly the masses within each partition defined by the feature values. However the converse is not true, that is a set of DAG-independent features is not independent in the induced distribution of a general $p(G)$ (i.e., DAG-independence does not imply independence in general). Note that local features can still be DAG-dependent, such as the directed edge features, so locality does not imply DAG-independence.

Finally, let S denote an elementary event (e.g., either G or $(G, \underline{\theta})$). A set of features $\{F_1, \ldots, F_L\}$ is called a *complete feature set*, if for each S the set of values $\{F_1(S), \ldots, F_L(S)\}$ identifies S (i.e., $(S_1 \neq S_2) \Rightarrow (\{F_1(S_1), \ldots, F_L(S_1)\} \neq \{F_1(S_2), \ldots, F_L(S_2)\})$). A set of features $\{F_1, \ldots, F_L\}$ is called complete w.r.t. a feature $F^*(S)$, if for each S the set of values $\{F_1(S), \ldots, F_L(S)\}$ identifies $F^*(S)$. In turn, a feature $F^*(S)$ is a *sufficient feature* for a set of features $\{F_1, \ldots, F_L\}$, if $\forall\ S, i :\ F_i(S) = F_i(F^*(S))$, consequently $p(F_1(S), \ldots, F_L(S))$ can be induced from the distribution of the complex feature $p(F^*(S))$. If additionally, the set of features are complete then the complex feature is called *exact feature* for the feature set (as it is a one-to-one/bijective relation).

FULLVERSION>

Now we consider quantitatively the issue of incompleteness, dependency and redundancy w.r.t. features vs. model and simple-features vs. complex-feature, using that $I(X;Y) = H(X) + H(Y) - H(X,Y) = \mathrm{KL}(p(X,Y) \| p(X)p(Y)) \leq 0$ and $H(f(X)) \leq H(X)$ see [32]. First note that $H(F_1(G), \ldots, F_L(G)) \leq H(G)$ and for a sufficient feature $H(F_1(F^*(G)), \ldots, F_L(F^*(G))) \leq H(F^*(G))$, equality with completeness only. If the features are dependent, then $H(F_1(G), \ldots, F_L(G)) \leq \sum_i H(F_i(G))$, equality with total independence only. The mutual information between the feature set and the model or complex feature quantifies the expected log loss of using the features ($\mathrm{KL}(p(G) \| p(F_1(G), \ldots, F_L(G)))$ and $\mathrm{KL}(p(F^*(G)) \| p(F_1(G), \ldots, F_L(G)))$). So, for the case of an exact complex feature, the mutual information $I(F_1(G), \ldots, F_1(G); F^*(G)) = \mathrm{KL}(p(F^*(G)) \| \prod_i p(F_i(G)))$ measures the dependency/redundancy between the (sub)features or in other

words the expected loss of the approximation of treating the features as if they were independent.

For example a complex feature, the Markov blanket of a target variable $\mathrm{MB}(Y)$ is an exact feature for the set of the Markov blanket membership pairwise features $\mathrm{MBM}(Y, X_i)$ for all $X_i \in V \backslash Y$. Then $\mathrm{KL}(p(\mathrm{MB}(Y)) \| \prod_i p(\mathrm{MBM}(Y, X_i, (G)))) = \sum_i H(\mathrm{MBM}(Y, X_i, G)) - H(\mathrm{MB}(Y, G))$ measures the expected loss of approximating the relevance (probability) of a subset of features with pairwise relevances as if they were independent (i.e., with product of probabilities of pairwise relevance) and the dependency/redundancy between the pairwise features.

As another example consider, that the set of all directed edge features $1(\mathrm{E}_{ij} \in Edge(G))$ or $\mathrm{E}_{ij}(G)$ for short for $i \neq j, (i, j) = (1, 2,), \ldots, (n, n-1)$ are a complete feature set for Bayesian networks $G(n)$. Then $\mathrm{KL}(p(G) \| \prod_i p(\mathrm{E}_{ij}(G))) = \sum_i H(\mathrm{E}_{ij}(G)) - H(G)$ measures the expected loss of approximating the distribution over the DAGs with a pairwise based product approximation and measures the dependency/redundancy between the edge features.

Finally, note that both in the investigation of completeness and redundancy of a feature set these quantities can be applied relatively by canceling the term $H(G)$ or $H(F^*(G))$. That is we can compare the completeness (i.e., loss of information) of feature sets F and F' by evaluating $H(F_1(G), \ldots, F_L(G))$ and $H(F'_1(G), \ldots, F'_K(G))$. For the redundancy of exact feature sets F and F' by evaluating $\sum_i H(F_i(G))$ and $\sum_i H(F'_i(G))$.

<FULLVERSION

2.2 The Markov Blanket (sub)Graph feature

In this section we propose a complex feature, Markov Blanket (sub)Graph feature $(\mathrm{MBG}(Y), \underline{\theta}_{\mathrm{MBG}})$, that includes all the direct causal and probabilistic relations corresponding to a given variable. This feature is at an intermediate level as its complexity is less than of the complete domain model. We show it is a necessary and sufficient feature w.r.t. classification of Y under the usual assumptions in the thesis, such as complete data, discrete values, multinomial local dependency models. The MBG feature can be equally derived from a causal point of view using the mechanism-interventionist interpretation as the minimal set of mechanisms directly relevant for Y, so we equally use the term Mechanism Boundary (sub)Graph feature. It means that the MBG feature represents such a fragment of the domain theory that its distribution is necessary and sufficient to induce the exact posteriors for any classification related feature, to support full scale Bayesian inference and to induce various priors for classifiers, such as logistic regression or multilayer perceptrons. In other words, the complex feature does not violate the dependency of (sub)features for these tasks by modeling them as independent (obviously the MBGs for different variables $(\mathrm{MBG}(X_i), \mathrm{MBG}(X_j), X_i \neq X_j)$ are dependent at the model level in general, so interpreting them as independent using $p(\mathrm{MBG}(X_i), \mathrm{MBG}(X_j)) = p(\mathrm{MBG}(X_i))p(\mathrm{MBG}(X_j))$ would be incorrect).

Definition 2.2.1 ([10, 9])**.** *The parametric Markov Blanket (sub)Graph feature or Mechanism Boundary Graph feature for a variable Y* $\mathrm{pMBG}(Y, G, \underline{\theta}_G)$ *maps Bayesian network models $(G, \underline{\theta}_G)$ to Markov Blanket Graphs of variable Y and to its parameters* $(\mathrm{MBG}(Y), \underline{\theta}_{\mathrm{MBG}(Y)})$*. The (non-parametric) Markov Blanket Graph feature for a given variable Y denotes the mapping of Bayesian network structures (G) to the Markov Blanket Graphs of variable Y (see Def. 1.1.11, Fig.* **??***, and Fig. 1.1).*

Because of our general assumptions of global parameter independence and parameter modularity, we always assume that the parameter transformation is a simple selection, so the parameter distribution is unchanged (i.e., $\underline{\theta}_{\mathrm{MBG}(Y,G)} = \{\underline{\theta}_Y, \underline{\theta}_{\mathrm{ch}(Y,G)_1}, \ldots, \underline{\theta}_{\mathrm{ch}(Y,G)_K}\}$ is equal to the corresponding parameters in $(G, \underline{\theta})$, where $K = |\,\mathrm{ch}(Y, G)|$). Note that these parameters by simple selection differ from the parameters from a marginalization over all the compatible BNs or over all the compatible BNs and a given ordering.

The characteristic property of the pMBG feature is that it completely defines the conditional distribution of Y given the other variables $V \setminus Y$ in a Bayesian network model $(G, \underline{\theta})$ by the local dependency models of Y and its children.

Proposition 2.2.1. *If $p(\underline{V}|G, \underline{\theta})$ is defined by a Bayesian network $(G, \underline{\theta})$, then the conditional distribution of the target variable $Y \in \underline{V}$ $p(Y|\underline{V} \setminus Y, G, \underline{\theta})$ is defined by its Markov Blanket (sub)Graph feature* $\mathrm{pMBG}(Y, G, \underline{\theta}_G)$*.*

Proof.

$$
\begin{aligned}
& p(Y|V \setminus Y, G, \underline{\theta}) \qquad (2.15)\\
&\quad = \; p(Y|\,\mathrm{MB}(Y, G), G, \underline{\theta}) = p(Y|\,\mathrm{pa}(Y, G), \mathrm{ch}(Y, G), \mathrm{pa}(\mathrm{ch}(Y, G), G), \underline{\theta})\\
&\quad \propto \; p(\mathrm{ch}(Y, G), Y|\,\mathrm{pa}(Y, G), \mathrm{pa}(\mathrm{ch}(Y), G), \underline{\theta})\\
&\quad = \; p(Y|\,\mathrm{pa}(Y, G), \underline{\theta}) \prod_{j=1}^{|\,\mathrm{ch}(Y,G)|} p(\mathrm{ch}(Y, G)_j|\,\mathrm{pa}(\mathrm{ch}(Y, G)_j), \underline{\theta}),
\end{aligned}
$$

where $\mathrm{ch}(X_i, G)_j$ denotes the children of X_i in a compatible ordering with G. □

For notational simplicity we assume a binary target varible Y. Let us define a vector-valued feature called *conditional distributional feature* $\mathrm{CD}(Y, G, \underline{\theta})$ denoting the conditional distribution $p(Y|V \setminus Y, G, \underline{\theta})$. Let $\mathrm{CD}(Y, G)$ denote the corresponding averaged conditional distribution $p(Y|V \setminus Y, G)$ (see Eq. 1.29 for the existence of an equivalent point parametrization).

Furthermore, we can state a Bayesian extension of Proposition 2.2.1.

Proposition 2.2.2. *In case of parameter independence, parameter modularity and Dirichlet parameter priors, the Markov Blanket structural and parametric marginals $p(\mathrm{MBG}(Y, G) = \mathrm{mbg})$ and $p(Y|\,\mathrm{MBG}(Y, G) = \mathrm{mbg})$ define the conditional distribution of Y given other variables $V \setminus Y$ in the Bayesian framework, where*

$$
p(\mathrm{MBG}(Y, G) = \mathrm{mbg}) = \sum_G 1(\mathrm{MBG}(Y, G) = \mathrm{mbg}) p(G) \qquad (2.16)
$$

and $p(Y|\,\mathrm{MBG}(Y,G)=\mathrm{mbg})$ *denotes the mean distribution* $\mathrm{E}_{\underline{\Theta}'}[p(Y|mbg,\underline{\Theta}')]$.

Proof.

$$
\begin{aligned}
& p(Y|V\setminus Y) \qquad (2.17)\\
&= \sum_G p(G)\int p(Y|G,\underline{\theta})p(\theta|G)\,\mathrm{d}\underline{\theta}\\
&= \sum_G p(G)\int p(Y|\,\mathrm{MBG}(Y,G),\underline{\theta}_{\mathrm{MBG}(Y,G)})p(\underline{\theta}_{\mathrm{MBG}(Y,G)}|G)\,\mathrm{d}\underline{\theta}_{\mathrm{MBG}(Y,G)}\\
&= \sum_G p(G)p(Y|\,\mathrm{MBG}(Y,G))\\
&= \sum_{\mathrm{MBG}(Y,G)=\mathrm{mbg}} p(\mathrm{mbg})p(Y|\,\mathrm{mbg}),
\end{aligned}
$$

□

Note that Proposition 2.2.2 also indicates that Bayesian model averaging for prediction can be performed in the MBG space, because the parametric marginal $p(Y|\,\mathrm{MBG}(Y,G)=\mathrm{mbg})$ is efficiently computable in case of Dirichlet parameter priors (see Eq. 1.29). However, in general there is no closed formula for the posterior $p(\mathrm{MBG}(Y,G)=\mathrm{mbg})$, but we can state the following theorem.

Theorem 2.2.1 ([10])**.** *If the parental set size is bounded by* k *and the scores* $p(pa(X_i)|D_N)$ *in Eq. 1.51 are available in* $\mathcal{O}(1)$*, then the ordering-conditional posterior* $p(\mathrm{MBG}(Y,G)=\mathrm{mbg}\,|\,\prec)$ *can be computed in polynomial time.*

Proof. If the parental set size is bounded by k, then

$$
\begin{aligned}
& p(\mathrm{MBG}(Y,G)=\mathrm{mbg}\,|D_N,\prec) \qquad (2.18)\\
&= p(\mathrm{pa}(Y,\mathrm{mbg})|D_N)\prod_{\substack{Y\prec X_i\\ Y\in\mathrm{pa}(X_i,\mathrm{mbg})}} p(\mathrm{pa}(X_i,\mathrm{mbg})|D_N)\prod_{\substack{Y\prec X_i\\ Y\notin\mathrm{pa}(X_i,\mathrm{mbg})}} p(Y\notin\mathrm{pa}(X_i,\mathrm{mbg})|D_N),
\end{aligned}
$$

where

$$
p(Y\notin\mathrm{pa}(X_i,\mathrm{mbg})|D_N)=\sum_{Y\notin\mathrm{pa}(X_i)} p(\mathrm{pa}(X_i)|D_N). \qquad (2.19)
$$

□

Clearly, for a given Markov Blanket structure and ordering Eq. 2.18 directly defines a conjunctive normal form, which gives the next property.

Corollary 2.2.1 ([10, 9])**.** *The Markov Blanket (sub)Graph feature* $MBG(Y,G)$ *is an ordering-modular feature.* □

The number of MBGs for a given variable $|\,\mathrm{MBG}(Y)|$ in case of n variables is still super-exponential (even if the number of parents is bounded above with k). Consider an ordering of the variables such that Y is the first and all the other

variables are children of it, then the parental sets can be selected independently, so the number of alternatives is in the order of $(n-1)^{n^2}$ (or $(n-1)^{(k-1)(n-1)}$). However, at the other extreme, if Y is last in the ordering, then the number of alternatives (i.e., parental sets) is in the order of 2^{n-1} or $(n-1)^{(k)}$). In case of MBG(Y, G), the types of the variable X_i can be (1) non-occurring in the MBG, (2) parent of Y ($X_i \in \text{pa}(Y, G)$), (3) children of Y ($X_i \in \text{ch}(Y, G)$) and (4) (pure) other parent in the MBG ($(X_i \notin \text{pa}(Y, G) \wedge (X_i \in \text{pa}(\text{ch}(Y, G)_j))))$). These types correspond to the categories irrelevant (1) and strongly relevant (2,3,4), as can be seen directly from the definitions of relevance (see, Def. 2.1.1). The number of DAG models $G(n)$ compatible with a given MBG and ordering $\prec$ can be computed as follows: the contribution of the variables $X_i \prec Y$ without any constraint and the contribution of the variables $Y \prec X_i$ that are not children of Y. Let us denote the number of such variables with N_B and N_A respectively, then assuming that the maximal number of parents is k, the number of compatible DAGs is $2^{\Theta((k-1)(N_B+N_A)\log(n))}$.

Proposition 2.2.1 and Proposition 2.2.2 offer two interpretations for the MBG feature. From a (conditional) probabilistic point of view the MBG(G) feature defines an equivalence relation over the DAGs w.r.t. the conditional distribution of Y given all the other variables under parameter modularity and global parameter independence. This is the consequence of Th. 1.1.3 and Th. 2.1.2, which allow the reduction of the space of DAGs to the space of MBGs from the point of view inferring a given variable. If the hypotheses are the observational classes (i.e. the parameter and structural priors are identical for observationally equivalent DAGs), then this conditionally induced equivalence relation is combined with the observational equivalence relation, which allows further reduction of the space of MBGs (for a partially oriented representation of the MBGs, see [2, 3]). We show certain properties of this combined equivalence, although in our exploratory context we assume causal priors, so we cannot simplify further the MBG space. In the non-Bayesian context let us define a pairwise relation over Bayesian networks as G_1 and G_2 are inferentially equivalent for variable Y, if they can encode the same set of conditional distributional features for Y (i.e., for each CD$(Y, G_1, \underline{\theta}_1)$, there exists a $\underline{\theta}_2$ such that CD$(Y, G_1, \underline{\theta}_1)$ = CD$(Y, G_2, \underline{\theta}_2)$). Clearly, observational equivalence and MBG equivalence of DAGs G_1, G_2 implies conditional distributional equivalence (IE), but MBG equivalence and conditional distributional equivalence does not imply observational equivalence. Interestingly, MBG equivalence is not implied by observational equivalence or by conditional distributional equivalence (i.e., the MBG feature is not a unique representant of an inferentially equivalent class of Bayesian networks and it can be different in observationally equivalent DAGs).

From a causal point of view, this feature uniquely represents the minimal set of mechanism including Y despite the non-uniqueness of the MBG feature w.r.t. the acausal conditional distributional equivalence. This offers the second interpretation of the MBG feature: the pMBG$(Y, G, \underline{\theta})$ feature includes exactly the mechanisms containing the variable Y, hence the name Mechanism Boundary (sub)Graph feature pMBG$(Y, G, \underline{\theta})$. The probability of an MBG is the sum of

probabilities of the causal domain models that are compatible with this causal subtheory for the variable Y (Eq. 2.16), which shows that for example inferentially equivalent MBGs may have different probabilities in a causal context (e.g., in case of causal prior or interventionist data).

From Proposition 2.2.1 we can conclude that the $\mathrm{MBG}(Y, G)$ feature is necessary and sufficient to represent the mechanisms directly relevant for the variable Y and from the point of view of prediction of Y, it is a *sufficient* feature for the conditional distributional features of Y. In other words, under the conditions such as parameter modularity, global parameter independence and complete data assumption, this structural and parametric feature of the causal BN domain model is necessary and sufficient to support the manual exploration and automated construction of a causal, probabilistic, interpretable conditional dependency model. This "ultimate" property of the MBG feature suggests the concept of conditional feature and the generalization of the feature subset selection problem.

Definition 2.2.2. *A feature (function) F is called conditional feature for a given variable Y, if it depends only on* $(\mathrm{MBG}(Y), \underline{\theta}_{\mathrm{MBG}(Y)})$

$$\mathrm{pMBG}(Y, G_1, \underline{\theta}_1) = \mathrm{pMBG}(Y, G_2, \underline{\theta}_2)) \Rightarrow (F(G_1, \underline{\theta}_1) = F(G_2, \underline{\theta}_2). \quad (2.20)$$

Definition 2.2.3. *In case of a stable distribution $p(Y, \underline{X})$, the feature (sub)graph selection problem (FGS) denotes the identification of a Markov Blanket subgraph $MBG(Y, G)$, where DAG G denotes a perfect map of distribution p (i.e., it includes the identification of a Markov Blanket set $\underline{X}' \subseteq \underline{X}$ w.r.t. p and Y, and a Bayesian network substructure over $\underline{X}'$ representing the dependencies between these variables, excluding incoming edges into the parents of Y).*

FULLVERSION> Later we will define more conditional features, an already defined such feature is the conditional distributional feature $\mathrm{CD}(Y, G, \underline{\theta})$. Because these features are determined by the pMBG feature, clearly the following statements hold for any set of conditional features $\underline{F}(G, \underline{\theta})$: (1) The pMBG feature is sufficient for $\underline{F}(G, \underline{\theta})$. (2) The generally dependent conditional features in $\underline{F}(G, \underline{\theta})$ are conditionally independent given the feature pMBG. (3) Given a data set of domain cases $D_N = \{\underline{C_1}, \ldots, \underline{C_N}\}$ (i.e., each variable $X_i \in \underline{V}$ is observed), the posterior distribution $p(MBG, \underline{\theta}_{MBG}|D_N, \xi)$ is sufficient for inducing the joint posterior for any set of conditional feature $p(\underline{F}(G, \underline{\theta}|D_N, \xi)$ (so it sufficiently summarizes the data set). (4) Given an i.i.d. data set of Bayesian network structures $D_N = \{G_1, \ldots, G_N\}$ from $p(G)$ the statistics $D'_N = \{\mathrm{MBG}(Y, G_1), \ldots, \mathrm{MBG}(Y, G_N)\}$ is a sufficient statistics for estimating? $\underline{F}(G, \underline{\theta})$. <FULLVERSION

FULLVERSION>

2.2.1 Challenges of the Bayesian application of the MBGs

Now we summarize the advantages and challenges of the application of the MBG feature in the Bayesian context.

Normativity The MBG feature has standard probabilistic and causal interpretation. The $p(\mathrm{MBG}(Y,G)|D_N,\xi)$ posterior is a normative confidence measure. The computation of the posterior probability of a given feature value mbg is a standard Bayesian task of computing the expectation

$$p(\mathrm{MBG}(Y,G)=\mathrm{mbg}\,|D_N,\xi)=\sum_G p(G|D_N,\xi)1(\mathrm{MBG}(Y,G)=\mathrm{mbg}). \quad (2.21)$$

The lack of a closed formula for the posterior $p(\mathrm{MBG}(Y)=\mathrm{mbg})$ excludes the direct use of the MBG space in optimization or in Monte Carlo methods. However, general MC methods can be applied successfully to approximate such expectations, but the task of finding the maximum a posteriori feature value $\mathrm{mbg}^{\mathrm{MAP}}$ and particularly the task of constructing a set of feature values with high posteriors (i.e., a good approximation of $p(\mathrm{MBG}(Y))$) requires new methods, because of the high number of feature values.

Feature dependency and complexity The MBG feature is a statistically motivated feature to balance between the size of the data and the modeled part of the domain model with intermediate complexity, while modeling jointly a compact part of the domain from the point of view of dependency analysis.

Offline probabilistic knowledge base Whereas the MBG space cannot be used directly in optimization or Monte Carlo methods, its comprehensiveness from the point of view of dependency analysis makes it an ideal candidate for embedding in probabilistic knowledge bases, such as in an ABN-KBs. Because of its sufficiency, it is in itself through its distribution $p(\mathrm{MBG}(Y))$ provides the basis of a probabilistic knowledge base. In the thesis we also developed a method to build an offline probabilistic knowledge base containing an approximation of the posterior distribution of the investigated dependent variable $p(\mathrm{MBG}(Y)|D_N,\xi)$. This knowledge base contains a set of $\mathrm{MBG}(Y)$s with high posteriors (HPD^{MBG}) and the inference about the conditional features $F_i(\mathrm{MBG}(Y,G), F_j(\mathrm{MBG}(Y,G)$ can be performed in the smaller space of MBGs, more exactly in its approximation,

$$\begin{aligned} p(F_i(\mathrm{MBG}(Y,G))=f_i, F_j(\mathrm{MBG}(Y,G))=f_j|D_N,\xi) & \quad (2.22)\\ = \sum_{\mathrm{mbg}} p(\mathrm{mbg}\,|D_N,\xi)1(F_i(\mathrm{mbg})=f_i, F_j(\mathrm{mbg})=f_j) & \quad (2.23)\\ \approx \sum_{\mathrm{mbg}\in HPD^{MBG}} \hat{p}(\mathrm{mbg}\,|D_N,\xi)1(F_i(\mathrm{mbg})=f_i, F_j(\mathrm{mbg})=f_j). & \quad (2.24) \end{aligned}$$

The importance of this offline MBG knowledge base is that the queries, such as $F_i(\mathrm{MBG}(Y,G)=f_i, F_j(\mathrm{MBG}(Y,G)=f_j$, do not have to be specified in advance of the costly MCMC simulation to approximate the corresponding posterior, but the MCMC simulation runs only once, constructs the knowledge base by averaging the parameters and partly the structures (actually estimating a sufficient posterior for any conditional feature), then the knowledge base can be queried multiple times in an explorative manner. For example the distribution

of the misclassification rate on an external test data set $MR(\mathrm{MBG}(Y,G))$ and the number of free parameters $|Param(\mathrm{MBG}(Y,G,)|$ can be generated offline (see Fig. **??**).

This MBG knowledge base is also integrated with the ABN-KB functionality of the system, so the queries on conditional features can be enriched possibly with textual annotations.

In fact, as the conditional distributional features $\mathrm{CD}(Y,G,\underline{\theta},v)$ are classification features, the full scale Bayesian inference to compute the conditional distribution $p(Y|V\setminus Y)$ similarly can be approximated using the offline MBG knowledge base

$$p(Y|V\setminus Y) \sum_{\mathrm{mbg}} \hat{p}(\mathrm{mbg}\,|D_N,\xi)p(Y|\,\mathrm{mbg},V\setminus Y) \tag{2.25}$$

$$\approx \sum_{\mathrm{mbg}\in HPD^{MBG}} \hat{p}(\mathrm{mbg}\,|D_N,\xi)p(Y|\,\mathrm{mbg},V\setminus Y), \tag{2.26}$$

where $p(Y|\,\mathrm{mbg},V\setminus Y)=\int p(Y|\,\mathrm{mbg},\underline{\theta}V\setminus Y)\,\mathrm{d}\underline{\theta}$. So, the advantage of the MBG representation exceeds the general advantages of being a Bayesian BN-feature, because of its sufficiency for the conditional distributional features allows its application in full Bayesian inference for variable Y. Whereas its direct use is not possible in the on-line inference, the ordering-MBG space can be used for this purpose and the MBG representation itself provides a representation that minimize the model space, so support the construction of offline knowledge bases for real-time full scale Bayesian inference. This is further discussed in Section 2.8.2.

Induced priors, two-phased and dual learning The direct use of MBG posteriors as priors or hard constraint in a second phase of Bayesian network learning is not investigated in the thesis, only derived, conditional features.

<FULLVERSION

FULLVERSION>

2.2.2 On the practical importance of conditional features

This focusing of the analysis to a given variable is not only statistically motivated to find a semantically meaningful feature with intermediate complexity, but from an application oriented point of view as well. The reasons can be (1) the goal of the data analysis itself (2) the prior background knowledge and (3) the data set. We illustrate these reasons using the IOTA project in the ovarian cancer domain, which is the application area of the thesis. First, the primary goal is the probabilistic (and possibly causal) understanding of the relevance of domain variables for the type of the mass. This is important from clinical point of view to improve diagnostic protocols, to design later data collections and experiments and from statistical point of view to learn better predictive diagnostic models (either using Bayesian networks or other model class as classifier). Second, the available prior knowledge is focused. Because of the central importance

of target variable (the type of the mass), the mechanisms directly related to this variable are better explored theoretically. This bias is further strengthened in practice by the usual diagnostic usage of the domain knowledge, which involves the refinement of the causal rules and the development of diagnostic short-cuts such as the diagnostic relevance of variables, risk ratios of variables, antagonist/protagonist interactions of variables, available performance with certain variables, value of further information in a certain situation (see [7]). Finally, the data set itself entail certain focusing in learning complete domain models) On the one hand, the recorded variables were selected to support the analysis of the target variable for example by including potential confounders, which helps the interpretation of the edges around the target variable (i.e., the Causal Markov condition validity is restricted to the relations of the target variable) On the other hand, only the explanatory variables has missing values, furthermore missing data is more frequent for the diagnostically less relevant variables.

<FULLVERSION FULLVERSION>

2.2.3 Derived conditional features

First we overview the relation of earlier introduced general Bayesian network features to the MBG feature, then we introduce new conditional features. Finally we indicate their use in manual dependency analysis and inducing priors for dependency models.

The Markov blanket set feature ($\mathrm{MB}(Y,G)$), the Markov blanket membership feature ($\mathrm{MBM}(Y,X_i,G)$, the parental subgraph feature ($PaG(Y,G)$), the parental set feature ($\mathrm{pa}(Y,G)$) are all determined by the $\mathrm{MBG}(Y,G)$ feature, so these are conditional features. Furthermore, the conditional relevance of X_i for Y ($CR(Y,X_i)$) is also represented by the $\mathrm{MBG}(Y,G)$ feature as pure other parent of a children of Y ($CR(Y,X_i) \Leftrightarrow ((X_i \notin \{\mathrm{pa}(Y,G),\mathrm{ch}(Y,G)\}) \wedge (X_i \in \mathrm{MB}(Y,G))))$.

The pairwise causal feature ($Y \prec_G X_i$ or $X_i \prec_G X_j$) only partially represented by the $\mathrm{MBG}(Y,G)$ feature. The compelled edge status ($RCEdge(,,G)$) is similarly only partially represented, even if one of the argument is Y. This can be seen by considering that an edge is reversible, if there is a series of reversion of covered edges, where covered edge means that its endpoints have the same set of parents except corresponding to the edge [74]; and the $\mathrm{MBG}(Y)$ feature represents the parents only for Y and its children and not for the parents of Y and for the pure other parents Y's children. Finally, the $\mathrm{MBG}(Y)$ feature similarly only partially represents the pairwise confounding relation ($Conf(,,G)$), because as in the earlier case distant common ancestors for the parents of Y and for the pure other parents Y's children are not represented. However, the partial ability of the MBG feature — as a comprehensive conditional feature — for representing Bayesian network features related to causality and confounding is understandable as these are borderline issues in dependency analysis as well (corresponding to complete domain modeling) [78].

The first family of conditional features corresponds to the conditional probability feature and its performance on a given data set. The Markov blanket

conditional probability feature $\mathrm{CD}(Y, G, \underline{\theta}, v)$ and its expectation $\mathrm{CD}(Y, G, v)$. The misclassification rate $MR(Y, G, \underline{\theta}, D_N)$ on a given external data set D_N using the $\mathrm{CD}(Y, G, \underline{\theta}, v)$ predicted conditional probabilities. The parameter-free analog misclassification rate $MR(Y, G, D_N)$ is based on the mean of predicted conditional probabilities the $\mathrm{CD}(Y, G, v)$ (note, that this is based on the optimally reported values under L_2 and not equal to the loss-free expectation of the earlier measure $\mathrm{E}[MR(Y, G, \underline{\theta}, D_N)]$). A more sophisticated measure of predictive performance in binary classification is the Area Under the (ROC) Curve $AUC(Y, G, \underline{\theta}, D_N)$ and its parameter-free analog $AUC(Y, G, D_N)$ based on the predicted conditional probabilities $\mathrm{CD}(Y, G, \underline{\theta}, v)$ or on its mean $\mathrm{CD}(Y, G, v)$, which are discussed in Section **??**).

The second family of new conditional features corresponds to the graphical structure and to the parameter structure of the $\mathrm{MBG}(Y)$ feature. Features related to the complexity of the dependency model are the size of the Markov blanket set ($|\,\mathrm{MB}(Y, G)|$) and the number of free parameters $|\underline{\theta}_{\mathrm{MBG}(Y,G)}|$. The set of the non-interacting inputs and its size are denoted with $NII(Y, G, X_i)$ and $|NII(Y, G, X_i)|$, which set contains the children without other parent(s) than Y and the single parent of Y (if any). The concept of interaction is discussed in Section **??**. The set of potential interacting variables for variable X_i is given by the feature $IT(Y, G, X_i)$, which set contains the other parents of the children of X_i. Finally, $|BANEdges(Y, G)|$ denotes the number of edges between the children of Y in the BAN converted MBG defined in Section **??** (not including the parental edges of Y).

The use of these features will be demonstrated in the exploratory data analysis for dependency models in Section **??**. Furthermore, these features can be incorporated in composite queries to study their joint distribution, for example the joint distribution of model complexity and classification performance ($p(|\underline{\theta}_{\mathrm{MBG}(Y,G)}|, AUC(Y, G, \underline{\theta}, D_N))$, either using a dedicated Monte Carlo simulation to estimate this distribution or using an offline MBG probabilistic knowledge base.

These features can be also used to induce informative prior distributions for classifiers for example the $\mathrm{MBM}(Y, X_i)$ feature as a prior probability for the relevance of the variable X_i or the number of free parameters $|\underline{\theta}_{\mathrm{MBG}(Y,G)}|$ to select optimal model complexity (see Section **??**).

<FULLVERSION

RELEVANT>

2.3 The bootstrap confidence measure

The *bootstrap* approach to induce confidence measures for Bayesian network features was investigated as an alternative to the Bayesian approach to support statistical inference from small sample [54, 53]. An important motivation was to avoid the Monte Carlo simulations usually necessary in the Bayesian approach by using a simple resampling scheme and optimization.

The bootstrap is a general purpose, computationally intensive statistical inference method using resampling to assess the accuracy of a statistical estimate given a finite sample [48, 70]. We discuss it here as we refer to it only in this context, but it is a general statistical methodology and applicable with arbitrary model classes (or without as a nonparametric bootstrap). Assume a fixed i.i.d. sample $D_N = \{\underline{X_1}, \ldots, \underline{X_N}\}$ and let us denote $\hat{\theta}(D_N)$ the statistical estimate of interest and θ_0, the unknown true parameter. For a given sample size N its distribution, particularly its deviation $\hat{\theta}(D_N) - \theta_0$ is also of interest for constructing confidence intervals and hypothesis testing. The standard frequentist approach analytically derives its distribution, confidence intervals for restricted sets of sampling models and estimates (e.g., Gaussian data generation and mean estimate). Note that if we had access to the generative model $p(X|\underline{\theta_0}))$, we could sample it for any complex estimate. The standard Bayesian approach would define a probabilistic model for the observations $p(X|\theta)$ with prior $p(\theta)$ providing a distribution for the estimate $\int p(\hat{\theta}(D_N)|\theta)p(\theta)\,\mathrm{d}\theta$, which can be analyzed or sampled to explore. The central idea of nonparametric bootstrap is the characterization of the distribution of the unobservable deviation $\hat{\theta} - \theta_0$ with the following distribution $\hat{\theta}^*(D_N^*) - \hat{\theta}(D_N)$, where the data set D_N^* of N samples (the boostrap replicate) is drawn uniformly from the observed D_N with replacement. That is, given a fixed sample D_N we define a bootstrap sample distribution over the finite (!) number of possible data sets D_N^*, which allows the assessment of the accuracy of the estimate $\hat{\theta}(D_N)$ by the distribution of $\hat{\theta}^*(D_N^*)$. In general, the bootstrap for $\hat{\theta}(D_N)$ is called consistent if

$$p(\hat{\theta}^*(D_N^*) - \hat{\theta}(D_N)) \to p(\hat{\theta}(D_N) - \theta_0) \text{ as } N \to \infty \text{ in distribution.} \tag{2.27}$$

For example, the *ideal (nonparametric) bootstrap estimate* of the variance $\mathrm{var}_{p(D_N)}(\hat{\theta}(D_N))$ is defined as $\mathrm{var}_{p(D_N^*)}(\hat{\theta}^*(D_N^*))$ (see [48]), which can be shown to provide a consistent estimate [48]. Because of the large number of bootstrap data sets with size N, the ideal bootstrap estimate is approximated by its Monte Carlo estimate using B number of randomly drawn bootstrap data sets $D_{b,N}^*$ for $b = 1, \ldots, B$ and the corresponding quantities $\hat{\theta}_b^*(D_{b,N}^*)$ as follows

$$\hat{\mathrm{var}}_B(\hat{\theta}^*) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_{(.)}^*)^2/(B-1) \text{ where } \hat{\theta}_{(.)}^* = \sum_{\mathrm{i}=1}^{\mathrm{B}} \hat{\theta}_{\mathrm{b}}^*/\mathrm{B}. \tag{2.28}$$

The Monte Carlo estimate of the ideal bootstrap estimate itself has a variance, which is asymptotically $c_1/N^2 + c_2/NB$, so relatively low number of bootstrap replicates suffices in practice [48]. This also indicate that the distribution of $\hat{\theta}_b^*(D_{b,N}^*)$ is more spread than of the target $\hat{\theta}(D_N)$, so it cannot be used directly (e.g., for constructing quantiles for $\hat{\theta}(D_N)$).

Now we can turn to the application of the bootstrap to induce confidence measures for model structures and its properties. This is not without problems as its first application in the model space of phylogenetic trees has shown (phylogenetic trees represent evolutionary relationships between entities corresponding to its nodes [46]). We will follow the terminology and explanations

from that field [77, 49, 13, 47, 4]. Assume that the i.i.d. data set D_N is generated from an unknown Bayesian network model $M_0 = (G_0, \underline{\theta}_0)$ and a fixed algorithm $\mathcal{C}$ induces the model structure $\hat{G}_{\mathcal{C}}(D_N)$, more exactly our hypothesis space are the observation equivalence classes of DAGs $G^{\sim}$. Because the estimate is a model structure without a semantic metric, we cannot define confidence intervals for models with an accuracy parameter, so the probably approximately correct (PAC) terminology is only partly applicable [140]. This frequentist definition of a confidence value is the probability of exact model induction with data sets of size N

$$p(D_N : \hat{G}^{\sim}_{\mathcal{C}}(D_N) = G^{\sim}_0 | M_0, N). \tag{2.29}$$

The essence of the argument for the assessment of Eq. 2.29 with bootstrap is as follows (adapted for discrete valued Bayesian network learning from [49, 47, 48]). By assuming the naive table representation with $d = \prod_i |X_i|$ entries we can interpret a complete (!) data set as corresponding empirical relative frequencies for the complete configurations denoted with $\underline{\hat{\theta}}$, which geometrically is located on the d-dimensional simplex. Note that for a fixed size N, this determines both Bayesian network learning scores, so we can write $\hat{G}^{\sim}(\underline{\hat{\theta}})$. Disregarding that this statistical estimate changes non-continuously across boundaries, it looks like a standard bootstrap problem to assess its accuracy, that is to estimate the probability that $\underline{\hat{\theta}}$ is in the same region as $\underline{\theta}_0$ (i.e., $\hat{G}^{\sim}(\underline{\hat{\theta}}) = G^{\sim}_0(\underline{\theta}_0)$). For a given fixed data set D_N and corresponding $\underline{\hat{\theta}}$, this is estimated using the bootstrap frequencies $\underline{\hat{\theta}}^*$, similarly to the standard case when the distribution of $\hat{\theta}(D_N) - \theta_0$ is assessed with the distribution of $\hat{\theta}^*(D^*_N) - \hat{\theta}(D_N)$. So the probability of exact model induction for an induced model $\hat{G}^{\sim}_{\mathcal{C}}(D_N)$ given a data set D_N theoretically can be characterized with the bootstrap probability and approximated with its Monte Carlo estimate

$$p(D^*_N : \hat{G}^{\sim}_{\mathcal{C}}(D^*_N) = \hat{G}^{\sim}_{\mathcal{C}}(D_N) | D_N) \approx \frac{1}{B} \sum_{b=1}^{B} 1(\hat{G}^{\sim}_{\mathcal{C}}(D^*_{b,N}) = \hat{G}^{\sim}_{\mathcal{C}}(D_N)). \tag{2.30}$$

For phylogenetic trees with model structure T it is shown that the posterior for $T = \hat{T}$ using uninformative prior is nearly equal to the bootstrap probability for $\hat{T}^* = \hat{T}$ (called the "poor man's" Bayes posterior [70]).

We can proceed analogously for the structural features for Bayesian networks. The ideal confidence value is the probability of the induction of the structural feature of the underlying essential graph $F(G_0) = f_0$ with data set of size N [54, 53]

$$p(D_N : F(\hat{G}^{\sim}_{\mathcal{C}}(D_N)) = f_0 | M_0, N). \tag{2.31}$$

This quantity is called "accuracy" in phylogenetics [77, 49]. As noted in [13, 49], this concept is still applicable for a non consistent induction algorithm widely used in a domain as indicating non-repeatability by the lack of support from a well-accepted method. With consistent structure learning algorithms as in the case of Bayesian networks (see Th. 1.5.3), this value will converge to 1 with increasing N. Though the theoretical background for the application

of bootstrap is still unsolved, because of the discrete valued estimate and the consistency properties of the induction algorithm $\mathcal{C}$, in empirical experiments the bootstrap probabilities of features were adopted as assessing the confidence values for features in the induced model $F(\hat{G}^{\sim}_{\mathcal{C}}(D_N)) = f_{D_N}$ given a data set D_N [54, 53].

$$p(D^*_N : F(\hat{G}^{\sim}_{\mathcal{C}}(D^*_N)) = f_{D_N}|D_N). \tag{2.32}$$

This is also backed by the arguments for phylogenetic trees. The bootstrap probabilities are approximated with their Monte Carlo estimates,

$$\frac{1}{B}\sum_{b=1}^{B} 1(\hat{G}^{\sim}_{\mathcal{C}}(D^*_{b,N}) = f_{D_N}), \tag{2.33}$$

with Monte Carlo variance rapidly decreasing with B, as mentioned above. However, the variation of the bootstrap probabilities depending on D_N in case of phylogenetic trees led to the concept of "repeatability" and its classical investigations empirically [77] and analytically [49]. An important clarification of a potential misuse of bootstrap was that the quantity

$$p(D^*_N : F(\hat{G}^{\sim}_{\mathcal{C}}(D^*_N)) = f_0|D_N) \tag{2.34}$$

is not approximating the accuracy (i.e., the probability of induction of "true" features in Eq. 2.31). As suggested [49], a bootstrap probability p for an induced feature can be interpreted as a 1-p-value for the hypothesis that the feature is not present. For phylogenetic trees, a (computationally intensive) correction of the bootstrap probability for its use in the standard hypothesis testing framework is suggested in [47].

For Bayesian networks the bootstrap approach was applied for the following structural features: compelled edges $CompE(X_i, X_j|G)$ (as direct causal relation), Markov blanket membership $\mathrm{MBM}(X_i, X_j|G)$ (as pairwise relevance), pairwise precedence $X_i \prec_G X_j$ (as causal relation) [54, 53] (see results for partly parametric features [117]). The bootstrap probabilities in Eq. 2.32 for the features were interpreted as "support from a given algorithm" [54] and later in testing various induction algorithms as the assessment of the confidence of the induced feature as defined in Eq. 2.31. The experiments were conducted on a gold standard model as reference, which allowed the generation of multiple data sets for proper evaluation of the bootstrap, and on data sets from a genomic and text domain as well.

In summary, earlier works provided an empirical support for the applicability of the bootstrap for Bayesian network features with the following conclusions [54, 53]. It yields a cautious, conservative estimate (no false-positive error) for the features, but its applicability seems sensitive to the domain (e.g., the selection of a confidence threshold for reporting the features), and to the optimization algorithm. Certain pairwise features can be more reliably estimated, especially the pairwise Markov blanket relation (MBM), which can be explained by the topological robustness of this feature (i.e., a given relation can occur in large number of DAGs). The induced confidence measures were reported visually

(as colors and thickness of the Bayesian network edges) to support efficient interpretation of the result of statistical inference from small amounts of data with large number of variables. Another use of the induced confidence measure also gave promising results, to support the second-phase learning of full Bayesian network models and subnetworks using the feature confidences as soft and hard constraints.

However, the relation of the bootstrap approach to the Bayesian approach is subtle w.r.t. the induced confidence measures for Bayesian network features, despite that under specific conditions the bootstrap probabilities approximate the corresponding posteriors [46, 47]. The Bayesian approach is capable to provide updated beliefs — the posterior — for an arbitrary fixed structural feature $F(G) = f_0$ given the observations D_N, either by Monte Carlo sampling or sometimes analytically. This posterior practically can be approximated by the set of models $\mathcal{G}_{\mathcal{C}}^{\text{HPD}}$ with high posteriors identified using an optimization algorithm $\mathcal{C}$ with some heuristic randomization (to correct its bias for local minima):

$$\begin{aligned} p(F(G) = f_0|D_N) &= \sum_G 1(F(G) = f_0)p(G|D_N) \\ &\approx \frac{1}{\sum_{G\in\mathcal{G}_{\mathcal{C}}^{\text{HPD}}} p(G|D_N)} \sum_{G\in\mathcal{G}_{\mathcal{C}}^{\text{HPD}}} 1(F(G) = f_0). \end{aligned} \tag{2.35}$$

Conversely, the bootstrap distribution can be used to assess the accuracy through the defined confidence expressing the effect of the sampling distribution on its identification. That is, the bootstrap approach can provide confidence values for features in the frequentist, hypothesis testing framework by the bootstrap probabilities (i.e., by its Monte Carlo estimates):

$$\begin{aligned} p(F(\hat{G}_{\mathcal{C}}^{\sim}(D_N)) = f_0|M_0, N) &\approx p(F(\hat{G}_{\mathcal{C}}^{\sim}(D_N^*)) = f_{D_N}|D_N) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} 1(\hat{G}_{\mathcal{C}}^{\sim}(D_{b,N}^*) = f_{D_N}). \end{aligned} \tag{2.36}$$

Indeed, as the similarity of the final sums suggests in Eq. 2.35 and Eq. 2.36, the bootstrap can be conceived as a heuristic method using perturbed data sets to generate a good subset of models $\mathcal{G}_{\mathcal{C}}^{\text{HPD}}$ with high posteriors around the maximum a posteriori or maximum likelihood Bayesian network structure. But it cannot be used in general as an approximation to the sampling distribution $p(D_N|M_0, N)$, consequently to sample $p(\hat{G}_{\mathcal{C}}(D_N))|M_0, N)$ or to approximate the posterior $p(G|D_N)$, particularly not in the small sample case, which is the primary goal of learning Bayesian network features.

Furthermore, as the learning of Bayesian network is NP-hard, the computational complexity of the heuristic algorithms used in practice is comparable to the computational complexity of the application of Monte Carlo methods for Bayesian networks with computationally efficient sampling. In fact, after the

investigation of the bootstrap approach [54, 53], the authors also reported an efficient Bayesian approach for inducing Bayesian confidence measures for certain Bayesian network features, which is applied in this thesis and described in the next section.

<RELEVANT

2.4 On the advantage of feature posteriors

After the overview of BN features, certain frequentist identification methods, and the bootstrap methodology to induce confidence measures, we now turn to the Bayesian approach.

The main disadvantage of the frequentist identification methods is that the significance level, if there is any or which in principle what could be derived with general aggregation methods of significances, is not model-based. Furthermore, the methods are fragmented by the type of the features (i.e., there are dedicated algorithms for the identification of local causal features ($RCEdge(X, Y)$), relevant variables and their subsets ($\mathrm{MB}(X), \mathrm{MBM}(X, Y)$) or subtheories ($G' \subseteq G$)).

The bootstrap methodology provides a model-based confidence value, its asymptotic behavior for increasing sample size is guaranteed with a consistent induction algorithm, although there are no theoretical results for its application on structural features for small sample size and it can be applied uniformly for arbitrary features. Furthermore, as it includes a model identification for each bootstrap replicates, its computational complexity can be considerable (e.g., compared to Bayesian Monte Carlo methods).

The introduction of Dirichlet parameter priors with parameter independence for Bayesian networks by Spiegelhalter et al. [131] (conjugate for the multinomial sampling, see Sections 1.2.1.2) provided an efficiently computable closed form for the posterior for the parental sets and for the structure conditional on a given ordering. Based on this, in the beginning of the 1990's the full Bayesian approach was proposed and advocated in a seminal paper [17]. In this paper Buntine proposed the posterior knowledge base view and analysis of the properties of the Bayesian network model conditioned on a given ordering. He also developed a construction method of an approximate posterior offline knowledge base to support theory (i.e., prior) refinement and full scale Bayesian inference. In [29], Cooper et al. discussed the general use of the posterior over Bayesian network structures as an inductive probabilistic knowledge base (i.e., to compute the posterior of arbitrary model properties). However this work had not proposed method to carry out the Bayesian inference. In [100], Madigan et al. proposed an MCMC scheme to approximate such Bayesian inference using the space of DAGs and PDAGS (utilizing also the orderings of the variables). They also developed the Ockham window algorithm for the construction of a small, selective set of models to support exploration of the posterior and inference with it. In [76], Heckerman considered the application of this full Bayesian approach to causal Bayesian networks (under the Causal Markov Condition).

The DAG-based MCMC method was improved by Castelo et al. [65]. In [36], Dash et al. reported a method to perform exact full Bayesian inference in a restricted case of naive Bayesian classifiers. In [55, 56], Friedman et al. reported another MCMC scheme utilizing the ordering of the variables (hence its name, ordering-based MCMC method), which used a closed form for the ordering-conditional posterior of Markov blanket membership, beside the earlier closed form for parental membership. In [87], Koivisto et al. reported a method to perform exact full Bayesian inference over modular features in $\mathcal{O}(n2^n)$ time. Note that the treatment of the submodels as independent hypotheses differs from our approach, which treats them as aggregates of compatible complete models. It would include the assumption of the existential uncertainty of the domain objects represented by the random variables (for the discussion of treating orderings as sets of compatible DAGs or as separate objects, see 2.5.2.2).

Before discussing these methods and their application for complex features, first we summarize the properties of the Bayesian approach and open issues.

1. *Normativity.* The Bayesian approach is a normative, model-based combination of prior and data, so the inputs and the outputs are probabilities conditional on the observed data, which are applicable in the Bayesian decision-theoretic framework. Consequently, its application and interpretation in the small sample region is unconstrained and general results about the behavior of Bayesian inference for large sample size can be applied as well.

2. *Probabilistic knowledge base.* The feature posteriors can be embedded into a probabilistic knowledge base, possibly with textual enrichment as in the case of ABN-KBs. More generally, the feature posteriors can provide the elementary building blocks for a probabilistic semantic web. An important question particularly for complex features is the efficient or approximate representation of the distribution over the feature space.

3. *Probabilistically linked model spaces and induced priors.* The feature posteriors can be used to induce priors for linked model spaces. For classifiers, see Chapter **??** and for the comparison of learning dual-Bayesian networks and the two-phased learning of Bayesian networks from literature data and clinical data, see Section **??**).

4. *Optimally selected feature complexity.* The induced posteriors for the features are dependent in general. A solution followed in the thesis is the definition of a semantically important complex feature, (i.e., subtheory), which includes many dependent simpler features and estimate its posterior distributions.

5. *Integrated estimate and search method.* An already investigated and solved question is the estimation of a moderate number of posterior values (expectations) (e.g., pairwise features such as edge relation or Markov blanket membership with $\mathcal{O}(n^2)$ cardinality). However, the number of values of a

complex feature can be exponentially large (e.g., the number of Markov blanket subsets is $\mathcal{O}(2^n)$), so search methods have to be integrated into the Monte Carlo inference methods to find feature values with relevant posterior. We will see in Section 2.6 and **??** that this issue is related to the estimation of the whole distribution over the feature values and the creation of an offline approximation for it.

FULLVERSION>

6. *Computational uniformity.* The Monte Carlo methods can be applied uniformly for arbitrary features with the same convergence diagnostics and confidence computations, though feature specific optimizations are possible.

7. *Computational complexity.* The used Monte Carlo methods on the orderings with restriction K on the maximal parental set size has $\mathcal{O}(n^{K+1})$ space complexity and the same time complexity for each sampling (typically for $N \approx 10^4$). <FULLVERSION

2.5 MC methods for a feature posterior

As we discussed in Section 2.1, there are two approaches to the use of BN features. The first approach (reported in [17, 102, 74, 54, 53, 55, 56]) is based on a set of simple features to construct a fragmentary representation for the distribution over the complete domain model from multiple, though simple aspects using various interdependent marginals, such as edge probabilities. The other approach is based on a complex feature (or subtheory), which is a focused representation from a restricted, but still comprehensive point of view. In our case, this is the MBG feature to support classifier construction. FULLVERSION>The general assumption and the final goal are the same in both approaches, such as the assumption of reduced statistical and computational complexity of the estimation of a feature w.r.t. and the support of human interpretation and later phases of learning. <FULLVERSION

In both cases we have to use Monte Carlo methods to perform the Bayesian inference, because of the lack of analytical formulas for the posterior of the features. So first, we summarize MC methods: the most direct DAG/PDAG-MCMC method and a latter developed method, the so-called ordering-based MCMC method to estimate the posterior of a limited number of features.

2.5.1 The DAG-based MCMC methods

The basic task is the estimation of the expectation of a given random variable $F(G)$ over the space of DAGs with a specified confidence level.

$$\hat{F} \approx \bar{F} = \mathrm{E}_{p(G|D_N)}[F(G)]. \tag{2.37}$$

In Eq. 1.76), we derived an efficiently computable closed formula for the (unnormalized) posterior of DAGs or for PDAGs in case of likelihood equivalent priors and our standard assumptions, such as complete data, discrete domain variables, multinomial local conditional distributions and Dirichlet priors at the parametric level. As the posterior over DAGs cannot be sampled directly in general and the construction of an approximating distribution to use in importance sampling is frequently not feasible, the standard approach is to use MCMC methods, such as the Metropolis-Hastings algorithm over the DAG or PDAG space (see Section **??**)FULLVERSION> with standard convergence diagnostic and confidence estimation (see Section **??** and **??**<FULLVERSION.

The first application of *DAG-based MCMC* methods for BN feature estimated the posterior of compelled edges [100]. It investigated two proposal distributions. The first constructs a candidate by perturbing directly the edges with insertions, deletions and reversals. The second constructs a candidate by perturbing the partial ordering of the variables and then perturbing the edges to be compatible with this candidate ordering.

2.5.2 The ordering-based MCMC methods

The DAG-based MCMC method for estimating a given expectation is generally applicable, but its statistical properties frequently can be improved by specializing it to a certain type of features. In this section we consider the *ordering-based MCMC* method, which is a hierarchic, semi-analytic MCMC method [55]. We shall see in Section 2.7 that this method can be utilized also to integrate the estimation and the search process in the case of large numbers of features.

2.5.2.1 The ordering-conditional feature posteriors

Assuming modular structure priors, parameter independence, and modularity and complete data, the structure posterior has the following product form:

$$p(G, D_N) = \prod_i^n p(D_N | \operatorname{pa}(X_i, G)) p(\operatorname{pa}(X_i, G)).$$

The ordering-based MCMC method relies on the following two uses of this product form [17, 36, 55]. First, we note that the set of DAGs compatible with an ordering $\prec$ can be constructed as the Descartes product of sets of parental sets compatible with the ordering, so combining this with the product form of

the probability of DAG G we have

$$\begin{aligned} p(D_N|\prec) &= \sum_{G\in\mathcal{G}^{k(n),\prec}} p(D_N,G|\prec) \\ &= \sum_{G\in\mathcal{G}^{k(n),\prec}} \prod_i^n p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec) \\ &= \prod_i^n \sum_{\operatorname{pa}(X_i,G)\sim\prec} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec), \end{aligned} \tag{2.38}$$

where $\operatorname{pa}(X_i,G)\sim\prec$ denotes the compatibility of a parental set $\operatorname{pa}(X_i,G)$ with ordering $\prec$. Second, for an *ordering-modular feature* $F(G)=f$ defined as $\bigwedge_1^n C_i(f,\prec,\operatorname{pa}(X_i,G))$, where C_i is true for some parental sets possibly conditionally on a given ordering $\prec$, we have

$$\begin{aligned} p(f,D_N|\prec) &= \sum_{\substack{G\in\mathcal{G}^{k(n),\prec}\\ F(G)=f}} p(D_N,G|\prec) \\ &= \sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ F(G)=f}} \prod_i^n p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec) \\ &= \prod_i^n \sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ C_i(f,\prec,\operatorname{pa}(X_i,G))}} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec). \end{aligned} \tag{2.39}$$

This gives the following proposition (the generalization of Th. 2.2.1).

Proposition 2.5.1. *For an ordering-modular feature $F(G)=f$ defined as $\bigwedge_1^n C_i(f,\prec,\operatorname{pa}(X_i,G))$, the ordering conditional posterior is decomposed as*

$$\begin{aligned} p(f|D_N,\prec) &= \frac{p(f,D_N|\prec)}{p(D_N|\prec)} \\ &= \prod_i^n \frac{\sum_{\substack{\operatorname{pa}(X_i,G)\sim\prec\\ C_i(f,\prec,\operatorname{pa}(X_i,G))}} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G),f|\prec)}{\sum_{\operatorname{pa}(X_i,G)\sim\prec} p(D_N|\operatorname{pa}(X_i,G))p(\operatorname{pa}(X_i,G)|\prec)} \\ &= \prod_i^n p(C_i(f,\prec,\operatorname{pa}(X_i,G))|D_N,\prec). \end{aligned} \tag{2.40}$$

□

The possible special ("complementer") value without such form can be managed by appropriate summations for the other feature values. Note that if the maximum number of parents is bounded by k, then the ordering conditional feature posterior in Eq. 2.40 can be computed in polynomial time $\mathcal{O}(n^{k+1})$ in

contrast to the exponential number of DAGs compatible with an ordering involved in the summations in Eq. 2.38, 2.39 [55].

FULLVERSION> Table **??** reports the number of orderings, DAGs, ordering-compatible DAGs, ordering-compatible DAGs with parental constraints and the total number of parental sets of an ordering-compatible DAG for the number of variables $i = 2, \ldots, 35$. <FULLVERSION

FULLVERSION>

Note that the ordering-modular property can be also defined for the loss functions as well $L(G)$, which allows the same *sum-product replacement* and so the efficient computation of the expected loss in Eq. 2.3.

<FULLVERSION

2.5.2.2 Advantages of ordering-based MCMC

The existence of the unnormalized posterior for the orderings and the normalized ordering-conditional posterior for a feature allows semi-analytic ordering-based MC methods with advantageous properties w.r.t. DAG-based MC methods.

First, consider the statistical effect of using orderings instead of DAGs and ignore the effect of the MC method used. By assuming a binary feature $F(G)$ and using the identity $\mathrm{E}[X] = \mathrm{E}_Y[\mathrm{E}_X[X|Y]]$ the target quantity can be rewritten as

$$\mathrm{E}[F(G)|D_N] = \mathrm{E}_{p(\prec,D_N)}[\mathrm{E}[F(G)|\prec, D_N]], \tag{2.41}$$

where the random variable $p(F(G)|\prec, D_N) = \mathrm{E}[F(G)|\prec, D_N]$ has variance $\mathrm{var}_{p(\prec|D_N)}(\mathrm{E}[F(G)|\prec, D_N])$. We can decompose it as follows, which directly follows from the identity $\mathrm{var}(X) = \mathrm{E}_Y[\mathrm{var}(X|Y)] + \mathrm{var}_Y(\mathrm{E}[X|Y])$ [64].

Proposition 2.5.2. *The variance of a binary feature $F(G)$ $\mathrm{var}_{p(G|D_N)}(F(G))$ using the augmented space of $\mathcal{G} \times \{\prec\}$ with the distribution $p(G|\prec)p(\prec)$ is the sum of its mean variance and the variance of its mean:*

$$\begin{aligned} &\mathrm{var}_{p(G|D_N)}(F(G)) \\ &= \mathrm{E}_{p(\prec|D_N)}[\mathrm{var}(F(G)|\prec, D_N)] + \mathrm{var}_{p(\prec|D_N)}(\mathrm{E}[F(G)|\prec, D_N]). \end{aligned} \tag{2.42}$$

□

Consequently, the availability of the ordering conditional posterior for a feature allows the cancellation of the term $\mathrm{E}_{p(\prec|D_N)}[\mathrm{var}(F(G)|\prec, D_N)]$ in the ordering-based MC approach compared to a DAG-based method with identical DAG posteriors. It can be a significant reduction because of the asymptotic behavior of the two terms. The expected variance of the ordering conditional probability of a feature is the expectation of the variance of a Bernoulli random variable with parameter $p(F(G)|\prec, D_N)$. In contrast, the other term can be close to zero, if the ordering-conditional posterior of a feature has a similar value for the orderings compatible with the essential graph generating the observations.

The decrease of the variance is not simply the consequence of "collapsing" the $\mathcal{G}(n)$ space into the space of orderings with smaller cardinality of $n!$, but of the

augmentated state space with the orderings $\mathcal{G} \times \{\prec\}$ and the analytic marginalization of the ordering conditional DAGs in the case of ordering modular features (for the general effects of hierarchical approaches and collapsing the state space by analytical marginalization, a.k.a. Rao-Blackwellisation, on MC sampling, see [63]). FULLVERSION> Note that the decrease of the variance is not the consequence of "collapsing" the $\mathcal{G}(n)$ space into the space of orderings with cardinality $n!$ by partitioning such that the feature value is constant in each partition and the posterior of the partition is available. $p(\hat{F}) = \mathrm{E}_{p(G|D_N)}[1(G,F)] = sum_G p(G|D_N)1(G,F) = sum_{R_i} p(R_i|D_N)p(F(G)|R_i)$ Though a reduced space can be useful in estimating a HPD region of feature values (see Section 2.8.2). <FULLVERSION

However, note that Proposition 2.5.2 treats the DAG space as part of an extended space and the explicit, autonomous use of the orderings in the joint distribution $p(G|\prec)p(\prec)$ can introduce a bias (cf. the implicit use of the orderings as sets of compatible DAGs with an induced distribution from $p(G)$). If the uniform distribution $p(\prec)$ is used as non-informative, then it has a bias towards DAGs compatible with many orderings. For example the empty graph is $n!$ times more probable than any complete graph. However, this bias is not related to standard measures of model complexity (i.e., to Ockham principle) as the number of compatible orderings is different for observationally equivalent DAGs (e.g., a Markov chain with different, but observationally equivalent orientations, see Example 1.1.2). An interesting direct consequence is the following proposition.

Proposition 2.5.3. *The induced prior $p(G) \propto \sum_{\prec \sim G} p(\prec)$ from a uniform $p(\prec)$ violates the structural prior equivalence (see Section 1.1.5.2.4).* □

A computationally expensive solution to maintain uniformity over the DAGs is to weight the DAGs through $p(G|\prec)$ properly.

Second, let us compare the ordering-based MC method against the DAG-MC method computationally. Assume that the posteriors of the ordering-conditional parental set are available in $\mathcal{O}(1)$ time (they can be precomputed in $\mathcal{O}(Nn^k)$ time and stored in $\mathcal{O}(n^{(k+1)})$ space, which is either directly acceptable or can be significantly decreased by caching only the high-scoring parental sets). Let $P(n)$ denote the time complexity of the drawing a sample or proposal, which is typically $\mathcal{O}(n^2)$, and $F(n)$ the time complexity of the target feature $F(G(n))$, which is $\mathcal{O}(1)$ for edges, $\mathcal{O}(n)$ for the MBM, MBG and MB features. The unnormalized posterior $p(G, D_N)$ can be computed in $\mathcal{O}(n)$ (assuming the precomputation and storage of the local scores). Thus the overall time complexity of one step of DAG-based MC method is $\mathcal{O}(n^2)$. For the ordering-based MC method this is $\mathcal{O}(n^{k+1})$, but it evaluates n^{kn} or $2^{\mathcal{O}(kn\log(n))}$ DAGs in one step.

Furthermore, to perform exact full Bayesian inference over modular features a dynamic programming method can be used over subsets instead of the naive enumeration of the orderings [87]. This method reduces the super-exponential $\mathcal{O}(n!)$ to $\mathcal{O}(n2^n)$ time, but it requires $\mathcal{O}(n2^n)$ space.

FULLVERSION>

Further factors mainly related to the space of orderings are as follows. (1)

In itself, it can support the design of better sampling methods even without a computable ordering-conditional feature posterior, meaning improved mixing coefficient, autocorrelation, and acceptance rate in MCMC methods. For example the work of Madigan is similarly used the augmented space of $\mathcal{G}(n)^k \times \prec (n)$ to define better proposal distributions in an MCMC method or it can be utilized in stratified sampling schemes without the use of a distribution $p(\prec)$ and ordering conditional feature posterior. (2) It can incorporate background information both directly as a prior and in proposal distributions in MCMC methods. (3) The posterior of the ordering $p(\prec |D_N)$ can be expected to be more insensitive to the data set, which expectation can be illustrated by the assumption of independent error terms $\epsilon(G, D_N, N)$ for $p(G|D_N)$ w.r.t. $\mathrm{E}[p(G|D_N)]$ showing a reduced error for $p(\prec |D_N)$ w.r.t. $\mathrm{E}[p(\prec |D_N)]$ with a factor $\sqrt{|\mathcal{G}^{k(n),\prec}|}$

$$p(\prec |D_N) = \sum_{G \in \mathcal{G}^{k(n),\prec}} p(G|D_N) = \sum_{G \in \mathcal{G}^{k(n),\prec}} \mathrm{E}[p(G|D_N)] + \epsilon(G, D_N, N). \quad (2.43)$$

(4) Finally, the computability of $p(f| \prec, D_N)$ allows the domain-specific interpretation (e.g., the sensitivity of the feature posterior over certain important causal orderings) and better algorithms for estimating a HPD region of feature values. Note that the real sample size effects the peakness of the distribution $p(f| \prec, D_N)$, the MC sample size influences the confidence of the estimate for $p(f| \prec, D_N)$.

<FULLVERSION

2.5.2.3 Estimating edge and pairwise relevance

In the proposal of the ordering-based MCMC method and in subsequent applications the setting was the following [55, 56]. The ordering prior $p(\prec)$ was uniform. The ordering-conditional structure prior $p(G| \prec)$ was a modular prior with uniform weights for the size of the parental sets up to a limit k and with uniform weights for the parental sets with a given size. The parameter independence and modularity were assumed, and the $\mathrm{BD_{eu}}$ parameter prior was used. The MCMC method in the ordering space used two kinds of operations in the proposal distribution: the replacement of pairs and the circular (modulo) shifting of the ordering. The number of variables was 35 in a medical domain, 100-1000 in the genetic and text-mining domains. The target features were the edges ($X_i \to X_j$), the pairwise relevance relations ($\mathrm{MBM}(X_i, X_j)$), the pairwise precedence relations ($X_i \prec X_j$) and the pairwise causal relations ($X_i \dashrightarrow X_j$). There is a closed form for the ordering-conditional posterior, except for the existence of a directed path between two nodes. By noting that the edge feature $f_{X_i \to X_j}$ is an ordering-modular feature and for a given ordering only one clause is relevant in Eq. 2.40, its ordering-conditional posterior is as follows:

$$p(f_{X_i \to X_j} | D_N, \prec) \tag{2.44}$$
$$= \frac{\sum_{\substack{X_i \in \mathrm{pa}(X_j,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}{\sum_{\mathrm{pa}(X_j,G) \sim \prec} p(D_N | \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}.$$

The ordering-conditional posterior of the Markov Blanket Membership feature $f_{\mathrm{MBM}(X_i,X_j)}$ given $\prec$ can be derived by noting that for a given $\prec$ the clauses in the conjunctive normal form for the false value are as follows (assuming $X_i \prec X_j$): earlier parental sets are irrelevant (empty for $X_i \prec X_j$), X_i is not parent of X_j (the clause for X_j includes the parental sets without X_i), and there is no common child of X_i and X_j (the clauses for variables after $X_j \prec X_l$ include the parental sets without X_i and X_l)

$$p(f_{\neg f_{\mathrm{MBM}(X_i,X_j)}} | D_N, \prec) \tag{2.45}$$
$$= p(X_i \notin \mathrm{pa}(X_j,G) | D_N, \prec) \prod_{l=j+1}^{n} p(X_i, X_j \notin \mathrm{pa}(X_l,G) | D_N, \prec),$$

where

$$p(X_i \notin \mathrm{pa}(X_j,G) | D_N, \prec) = \frac{\sum_{\substack{X_i \notin \mathrm{pa}(X_j,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}{\sum_{\mathrm{pa}(X_j,G) \sim \prec} p(D_N | \mathrm{pa}(X_j,G)) p(\mathrm{pa}(X_j,G) | \prec)}$$
$$p(X_i, X_j \notin \mathrm{pa}(X_l,G) | D_N, \prec) = \frac{\sum_{\substack{X_i, X_j \notin \mathrm{pa}(X_l,G) \\ \mathrm{pa}(X_j,G) \sim \prec}} p(D_N | \mathrm{pa}(X_l,G)) p(\mathrm{pa}(X_l,G) | \prec)}{\sum_{\mathrm{pa}(X_l,G) \sim \prec} p(D_N | \mathrm{pa}(X,G)) p(\mathrm{pa}(X_l,G) | \prec)}.$$

The summations involve a polynomial number of terms if the parental set is bounded by k. For approximations using a restricted set of parental sets with high probability, see [55].

For the features which are not ordering-modular, such as the existence of a directed path between a given variable pair, a direct sampling method over the ordering-compatible DAGs is possible using that

$$p(f | D_N, \prec) = \mathrm{E}_{p(G^\prec | D_N, \prec)}[f(G^\prec)] \tag{2.46}$$

and because the ordering-conditional posterior of a given structure G (as an ordering-modular feature) is the product of the ordering-conditional posteriors

$$p(G | D_N, \prec) = \prod_{1}^{n} p(\mathrm{pa}(X_i,G) | D_N, \prec). \tag{2.47}$$

FULLVERSION> This hierarchical, two-layered sampling method will not benefit from the decrease of the variance of exact averaging using a closed form,

though the advantages of the ordering space discussed above and the efficiency of direct sampling are still considerable. <FULLVERSION

A related case is if a not ordering-modular feature F' (such as the Markov Blanket feature $\mathrm{MB}(Y,G)$) is completely defined by an ordering-modular feature F (such as the Markov Blanket subgraph feature $\mathrm{MBG}(Y,G)$), because the ordering-conditional posterior $p(F')$ can be approximated by averaging over a set of features with high ordering-conditional posteriors:

$$p(f'|D_N,\prec) = \sum_f 1(F(f)=f')p(f|D_N,\prec) \quad (2.48)$$

$$\approx \sum_f 1(F(f)=f')p(f|D_N,\prec). \quad (2.49)$$

FULLVERSION> Note that whereas the smaller space of the ordering-compatible features $\mathcal{F}'^{\prec}$ can have a beneficial effect on identification of such a HPD region w.r.t. the space of ordering-compatible DAGs $G^{\prec}$, the direct averaging over $\mathcal{F}'^{\prec}$ instead of $G^{\prec}$ (i.e., averaging using the ordering-conditional feature posterior $p(F|\prec,D_N)$ instead of the ordering-conditional posterior for parental sets $p(\mathrm{pa}(X_i)|\prec,D_N)$) has no effect on the efficiency of the MC estimation

$$p(G:F'(G)=f'|D_N,\prec) = p(F:F'(F)=f'|D_N,\prec) \quad (2.50)$$

$$\approx \frac{1}{M}\sum_{\substack{i=1\\ G_i\sim p(G|D_N,\prec)}}^{M} 1(F'(G_i)=f') \quad (2.51)$$

$$\approx \frac{1}{M}\sum_{\substack{i=1\\ f_i\sim p(f|D_N,\prec)}}^{M} 1(F'(f_i)=f'). \quad (2.52)$$

<FULLVERSION

2.6 Decision over features using MC estimates

In the previous overview of estimation methods of the posteriors of pairwise features, we ignored that the estimated feature posteriors are usually used jointly and we simplified the problem to the estimation of a single posterior. However, the number of target features can be as high as 10^4-10^6 features even for a given type of pairwise features and moderate domain complexity with $100-1000$ variables. For complex features the number of feature values is exponential in the number of variables. Such a high number of feature values makes for example the manual analysis of the estimated edge posteriors intractable. It is thus a typical expectation that the MCMC method should estimate the posteriors uniformly well for all the n^2 features or over a predefined set of features rated a priori as highly relevant. Another typical expectation in bioinformatics is that

the estimates allow the correct ranking of the features or at least the selection of the most probable K feature values. These expectations indicate that the problem of the joint usage of the estimated posteriors in case of large number of features requires an additional level of analysis of the overall MCMC process. This analysis should investigate the effects of the large number of estimates on the MCMC convergence, the estimation and the confidence estimation for the estimates as well, particularly w.r.t. the typical joint usages of the estimates in bionformatics, such as exploration, ranking and selection. In a formal approach we will define an additional frequentist decision-theoretic level over the Bayesian layer of posteriors and their MC estimates. We formalize appropriate losses for joint usages of the estimates typical in bioinformatics, and analyze the effect of feature cardinality on the error of selecting the most probable features. Note that this integrated estimation and decision problem (and the subsequently discussed estimation and search problem) is present also at the level of domain values (i.e., if the goal is the selection of the set of the most probable configurations of values of target variables with a given condition, and Monte Carlo methods are applied for the estimates of their joint conditional probability).

2.6.1 The Most Probable Features problem

We consider the case of a single complex feature with set of values $\mathcal{F}$, when the unknown feature posteriors form a single multinomial distribution $\mathcal{P} = p(F|D_N)$. OPTIONAL> The general case means that the unknown parameters describe the joint distribution of a set of features $p(F_1, \ldots, F_L|D_N)$. <OPTIONAL The *decision problem of feature selection* includes the feature posteriors $\mathcal{P}$ as the unknown parameters, the event space consists of M (possibly dependent) samples D'_M given by a MC method $\mathcal{A}$ as a sampling distribution, and the set of actions consists of the report of the estimates and selections of the parameters. The decision rule $\delta(D'_M) = (I, \hat{\mathcal{P}}_M)$ in general can give a binary vector I indicating the selection and a scalar vector $\hat{\mathcal{P}}_M$ containing the estimates $\hat{p}_M(f|D_N)$.

If the overall estimation is important, then general distance measures such as $L_2(\mathcal{P}, \hat{\mathcal{P}}_M)$ can be adopted as loss function. However, frequently the overall estimates or rankings of the feature values are irrelevant and only the selection of feature values with high posteriors is important.

Definition 2.6.1. *The* Most Probable Features *problem (MPFs) consists of the selection of a predefined K number of feature values $f \in \mathcal{F}$ with high posteriors $p(f|D_N)$, which minimize the following loss based only on $I \in \mathcal{I}^K$ ($\mathcal{I}^K$ denotes the set of $|\mathcal{F}|$ dimensional binary vectors with exactly K ones)*

$$L(I) = L(\mathcal{P}, I) = \sum_i I_i L(s_i), \text{ where } L(s_i) = 1 - \mathcal{P}_i. \tag{2.53}$$

Note that the estimates of the selected feature values are secondary and not involved in the loss function, and with this decomposable loss function this problem is not a set selection problem. The Most Probable Features problem

with the Markov Blanket subset feature generalizes the feature subset selection problem and reformulates it in the Bayesian framework. With the Markov Blanket subgraph feature it generalizes and reformulates the feature subgraph selection problem Def. 2.2.3.

FULLVERSION>

By combining the distance approach and the selective approach, we get the loss

$$L_2^K(\mathcal{P},\mathcal{I},\mathcal{Q}) = \begin{cases} \sum_i \mathcal{I}_i((\mathcal{P}_i-\mathcal{Q}_i)^2-\mathcal{P}_i^2)+\mathcal{P}_i^2 \text{ if } \sum_{\mathrm{i}} \mathcal{I}_{\mathrm{i}} = \mathrm{K} \\ \infty \text{ else} \end{cases} \tag{2.54}$$

Another interesting loss is based on threshold t and binary cost matrix C

$$L_{t\leq,C}(\mathcal{P},\mathcal{I}) = \sum_i C_{1(t\leq\mathcal{P}_i),\mathcal{I}_i}, \text{ denoted with } \mathrm{L}_{\mathrm{t}\leqslant}(\mathcal{P},\mathcal{I}) \text{ if } \mathrm{C} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) \tag{2.55}$$

Finally we define a loss in the general case over a set of features $p(F_1,\ldots,F_L|D_N)$. Assuming L binary features, binary actions of reporting I_i and some predefined losses $L_i(p(G|D_N,I_i)$ for each, the overall loss of reporting them is

$$L_L^K(p(F_1,\ldots,F_L|D_N),\mathcal{I}) = \sum_i L_i(p(G|D_N,I_i). \tag{2.56}$$

<FULLVERSION

2.6.2 Effect of feature cardinality in MPFs

First, assume that the MC estimates of the posteriors are available for all the feature values and let us investigate the statistical consequences of using these estimates of the feature posteriors in the most probable feature selection problem with loss Eq. 2.53. That is we neglect momentarily the computational aspects of the search of the most probable features, and the integrated estimate and search problem. Specifically, we investigate the effect of the cardinality of the feature values $|\mathcal{F}|$ on the mean error of the selected set of features.

Theorem 2.6.1 ([108]). *Let us assume that we solve the K Most Probable Features problem in Def. 2.6.1 using an i.i.d. data set D'_M containing M samples from the feature posterior $\mathcal{P} = p(F|D_N)$ and applying the following decision rule $\delta(D'_M) = I^*_M$ defined as $I^*_M = \arg\min_{I\in\mathcal{I}^\kappa} L(\hat{\mathcal{P}}_M, I)$ (i.e., we select the most probable feature values). The loss function is defined in Eq. 2.53. Let $\hat{L}(I), \hat{L}(s_i)$ denote the corresponding estimated losses based on $\hat{\mathcal{P}}_M$, $I^* = \arg\min_{I\in\mathcal{I}^\kappa} L(\mathcal{P}, I)$ denotes an optimal set, and $I^*_M = \arg\min_{I\in\mathcal{I}^\kappa} L(\hat{\mathcal{P}}_M, I)$ denotes an empirically[†] optimal set. The error is defined as $1/K(L(I^*_M) - L(I^*))$. Then the sample complexity and the expected error of the selection of the K most probable features*

[†]We use the empirical term w.r.t. the stochastic simulations as well.

are proportional to the logarithm of the number of feature values $|\mathcal{F}|$*:*

$$p(\frac{1}{K}|L(I_M^*) - L(I^*)| \geq \epsilon) \leq \delta, \text{ if } M \geq 2/\epsilon^2(\log(2|\mathcal{F}|) + \log(1/\delta)), \quad (2.57)$$

$$\mathrm{E}_{p(D'_M)}[\frac{1}{K}(L(I_M^*) - L(I^*))] \leq \sqrt{\frac{\log(2|\mathcal{F}|) + 1}{M/2}}. \quad (2.58)$$

Proof. We proceed analogously as in the case of selecting the best (binary) classifier, in fact we treat each feature value as a classifier and this theorem is the generalization of the earlier results for selecting the single best classifier [42].

$$\begin{aligned}
&\tfrac{1}{K}(L(I_M^*) - L(I^*)) \\
&= \tfrac{1}{K}(L(I_M^*) - \hat{L}(I_M^*) + \underbrace{\hat{L}(I_M^*) - \hat{L}(I^*)}_{\leq 0} + \hat{L}(I^*) - L(I^*)) \\
&\leq \tfrac{1}{K}(L(I_M^*) - \hat{L}(I_M^*) + \hat{L}(I^*) - L(I^*)) \\
&\leq \tfrac{1}{K}|L(I_M^*) - \hat{L}(I_M^*)| + |\hat{L}(I^*) - L(I^*)| \\
&\leq 2 \max_{f \in \mathcal{F}} |p(f|D_N) - \hat{p}_M(f|D_N)|. \qquad (2.59)
\end{aligned}$$

It means that if we can estimate uniformly well the probabilities of the features, then we can bound the error of the selected set of features. Using the Hoeffding inequality [42], we get for ϵ accuracy and δ confidence

$$\begin{aligned}
&p(\frac{1}{K}|L(I_M^*) - L(I^*)| \geq \epsilon) \\
&\leq p(\max_{f \in \mathcal{F}} |p(f|D_N) - \hat{p}_M(f|D_N)| \geq \epsilon/2) \leq 2|\mathcal{F}|e^{-M\epsilon^2/2} \leq \delta,
\end{aligned}$$

which shows that the sample complexity is

$$M \geq 2/\epsilon^2(\log(2|\mathcal{F}|) + \log(1/\delta)). \quad (2.60)$$

Furthermore, the expected average error of the selected set of features can be bounded as follows using the inequality $\mathrm{E}[Z] \leq \sqrt{\frac{\log(ce)}{2M}}$ (which holds if $p(Z \geq \epsilon) \leq ce^{-2M\epsilon^2}$ for all $0 \leq \epsilon$ and some $0 \leq c$) [42]:

$$\mathrm{E}_{p(D'_M)}[1/K|L(I_M^*) - L(I^*)|] \leq \sqrt{\frac{\log(2|\mathcal{F}|) + 1}{M/2}}. \quad (2.61)$$

□

Note that here the cardinality of the set for selection $|\mathcal{F}|$ is independent of the sample size M. Note that the best K-term approximation of $\mathcal{P}$ in L_1 is the K MAP feature posterior represented by I^*.

This result was derived assuming an i.i.d. sample from the feature posterior. Analogic results for estimates based on dependent MCMC samples can be derived using MCMC variants of the Hoeffding inequality (e.g., see [67]).

2.7 Integrating estimation and search of MBGs

Until now we have assumed that the estimates of the posteriors OPTIONAL> or the single-feature scores corresponding to a decision rule <OPTIONAL are available for all the feature values. As discussed below this assumption is implicitly fulfilled by DAG-MC methods, but it is computationally prohibitive for ordering-based MC methods. The DAG-MC methods perform an implicit feature selection by generating a sample $D'_M = G_1, \ldots, G_M$, which can be used to construct a feature-tree containing the maximum M number of distinct feature values usually in $\mathcal{O}(Mn^2)$ time to compute the non-zero single-feature scores in $\mathcal{O}(M)$, and to select the K optimal feature values in $\mathcal{O}(M \log(K))$ time. This total $\mathcal{O}(M(n^2+log(K)))$ time and $\mathcal{O}(Mn^2)$ space complexity is usually acceptable in practice, although the additional costs of confidence estimation methods, the extra cost of achieving convergence for features that are not part of the solution and the space requirement suggest some selection or search method to process only the promising features (ideally only the finally reported K features).

On the contrary, the issue of an integrated feature selection method within the ordering-based MC method is relevant, because an ordering-based MC method does not generate implicitly a feature set, as usually an exponential number of features are compatible with an ordering. The alternative approaches are as follows: (1) we treat estimation embedded in a search method, (2) we perform an implicit estimation by sampling, precomputing, and storing to support the subsequent search, or (3) we perform an integrated estimation and search method. We investigate these options in turn focusing on the MBG feature.

First, we consider the separation of estimation and search (cases (1) and (2)). The time complexity of the computation of the posterior of an ordering $p(\prec |D_N)$ and an ordering-conditional posterior $p(f| \prec, D_N)$ of a modular or ordering-modular feature is $\mathcal{O}(n^{k+1})$, where the effect of the real sample size in computing the likelihood terms for a parental set is $\mathcal{O}(Nk)$. We will assume that this polynomial number of scores for the parental sets (or at least for the high-scoring sets) is cached in $\mathcal{O}(n^{k+1})$ space. We also consider the advantages of precomputing ordering-conditional factors for subsequent feature search. OPTIONAL> For an ordering-modular feature value $F(G) = f$, the set of ordering compatible DAGs having this value for a given ordering $\prec$ is defined as

$$\{G : (F(G) = f) \wedge (G \in G^{\prec})\} = \times_{i=1}^{n} S_i(f, \prec), \tag{2.62}$$

where $S_i(f, \prec)$ is the set of valid parental sets of X_i in feature f given ordering $\prec$. Let $|S_F^{Pa}(\prec, X_i)|$ denote the number of ordering-specific sets of parental sets for variable X_i over all feature values f for a fixed ordering $\prec$ and the total with $|S_F^{Pa}(\prec)|$. <OPTIONAL For example, the sets of parental sets for a fixed ordering and for a given MBG feature value $S_i(f, \prec)$ can be either completely independent of the feature value (i.e., containing all the parental sets compatible with the ordering), completely determined by the MBG value (i.e., containing the parental set specified by it) or they can be dependent on both the ordering and the MBG value. However, this last option means less than n distinct sets of

parental sets for each ordering (despite the exponential number of feature value, see Eq. 2.18). This shows that in the case of MBG feature we can precompute also n ordering-conditional factors with $\mathcal{O}(1)$ computational overhead and store in $\mathcal{O}(Mn)$ space together with the $\mathcal{O}(n^{k+1})$ ordering-free parental scores and M orderings in case (2). If the search process evaluates L number of feature value in cases (1) and (2), the overall time complexities are $\mathcal{O}(LMn^{k+1})$ and $\mathcal{O}(M(n^{k+1}+Ln))$ ($\mathcal{O}(n^{k+1}+n)$ corresponds to a separate ordering-based MCMC step).

Second, now we consider the embedding of search into the estimation to overlap them computationally and to decrease the number of estimated feature values L close to the number of selected feature values K (i.e., case (3)). This is particularly relevant if K is large (i.e., it is in the range of n^k), which is the case if our goal is the construction of an offline knowledge base for exploring the MBG space. Another reason is that features that are not part of the solution cause not only extra computational costs because of the computation of their estimates, but can delay the convergence of the MCMC simulation.

FULLVERSION>

First let us consider a non-iterative heuristic method for the selection and estimation of K optimal feature values for the L^K, L_2^K losses (i.e., for avoiding the estimation of all the features). It relies on the existence of efficient greedy algorithms for finding DAGs/PDAGs with high posteriors, which can be used to identify MAP feature values, though $f^{\mathrm{MAP}} = \arg\max_{f\in\mathcal{F}} p(F(G) = f|D_N) neq F(G^{\mathrm{MAP}})$ particularly in the small sample range. These suggest the following two-phased *Select-Estimate features* strategy. First high-scoring structures $\hat{G}$ are identified using heuristic structure optimization methods with some randomization in their initialization, parameterization or in the data set itself (e.g., in a bootstrap scheme), then the corresponding feature values $F(\hat{G})$ are estimated. However, this two-phased Select-Estimate method has serious limitations, because the selection of the candidate feature set is heuristic without semantic interpretation and it cannot be adapted for general losses.

Second let us consider the iterative selection (i.e., search method) and estimation of K optimal feature values for the L^K, L_2^K losses (i.e., for avoiding the estimation of all the features) using DAG-MC sampling. Note that in this case the computation of the averaged functions for the features f, f' for the sample G_i usually are simple and do not overlap significantly $1(F(G_i) = f), 1(F(G_i) = f')$ nor with the sampling itself. Consequently by storing the relevant part of the sample for the feature F ($D'_M = F(G_1), \ldots, F(G_M)$) in $\mathcal{O}(Mn^2)$ space, any search method can be applied offline after the DAG-MC sampling without extra time costs. Because treating the estimations separately embedded in a search/optimization method is not reasonable, these suggest the following methods: 1, if we process (estimate and evaluate) all the features implicitly selected by the sampling process (see Alg. 1) and 2, if we generate and store an MC-sample to support a subsequent search/optimization (see Alg. 2).

The search is over the indicators $\mathcal{I}_{i+1}^K$, which can be based on the currently estimated feature values $\mathcal{I}_i^K$, their estimates and scores and domain-specific

Algorithm 1 A BN-feature selection/learning using unconstrained estimation

Require: $p(F|\prec, D_N), p(\prec|D_N), K, I_{L_2}^{\mathrm{MAP}}, M, L$
Ensure: optimal I^K
  Precomputation of $p(X_i(D_N)|\,\mathrm{pa}(X_i)(D_N))\forall\, i, |\,\mathrm{pa}(X_i)| < k$
  Initialize DAG-MC sampling
  for $i = 0$ to M **do** {the sampling cycle}
    Draw G_i with DAG-MC
    Update statistics for all (already occurred) feature value
  Select K optimal feature (I^K) from the collected statistics.
  Estimate confidence for I^K.

Algorithm 2 An offline estimate/precompute-then-select BN-feature selection/learning

Require: $p(F|\prec, D_N), p(\prec|D_N), K, I_{L_2}^{\mathrm{MAP}}, M, L$
Ensure: optimal I^K
  Precomputation of $p(X_i(D_N)|\,\mathrm{pa}(X_i)(D_N))\forall\, i, |\,\mathrm{pa}(X_i)| < k$
  Generate ordering-MC sample D'_M with precomputed ordering-free and ordering-specific conditionals
  $S_0 \leftarrow \{\}$ {Initialize processed feature set}
  for $i = 0$ to ∞ **do** {the search cycle}
    Estimate S_i with using D'_M (extending D'_M if necessary)
    S_i=Search(S_{i-1},$p(F|\prec_i, D_N)$,$\underline{\hat{p}}$,$\underline{\hat{\mathrm{var}}}$)

heuristics. Standard choices are deterministic greedy methods and randomized schemes such as the simulated annealing. For example, in this later global optimization method, a proposal distribution selects the next feature set $p(I_i^K|I_{i-1}^K)$ and the difference of the scores $e^{-\frac{\hat{L}_M(I_i^K)-\hat{L}_M(I_{i-1}^K)}{T}}$ determines the acceptance probability through a "temperature" parameter, which is gradually decreased. An additional phase may be necessary if the convergence diagnostic for a newly selected feature or its confidence estimation requires larger MC-sample size, but this can be solved by extending the sample.

Third, let us consider the estimation and selection of K optimal feature values for the L^K losses (i.e., for avoiding the estimation of all the features) using ordering-based MC sampling. Assuming that the ordering-free factors are precomputed and stored in $\mathcal{O}(n^{k+1})$, we identified three options with different space-time complexities: (1) we treat estimation embedded in a search method, (2) we perform an implicit estimation by sampling, precomputing, and storing to support a subsequent search, or (3) we perform an integrated estimation and search method. These correspond to Alg. 3, Alg. 4 and Alg. 5.

Algorithm 3 A BN-feature selection/learning with embedded estimation

Require: $p(F|\prec, D_N), p(\prec|D_N), K, I_{L_2}^{\text{MAP}}, M, L$
Ensure: optimal I^K
  Precomputation of $p(X_i(D_N)|\operatorname{pa}(X_i)(D_N))\forall\, i, |\operatorname{pa}(X_i)| < k$
  Initialize ordering-MC sampling
  $S_0 \leftarrow \{\}$ {Initialize processed feature set}
  for $i = 0$ to M **do** {the search cycle}
    Estimate S_i with ordering-MC
    S_i=Search(S_{i-1},$p(F|\prec_i, D_N)$,$\underline{\hat{p}}$,$\underline{\text{vâr}}$)

Algorithm 4 An offline estimate/precompute-then-select BN-feature selection/learning

Require: $p(F|\prec, D_N), p(\prec|D_N), K, I_{L_2}^{\text{MAP}}, M, L$
Ensure: optimal I^K
  Precomputation of $p(X_i(D_N)|\operatorname{pa}(X_i)(D_N))\forall\, i, |\operatorname{pa}(X_i)| < k$
  Generate ordering-MC sample D'_M with precomputed ordering-free and ordering-specific conditionals
  $S_0 \leftarrow \{\}$ {Initialize processed feature set}
  for $i = 0$ to M **do** {the ordering-MCMC cycle}
    Estimate S_i with using D'_M (extending D'_M if necessary)
    S_i=Search(S_{i-1},$p(F|\prec_i, D_N)$,$\underline{\hat{p}}$,$\underline{\text{vâr}}$)

<FULLVERSION

In such an integrated scheme the search method at step i can be based on the sequentially refined estimates of earlier selected features and on the cur-

rently available ordering-conditional posteriors $p(F|\ \prec_i, D_N)$. By noting that the extra cost of an additional feature statistics collection is negligible (i.e., L can be increased to n^k without having significant effect), a robust strategy applies a search method on $p(F|\ \prec_i, D_N)$ for collecting high-scoring features using constraints from the earlier selected features (e.g., threshold for the score). The selected features are estimated, convergence and confidence quantities are computed (note that automated methods are necessary for convergence diagnostics, such as described in Section **??**). If the number of features grows over a given limit L, then they are pruned to maintain efficiency and space limits. In fact this approach can be conceived as a two phased sample-then-search method with a special search method exploiting the estimation steps and using increasing prefixes of an offline sample to decrease time complexity. FULLVERSION> The main steps of the method is reported in Alg. 5. <FULLVERSION

FULLVERSION>

Algorithm 5 An integrated decision theoretic BN-feature selection and estimation using ordering-based MCMC

Require: $p(F|\ \prec, D_N), p(\prec |D_N), K, I_{L_2}^{\mathrm{MAP}}, M, L$
Ensure: optimal I^K
  Precomputation of $p(X_i(D_N)|\,\mathrm{pa}(X_i)(D_N))\forall\ i, |\,\mathrm{pa}(X_i)| < k$
  $S_0 \leftarrow \{\}$ {Initialize processed feature set}
  for $i = 0$ to M **do** {the ordering-MCMC cycle}
    Draw $\prec_i$
    Precomputation for $p(F|\ \prec, D_N)$
    S'=Search(S_{i-1},$p(F|\ \prec_i, D_N)$,$\underline{\hat{p}}$,$\underline{\hat{\mathrm{var}}}$)
    $S_i += S'$ {store new features}
    Update average, variance for $f \in S_i$ $(\hat{p}(f|D_N), \hat{\mathrm{var}}(p)(f|\ \prec, D_N))$
    [Compute convergence diagnostic parameters for $f \in S_i$]
    if $I_{L_2}^{\mathrm{MAP}}$ **then**
      S_i=PruneMAP(S_i,$\underline{\hat{p}}$,i,M)
    else if $L \leq |S_i|$ **then**
      S_i=PruneL2(S_i,$\underline{\hat{p}}$,$\underline{\hat{\mathrm{var}}}$)
  Select K optimal feature (I^K) from S_M using $\underline{\hat{p}}$,$\underline{\hat{\mathrm{var}}}$,$I_{L_2}^{\mathrm{MAP}}$.
  Estimate confidence for I^K.

<FULLVERSION

The search method for finding high-probability features can be any general search such as the deterministic greedy beam search or just the sampling of the ordering-conditional posterior $p(F|\ \prec_i, D_N)$ in each step i or an overpeaked $p(f|\ \prec_i, D_N)^\alpha$ with $1 < \alpha$. Note that the goals of exploring the space of feature values and estimating their posteriors are distinct for ordering-modular features. This is even so if our goal is to generate a set of feature values with cardinality K approximating $p(F|D_N)$ in L_1, because this corresponds to the K-MAP feature finding-estimation problem with the L^K loss.

To develop better estimate and search methods the following observations and constructs can be exploited. First, the product form of the ordering-conditional posterior of an ordering-modular feature allows a decomposed identification of the feature with maximal posterior for a given ordering $\prec_i$.

Lemma 2.7.1. *For an ordering-modular feature function F the most probable feature value f^* compatible with a given ordering $\prec$ can be found by independent optimizations per variable using the posterior $p(F|\prec, D_N)$.*

Proof. It is the direct consequence of the existence of decomposed ordering conditional posterior

$$f^* = \arg\max_{f\sim\prec} p(f|\prec, D_N) = \arg\max_{f\sim\prec} \prod_{i=1}^{n} p(S_i(f,\prec)|\prec, D_N) \quad (2.63)$$

$$= \prod_{i=1}^{n} \arg\max_{S_i(f,\prec)} p(S_i(f,\prec)|\prec, D_N). \quad (2.64)$$

The possible special ("complementer") value without such form can be managed by appropriate summations per variable. □

Furthermore, this decomposed form allows the sorting of the set of potential parental sets $S_i(F,\prec) = \{S_i(f,\prec) : \forall f \in \mathcal{F}\}$, which allows specialized search techniques in the space of $S_1(F,\prec) \times \ldots \times S_n(F,\prec)$.

Based on this observation we introduce the following concepts.

Definition 2.7.1. *The ordering conditional (truncated) MBG space for variable Y is the most probable subspace of $S_1(\mathrm{MBG}(Y),\prec) \times \ldots \times S_n(\mathrm{MBG}(Y),\prec)$ (the truncation in each dimension and the optional sorting is discussed below).*

An MBG state is represented by an $n' \leq n$ dimensional vector $\underline{s}$, where n' is the number of variables not preceding the target variable Y in the ordering $\prec$:

$$n' = \sum_{i=1}^{n} 1(Y \preceq X_i). \quad (2.65)$$

In each dimension, the range of the values are integers $s_i = 0, \ldots, r_i$ representing either separate parental sets or a special set of parental sets not including the target variable. This special value is present only for variables after the target variable and not for the target variable. So $|S_i(\mathrm{MBG}(Y,G),\prec)|$ is $\mathcal{O}(n^k)$, which implies that f^* in Lemma 2.7.1 from the potentially exponential number of features ($\mathcal{O}(n^{nk})$)) can be found in polynomial time $\mathcal{O}(n^{k+1})$, which drops to $\mathcal{O}(1)$ extra time factor if it is done in parallel with the ordering-based MCMC simulation. The product of the ordering conditional posteriors of the represented sets of parental sets gives the ordering conditional posterior of the represented MBG state. We assume that the conditional posteriors of the represented sets of parental sets are monotone decreasing w.r.t. their indices:

$$\forall s_i < s_i' : p(s_i|D_N,\prec) \geq p(s_i'|D_N,\prec). \quad (2.66)$$

which can be constructed in $\mathcal{O}(n^{k+1}\log(\max_i r_i))$ time under the standard assumption of maximum parental set size smaller than k.

Note that if the not ordering-modular feature F' is an aggregate of an ordering-modular feature F, then the search can be performed in the smaller space of ordering-compatible features instead of the space of ordering-compatible DAGs FULLVERSION> (see Eq. 2.50). <FULLVERSION

Second, in the most probable features problem the loss of the selected features in Eq. 2.53 is a sum of non-negative terms, which allows an exact (!) prefiltering (i.e., thresholds t_i to select only the potentially optimal features). Clearly, it is enough to process features with ordering-conditional posteriors above $\tau = \max\mathrm{K}^{\mathrm{th}}{}_{f\in F}\,\hat{p}_M(f|D_N)$ (where $\max\mathrm{K}^{\mathrm{th}}$ denotes the Kth value in a set in decreasing ordering), because for a feature value f part of the set of K features with maximal MC-estimate

$$\tau \leq \hat{p}_M(f|D_N) = \frac{1}{M}\sum_{i=1}^{M} p(f|\prec_i, D_N) \leq \max_{i=1,\ldots,M} p(f|\prec_i, D_N). \tag{2.67}$$

Because such a threshold τ usually is not available a priori, a sample specific threshold τ_i can be used at sample i as the following lemma shows.

Lemma 2.7.2. *If for all MCMC sample $\prec_i$ $i = 1,\ldots,M$ a feature value f is always below a threshold $\tau_i = \max\mathrm{K}^{\mathrm{th}}{}_{f\in F_i^{\prec}}\, p(f|\prec_i, D_N)/M$, then f cannot be part of the set of K features with maximal MC-estimate, because there are at least K feature with larger estimate.*

$$(\forall_{i=1}^{M}\ \prec_i\colon p(f|\prec_i, D_N) < \tau_i) \Rightarrow (\hat{p}_M(f|\prec, D_N) \leq \max\mathrm{K}^{\mathrm{th}}{}_{f'\in F^{\prec_j}}\, p(f'|D_N))$$

Proof.

$$\begin{aligned}
\hat{p}_M(f|D_N) = \frac{1}{M}\sum_{i=1}^{M} p(f|\prec_i, D_N) &\leq \max_{i=1,\ldots,M} p(f|\prec_i, D_N) \\
&< \max\mathrm{K}^{\mathrm{th}}{}_{f'\in F^{\prec_j}}\, p(f'|\prec_j, D_N)/M \\
&\leq \frac{1}{M}\sum_{i=1}^{M} p(f^*|\prec_i, D_N) = \hat{p}_M(f^*|D_N),
\end{aligned} \tag{2.68}$$

where $j = \mathrm{argmax}_{i=1,\ldots,M}\, p(f|\prec_i, D_N)$ and f^* can be any feature in the set

$$\{f'' \in F^{\prec_j} : \max\mathrm{K}^{\mathrm{th}}{}_{f'\in F^{\prec_j}}\, p(f'|\prec_j, D_N) \leq p(f''|\prec_j, D_N)\}.$$

□

Eq. 2.68 also shows that with small variance $\mathrm{var}_{p(\prec_i|D_N)}(p(f|\prec_i, D_N))$ the threshold factor $\frac{1}{M}$ can be selected in practice to be smaller (i.e., when the maximum value is closer to the mean). Note that this filtering based on thresholds τ or τ_i can be extended to importance sampling in which the estimate is a weighted sum.

This truncation per orderings can be specialized for ordering-modular features to truncation per orderings and variables, because of their decomposed score in Eq. 2.63. In the case of MBG(Y, G) feature, this specialized filtering can guide the truncation of the MBG space as follows. We can apply the thresholds per variable j at step i with a given ordering for limiting the $\mathcal{O}(n^k)$ number of set of parental sets to $r_{i,j}$. Furthermore, these can be sorted, which means an $\mathcal{O}(r_{i,j} \log(r_{i,j}))$ extra time factor if it is done in parallel with the ordering-based MCMC simulation). This allows a uniform-cost search or a cost-limited depth-first search. A corresponding estimation and search algorithm based on the orderings and on the ordering-conditional MBG spaces is reported in Section **??**. Note that the overall accuracy-confidence analysis of an integrated estimation and selection method is the same as discussed in Section 2.6.1 assuming that the optimal feature values are identified and not pruned.

FULLVERSION> The pruning method is fitted to the loss function, for heuristic pruning of offline probabilistic KBs, see [17]; for pruning ordering-conditional DAGs and for pruning to "Ockham windows" for model averaging, see [103]. <FULLVERSION

FULLVERSION>

2.8 Applications of the ordering-conditional estimation[/decision] method

In the thesis the primary goal related to the applications of the estimation[/decision] method is the Bayesian, domain model-based, sequential analysis of conditional features with particular emphasis on the incorporation of prior knowledge. This includes on the one hand the estimation of posteriors of fixed set of structural features such as edges, Markov Blanket Membership, and ABN ordering-modular features. On the other hand it includes the estimation/identification of the most relevant classification oriented, complex feature values and particularly the construction of an offline MBG collection/knowledge base approximating the feature posteriors with fixed cardinality, which can be used for an offline induction of posteriors for conditional features such as MB features, $|TANEdge|$ and ABN features. The idea of offline probabilistic KB has appeared in [17] containing ordering-compatible DAGs, a related concept was the model-averaging using Ockham-window [103].

First we describe the reasons/advantages and applications of the ordering-based MC methods for these fixed sets of features. Second we describe the reasons/advantages and applications of the ordering-based MC based estimation/decision methods for the MBG feature, which can be used to explore the high-scoring MBGs and to construct an offline MBG-based knowledge base/data structure for an offline induction of posteriors for conditional features.

The general conditions of the domain relevant for both directions are the following. The number of variables in the main, clinical domain is 35 and the analysis of the size of the parental sets shows that $k < 5$ is acceptable (see

Table **??**). The $k = 4$ selection gives $n^{k+1} \sim 10^7$ parental sets and a single term can computed in $\mathcal{O}(Nk)$ time, which complexities are acceptable in a standard computational environment, so we always assume complete precomputation and storing of these terms, which technically is implemented as online caching (for using only a smaller set of high-scoring parental sets, see [55]). The prior knowledge from a domain expert contains various partial orderings, such as the four embedded denoted with $\prec_t, \prec_w, \prec_r, prec_h$ allowing $1, 10^5, \ldots$ orderings or the identification of "causes" and "effects" for the central disease variable providing 11! and 13! independent orderings (decreasing $\sim 10^{40}$ orderings to $\sim 10^{17}$) and the orderings of groups of variables. Additionally, both parental and deviation structure priors are available. Finally, the size of the clinical data sets is 782 and $4000, 10000, 40000, 400000$ for the literature data sets, but the intended sequential application of the methods for increasing prefix data sets or using moving windows requires further scalability.

2.8.1 Estimation of simple conditional features

In the case of target features with low cardinality such as edges, Markov Blanket Membership, and ABN ordering-modular features the appropriate ordering-conditional posteriors are computable in $\mathcal{O}(n^{k+1})$ time. Furthermore, because of the common factors with the $p(\prec |D_N)$ computation, assuming that these are stored and available in $\mathcal{O}(1)$, the ordering-conditional posterior for an edge can be computed in $\mathcal{O}(n^k)$ time, for an MBM relation or an ABN ordering-modular feature in $\mathcal{O}(n^k)-\mathcal{O}(n^{k+1})$. The availability of the ordering-conditional posteriors, the acceptability of the limit for the parental set size and the general advantages of the ordering-based MC methods described (with the implied space/time complexities) proposed the usage of the ordering-based MC (ordering-conditional MC) method. Additional factor against a DAG-based or DAG-ordering based method was the existence of prior knowledge on the orderings, particularly the existence of small sets of orderings, because these allows exact computations (i.e., exhaustive summation without MC part). Consequently, we used the ordering-conditional approach with the following ordering generation methods to compute/approximate expectations: exhaustive/sparse-enumeration with hard/logical prior, importance sampling with uniform soft prior and hard/logical prior and MCMC sampling with uniform and informative prior. In each of these methods the caching of the parental set scores through the complete computation and the caching of the common factors for a given ordering are always present.

The ordering-conditional enumeration method recursively enumerates all the orderings compatible with a partial ordering defined by a given DAG G(n) and a DAG G^c over the classes of nodes, and exactly compute the expectation $p(F|D_N, \prec (G)) = \mathrm{E}_{p(\prec|D_N, \prec(G))}[p(F|D_N, \prec)]$. The sparse-enumeration method heuristically evaluates only each Mth ordering.

The ordering-conditional importance sampling samples the orderings compatible with a partial ordering defined by a given DAG G(n) and a DAG G^c over the classes of nodes using a uniform distribution (the combination with the

earlier deterministic method leads to a stratified sampling scheme).

The advantage of these algorithms that they can exploit the logical constraints on the ordering (and implicitly on the DAGs), whereas the ordering-conditional MCMC needs irreducibility and aperiodicity, which is hard to guarantee if the space of orderings contains distantly separated small regions of allowed orderings. However the ordering-conditional MCMC method can incorporate "soft" prior information either semantically as prior or computationally as its proposal distribution (e.g., independence sampler with the prior), which is harder/not possible for the earlier methods.

Note that these algorithms can be applied for a fixed set of MBG values as well, but not for MB values (see Eq. 2.70 for the ordering-conditional posterior for an mbg value).

2.8.2 Estimations/decisions over complex conditional features

In the case of simple features related to conditional modeling such as the edge or MBM pairwise feature and complex features such as Markov Blanket membership, interaction substructures, or Markov Blanket subgraphs, we discussed that the MBG feature is complete. This shows the importance of the exploration of high-scoring MBGs and furthermore completeness means the sufficiency to induce a joint distribution for an arbitary set of so-called "classification" features from $p(MBG|D_N)$ or from its K cardinality L_1 approximation $\hat{p}^K_{L_1}(MBG|D_N)$ as

$$\hat{p}(F_1 = f_1, \ldots, F_L = f_L|D_N) \approx \sum_{\text{mbg}} \hat{p}^K_{L_1}(\text{mbg}\,|D_N) 1(F_1(\text{mbg}) = f_1), \ldots, 1(F_L(\text{mbg}) = f_L). \tag{2.69}$$

Such an approximation from an offline $\hat{p}^K_{L_1}(MBG|D_N)$ can be particularly useful in case of multiple, sequential queries, such as the series of ABN-queries in exploring a domain. The estimation/decision method for complex features applied for the MBG feature can serve both purposes, because its ordering-conditional posteriors is computable in $\mathcal{O}(n^k) - \mathcal{O}(n^{k+1})$ assuming that the parental set scores and the common factors with the $p(\prec |D_N)$ computation are stored in $\mathcal{O}(n^{k+1})$,$\mathcal{O}(n)$ space and available in $\mathcal{O}(1)$. As earlier for simple features, the acceptability of the limit for the parental set size and the general advantages of the ordering-based MC methods proposed the usage of the ordering-conditional method, particularly the existence of prior knowledge on the orderings, as small sets of highly relevant orderings.

Because of the exponential number of feature values a search method has to be applied either subsequently-iteratively or in an integrated fashion. The first approach requires the offline storage of the orderings in D'_M, the storage of $p(|\operatorname{pa}(X_i)| \leq k| \prec_l)$ for $X_i \prec_l Y$ and the storage of $p(|Y \notin \operatorname{pa}(X_i)| \leq k| \prec_l)$ for $Y \prec_l X_i$, in $2nM$ space in total, which allows the $\mathcal{O}(nM)$ time computation of the approximation of a given $p(\text{mbg}\,|D_N)$. Because of the product form of the ordering-conditional posterior and the sum form of its approximation,

special search methods are possible, which are better suited to the integrated method. Furthermore, the convergence diagnostics can be better incorporated in the integrated method and its space complexity is smaller without convergence diagnostics and confidence estimation. The two main ingredients are the ordering generation method and a search method.

The goal of the search method is the generation of MBGs with high ordering-conditional posterior, potentially using the already generated MBGs and the posteriors $p(MBG|\prec_l, D_N)$ at step l. We experimented with the following three search algorithms using only the posteriors $p(MBG|\prec_l, D_N)$: direct sampling, top-sampling and a uniform-cost search. In each of these methods we can define a state space for an ordering $\prec$ with the coordinates $S_(^{Pa}Y, \text{mbg}), \ldots, S_(^{Pa}X_{\prec[n]}, \text{mbg})$, where X_i with $Y \prec_l X_i$ has a special value the set of parental sets $Y \notin \text{pa}(X_i)$, the rest of the values are singular sets. The corresponding scores for the values are the ordering-conditional posteriors of the sets of parental sets and the scores for the states (i.e., for the MBGs) are the ordering-conditional MBG posteriors. The direct sampling uses the cached $p(S_(^{Pa}X_{\prec[n]})|\prec, D_N)$ $\mathcal{O}(n^{k-1})$ posteriors, but it could equally use the $\mathcal{O}(n^k)$ cached parental sets $p(\text{pa}(X_i)|\prec, D_N)$. The top-sampling method is biased towards sampling MBGs with high ordering-conditional posterior, by sampling only from the K (typically 2-3) most probable sets of parental sets for each $Y \precsim X_i$ (i.e., from $\text{maxK}^{\text{th}}_{s_(^{Pa}X_i)} p(s_(^{Pa}X_i)|\prec, D_N)$). The uniform-cost search first finds the conditionally MAP MBG (in $\mathcal{O}(n^{k+1})$ time), the K (typically 2-3) most probable sets of parental sets for each $Y \precsim X_i$, and then performs a uniform-cost search to a maximum number of MBGs or to threshold $p(MBG^{MAP,\prec}|\prec, D_N)/M'$, where M' combines the estimated number of orderings and the target number of selected MBGs according to Section **??** (in case of deterministic summation or in importance sampling the weight of the sample contributes as well appropriately).

The estimated MBGs are stored in an n-depth tree with branching on level i by the parental set of X_i, where the indices are sorted lists. The MBGs generated at step l can be found in roughly $\mathcal{O}(n\log(\sqrt[n]{K}))$ assuming a balanced tree with K MBG leaves or inserted in $\mathcal{O}(n\sqrt[n]{K}\log(\sqrt[n]{K}))$ time if not present.

The estimates of the L_l number of MBGs in the updated MBG-tree are all updated with the appropriate ordering-conditional posteriors in $\mathcal{O}(L_l n)$ time. Note that not only the currently generated MBGs are updated, because it would cause an underestimation bias. The exact conditionals allow the computation of the ordering-conditional probabilities of being generated $p(\text{mbg} : \text{Search} - \text{generatedatstepl}|\ \prec_\text{l}, \text{D}_\text{N})$ and being in the tree $p(\text{mbg} \in MBG - tree_l|\prec_l, D_N)$ (i.e., estimated) at step l, which can be used to parameterize the method even run-time.

The orderings are generated with exhaustive/sparse-enumeration with hard/logical prior, importance sampling with uniform soft prior and hard/logical prior and MCMC sampling with uniform and informative prior. In both case the the orderings and the common factors are stored for the analysis. Subsequently, we assume L^K loss (i.e., the goal is the selection of K MAP MBGs).

The integrated, ordering-conditional MBG selection and estimation method

using a deterministic ordering-generation recursively enumerates all or Mth orderings compatible with a partial ordering defined by a given DAG G(n) and a DAG G^c over the classes of nodes.

The stochastic version of the earlier uses importance sampling to sample the orderings compatible with a partial ordering defined by a given DAG G(n) and a DAG

G^c over the classes of nodes using a uniform distribution.

As in the case of simple features, the advantage of these algorithms that they can exploit the logical constraints on the ordering (and implicitly on the DAGs), whereas the ordering-conditional MCMC needs irreducibility and aperiodicity, which is hard to guarantee if the space of orderings contains distantly separated small regions of allowed orderings. However the ordering-conditional MCMC method can incorporate "soft" prior information either semantically as prior or computationally as its proposal distribution (e.g., independence sampler with the prior), which is harder/not possible for the earlier methods.

In the thesis derived classification related features are approximated using such offline MGB collections. However, these methods can be easily specialized for a given conditional features, such as the traditionally important complex, non-ordering-modular MB feature. The ordering generation method is unchanged and the MBG generation heuristic search methods can be applied as heuristic search methods for MB values. However, the method has to be expanded for estimating the ordering-conditional posterior for promising feature values. For example, this high-scoring MBG set S_l selected at step l can be used to induce an approximation as

$$p(\mathrm{mb} \mid \prec_l, D_N) \approx \sum_{\mathrm{mbg} \in S_l} p(\mathrm{mb} \mid \prec_l, D_N) 1(\mathrm{MB}(\mathrm{mbg}) = \mathrm{mb}) \tag{2.70}$$

or its normalized ordering-conditional to avoid underestimation, because of the incompleteness of S_l (i.e., using p' that $1 = \sum_{\mathrm{mbg} \in S_l} p'(\mathrm{mb} \mid \prec_l, D_N)$). Another approach is to introduce an extra inner cycle for direct sampling the ordering-compatible DAGs or MBGs for estimating the ordering-conditional MB posteriors as

$$p(\mathrm{mb} \mid \prec_l, D_N) \approx \frac{1}{M'} \sum_{i=1}^{M'} 1(\mathrm{MB}(\mathrm{mbg}_i) = \mathrm{mb}) \tag{2.71}$$

or using a single cycle for the search and the estimation with the disadvantage that these are joined and cannot be specialied separately for example by using the ordering-conditional MBG posterior and its product form in the search and in the estimation.

<FULLVERSION

Bibliography

[1] B. Abramson and K.-C. Ng. Towards an art and science of knowledge engineering. *IEEE Transactions on Knowledge and Data Engineering*, 5(4):705–711, 1993.

[2] S. Acid and L. M. de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445490, 2003.

[3] S. Acid, L. M. de Campos, and J. G. Castellano. Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.

[4] M. E. Alfaro, S. Zoller, and F. Lutzon. Bayes or bootstrap? a simulation study comparin the performance of bayesian mcmc sampling and bootstrapping in assesing phylogenetic confidence. *Mol. Biol. Evol.*, 20(2):255–266, 2003.

[5] C.F. Aliferis, I. Tsamardinos, and A. Statnikov. Large-scale feature selection using markov blanket induction for the prediction of protein-drug binding, 2003.

[6] T. Van Allen, R. Greiner, and P. Hooper. Bayesian Error-Bars for belief net inference. In *Proc. of the 17th Conf. on Uncertainty in Artificial Intelligence (UAI-2001)*, pages 522–529. Morgan Kaufmann, 2001.

[7] J. R. Anderson. *Rules of the mind*. Lawrence Erlbaum Associates, Hillsdale,NJ, 1993.

[8] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004.

[9] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[10] P. Antal and A. Millinghoffer. A probabilistic knowledge base using annotated bayesian network features. In *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, pages 1–12, 2005.

[11] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote. Bayesian networks in ovarian cancer diagnosis: Potential and limitations. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med. Sys. (CBMS-2000)*, pages 103–109, 2000.

[12] J. M. Bernardo. *Bayesian Theory*. Wiley & Sons, Chichester, 1995.

[13] V. Berry and O. Gascuel. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.*, 13(7):999–1011, 1996.

[14] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[15] R. R. Bouckaert. *Bayesian Belief Networks: From construction to inference*. Ph.D. Thesis, Dept. of Comp. Sci., Utrecht University, Netherlands, 1995.

[16] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks. In Eric Horvitz and Finn V. Jensen, editors, *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 115–123. Morgan Kaufmann, 1996.

[17] W. L. Buntine. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.

[18] R. Castelo and A. Siebes. Priors on network structures. biasing the search for Bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

[19] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. *Expert systems and probabilistic network models*. Springer, Berlin, 1997.

[20] P. Cheeseman. In defense of probability. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 1002–1009. Morgan Kaufmann, 1985.

[21] J. Cheng, D. A. Bell, and W. Liu. Learning belief networks from data: an information theory based approach. In *Proc. of the 6th ACM International Conference on Information and Knowledge Management, CIKM'97*, pages 325–331, 1997.

[22] J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI'99)*, pages 101–107. Morgan Kaufmann, 1999.

[23] J. Cheng and R. Greiner. Learning Bayesian belief network classifiers: Algorithms and system. *Lecture Notes in Computer Science*, 2056:141–151, 2001.

[24] D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proc. of 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 87–98. Morgan Kaufmann, 1995.

[25] D. M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks: Search methods and experimental results. In *Proceedings of Fifth Conference on Artificial Intelligence and Statistics*, pages 112–128, 1995.

[26] G. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 2:203–224, 1997.

[27] G. F. Cooper. The computational complexity of probabilistic inference using bayesian belief network. *Artificial Intelligence*, 42:393–405, 1990.

[28] G. F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.

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

[30] G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 116–125. Morgan Kaufmann, 1999.

[31] V. Coupe, L. van der Gaag, and J. Habbema. Sensitivity analysis: an aid for belief-network quantification. *Knowledge Engineering Review*, 15:1–18, 2000.

[32] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2001.

[33] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, New York, 1999.

[34] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

[35] S. Dasgupta. The sample complexity of learning fixed-structure Bayesian networks. *Machine Learning*, 29:165–180, 1997.

[36] D. Dash and G. F. Cooper. Exact model averaging with naive bayesian classifiers. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 91–98, 2002.

[37] R. Davis, H. Schrobe, and P. Szolovits. What is knowledge representation? *AI Magazine*, 14(1):17–33, 1993.

[38] A. P. Dawid. Conditional independence in statistitical theory. *J. of the Royal Statistical Soc. Ser.B*, 41:1–31, 1979.

[39] A. P. Dawid. Discussion of 'causal diagrams for empirical research' by j. pearl. *Biometrika*, 82(4):689690, 1995.

[40] F. T. de Dombal, D. J. Leaper, J. C. Horrocks, and J. R. Staniland. Human and computer-aided diagnosis of abdominal pain. *British Medical Journal*, 1:376–380, 1974.

[41] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley & Sons, 2002.

[42] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, Berlin, 1996.

[43] M. J. Druzdzel, A. Oniśko, D. Schwartz, J. N. Dowling, and H. Wasyluk. Knowledge engineering for very large decision-analytic medical models. In *Proc. of the 1999 Annual Meeting of the American Medical Informatics Association (AMIA-99)*, page 1049, Washington, D.C., November 6-10 1999.

[44] M. J. Druzdzel and H. Simon. Causality in bayesian belief networks. In David Heckerman and Abe Mamdani, editors, *Proceedings of the 9th Conf. on Uncertainty in Artificial Intelligence (UAI-1993)*, pages 3–11. Morgan Kaufmann, 1993.

[45] M. J. Druzdzel and L. C. van der Gaag. Building probabilistic networks: 'where do the numbers come from? *IEEE Trans. on Knowledge and Data Engineering*, 12(4):481–486, 2000.

[46] R. Durbin, S. R. Eddy, and A. Krogh anf G. Mitchison. *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Chapman & Hall, London, 1995.

[47] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence level for phylogenetic trees. *Proc. Natl. Acad. Sci.*, 93:13429–34, 1996.

[48] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, 1993.

[49] J. Felsenstein and H. Kishino. Is there something wrong with bootstrap on phylogenies? *Syst. Biol.*, 42(2):193–200, 1993.

[50] N. Friedman. The bayesian structural EM algorithm. In *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence(UAI-1998)*, pages 129–138. Morgan Kaufmann, 1998.

[51] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1997.

[52] N. Friedman and M. Goldszmidt. Learning Bayesian networks with local structure. In Eric Horvitz and Finn V. Jensen, editors, *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 252–262. Morgan Kaufmann, 1996.

[53] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A Bootstrap approach. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence(UAI-1999)*, pages 196–205. Morgan Kaufmann, 1999.

[54] N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced bayesian networks. In *AI&STAT VII.*, 1999.

[55] N. Friedman and D. Koller. Being Bayesian about network structure. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence(UAI-2000)*, pages 201–211. Morgan Kaufmann, 2000.

[56] N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

[57] N. Friedman and Z. Yakhini. On the sample complexity of learning Bayesian networks. In *Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, pages 274–282. Morgan Kaufmann, 1996.

[58] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[59] D. Geiger and D. Heckerman. A characterization of the Dirichlet distribution with application to learning Bayesian networks. In Philippe Besnard, Steve Hanks, Philippe Besnard, and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 196–207. Morgan Kaufmann, 1995.

[60] D. Geiger and D. Heckerman. A characterization of the Dirichlet distribution with application to learning Bayesian networks. In Philippe Besnard and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 196–207. Morgan Kaufmann, 1995.

[61] D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82:45–74, 1996.

[62] D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characteriration of several probability distributions. *The Annals of Statistics*, 30(2):216–225, 2002.

[63] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[64] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[65] P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.

[66] C. Glymour and G. F. Cooper. *Computation, Causation, and Discovery*. AAAI Press, 1999.

[67] P. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics and Probability Letters*, 56:143–146, 2002.

[68] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[69] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 17(2):37–43, 2002.

[70] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data minig inference and prediction*. Springer-Verlag, 2001.

[71] D. Haussler. Bounds on the sample complexity of Bayesian learning using information theory and the vc dimension. *Machine Learning*, 14:83–113, 1994.

[72] D. Heckerman and J. S. Breese. Causal independence for probability assesment and inference using bayesian networks. *IEEE, Systems, Man, and Cybernetics*, 26:826–831, 1996.

[73] D. Heckerman and D. Geiger. Likelihoods and parameter priors for Bayesian networks, 1995. Tech. Rep. MSR-TR-95-54, MicroSoft Research.

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

[75] D. Heckermann. A tutorial on learning with Bayesian networks., 1995. Technical Report, MSR-TR-95-06.

[76] D. Heckermann, C. Meek, and G. Cooper. A bayesian aproach to causal discovery. Technical Report, MSR-TR-97-05, 1997.

[77] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogeneteci analysis. *Syst. Biol.*, 42(2):182–192, 1993.

[78] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley & Sons, Chichester, 2000.

[79] C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. In *International Journal of Approximate Reasoning*, volume 15, pages 225–263. Elsevier Science Inc., 1996.

[80] I. Inza, P. Larranaga, and B. Sierra. Bayesian networks for feature subset selection. In *Proceedings of the Workshop on Bayesian and Causal Networks (CaNew2000), ECAI2000*, pages 143–164, 1997.

[81] Manfred Jaeger. Relational bayesian networks. *Proc. of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 266–273, 1997.

[82] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the feature subset selection problem. In *Proc. of the 11th International Conference on Machine Learning*, volume 97, pages 121–129. Morgan Kaufmann, 1994.

[83] D. Kahneman, P. Slovic, and A. Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press, 2001.

[84] E. Keogh and M. Pazzani. Learning the structure of augmented bayesian classifiers. *International Journal of Artificial Intelligence Tools*, 11(4):587–601, 2002.

[85] T. Kocka and R. Castelo. Improved learning of Bayesian networks. In Jack S. Breese and Daphne Koller, editors, *Proc. of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-2001)*, pages 269–276. Morgan Kaufmann, 2001.

[86] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

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

[88] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In Dan Geiger and Prakash P. Shenoy, editors, *Proc. of the 13th Conf. on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 302–313. Morgan Kaufmann, 1997.

[89] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI), Madison, Wisconsin*, pages 580–587, 1998.

[90] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[91] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On supervised selection of bayesian networks. In *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 334–342. Morgan Kaufmann, 1999.

[92] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Urban legends in bayesian network research i: Model selection for supervised problems. *Arpakannus*, 1:8–14, 1999.

[93] P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing prequential model selection criteria in supervised learning of mixture models. In *Proc. of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2001)*, pages 233–238. Morgan Kaufmann, 2001.

[94] W. Lam and F. Bacchus. Using causal information and local measures to learn Bayesian networks. In David Heckerman and Abe Mamdani, editors, *Proc. of the 9th Conference on Uncertainty in Artificial Intelligence (UAI-1993)*, pages 243–250. Morgan Kaufmann, 1993.

[95] P. Larrañaga, C. M. H. Kuijpers, R. H. Murga, and Y. Yurramendi. Learning bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(4):487–493, 1996.

[96] K. Laskey and S. Mahoney. Network fragments: Representing knowledge for constructing probabilistic models. In Dan Geiger and Prakash P. Shenoy, editors, *Proc. of the 13th Conf. on Uncertainty in Artificial Intelligence (UAI-1997)*, pages 334–341. Morgan Kaufmann, 1997.

[97] K.B. Laskey and S.M. Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):487–498, 2000.

[98] S. L. Lauritzen. *Graphical Models*. Oxford, UK, Clarendon, 1996.

[99] T. Y. Leong. Representing context-sensitive knowledge in a network formalism: A preliminary report. In *Proc. of the 8th Conference on Uncertainty in Artificial Intelligence (UAI-1992)*, pages 166–173. Morgan Kaufmann, 1992.

[100] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520, 1996.

[101] D. Madigan, J. Gavrin, and A. E. Raftery. Eliciting prior information to enhance the predictive performance of bayesian graphical models. *Comm.Statist. Theory Methods*, 24:22712292, 1995.

[102] D. Madigan and J.York. Bayesian graphical models for discrete data. *Internat. Statist. Rev.*, 63:215–232, 1995.

[103] D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using occams window. *J. Amer. Statist. Assoc.*, 89:15351546, 1994.

[104] S. Mahoney and K. B. Laskey. Network engineering for complex belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, pages 389–396, 1996.

[105] S. Mani and G. F. Cooper. Causal discovery from medical textual data. In *AMIA Annual Symposium*, pages 542–6, 2000.

[106] Subramani Mani and Gregory F. Cooper. A simulation study of three related causal data mining algorithms. In *International Workshop on Artificial Intelligence and Statistics*, pages 73–80. Morgan Kaufmann, San Francisco, CA, 2001.

[107] C. Meek. Causal inference and causal explanation with background knowledge. In Philippe Besnard, Steve Hanks, Philippe Besnard, and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 403–410. Morgan Kaufmann, 1995.

[108] A. Millinghoffer, G. Hullám, and P. Antal. On inferring the most probable sentences in bayesian logic. In *Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007)*, pages 13–18, 2007.

[109] Stefano Monti and Giuseppe Carenini. Dealing with the expert inconsistency in probability elicitation. *IEEE Trans. on Knowledge and Data Engineering*, 12(4):499–508, 2000.

[110] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, Berlin, 1996.

[111] M. Neil, N. E. Fenton, and L. Nielsen. Building large-scale Bayesian networks. *The Knowledge Engineering Review*, 15(3):257–284, 2000.

[112] D. Nikovski. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):509–516, 2000.

[113] J. Forster P. Dellaportas and I. Ntzoufras. On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12:27–36, 2002.

[114] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

[115] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669710, 1995.

[116] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[117] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics, Proceedings of ISMB 2001*, 17(Suppl. 1):215–224, 2001.

[118] E. Segal D. Peer, A. Regev, D. Koller, and N. Friedman. Learning module networks. In *Proc. of the 19th Conf. on Uncertainty in Artificial Intelligence (UAI-2003)*, pages 525–534. Morgan Kaufmann, 2003.

[119] M. Pradhan, G. Provan, B. Middleton, and M. Henrion. Knowledge engineering for large belief networks. In *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence*, pages 484–490, 1994.

[120] G. M. Provan and M. Singh. Learning bayesian networks using feature selection. In *Proc. of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 450–456, 1995.

[121] S. Renooij, S. Renooij, and L. van der Gaag. Context-specific sign-propagation in qualitative probabilistic networks. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 667–672, 2000.

[122] A. Rényi. *Probability Theory*. Akadémiai Kiad, Budapest, 1970.

[123] A. Rosenberg. *Philosophy of Science: A contemporary introduction*. Routledge, 2000.

[124] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[125] D. Rusakov and D. Geiger. On parameter priors for discrete dag models, 2000.

[126] S. Russel and P. Norvig. *Artificial Intelligence*. Prentice Hall, 2001.

[127] Sumit Sarkar and Ishwar Murthy. Constructing efficient belief network structures with expert provided information. *Knowledge and Data Engineering*, 8(1):134–143, 1996.

[128] E. Segal, M. Schapira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34(2):166–76, 2003.

[129] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2/3):163–192, 2000.

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

[131] D. J. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

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

[133] P. Spirtes and G. Cooper. An experiment in causal discovery using a pneumonia database, 1998.

[134] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2001.

[135] S. Srinivas, S. Russell, and A. Agogino. Automated construction of sparse Bayesian networks for unstructured probabilistic models and domain information. In *Proc. of the 5th Conference on Uncertainty in Artificial Intelligence (UAI-1990)*, pages 295–308. North-Holland, 1990.

[136] D. Subramanian, R. Greiner, and J. Pearl. The relevance of relevance. *Artificial Intelligence*, 97:l–5, 1997.

[137] A. Tanay and R. Shamir. Computational expansion of genetic networks. *Proc. of Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'01)*, 17(Suppl. 1):270–278, 2001.

[138] I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.

[139] I. Tsamardinos, C.F. Aliferis, and A. Statnikov. Algorithms for large-scale local causal discovery and feature selection in the presence of limited sample or large causal neighbourhoods. In *The 16th International FLAIRS Conference*, 2003.

[140] L. Valiant. A theory of the learnable. *Comm. of the ACM*, 27:1134–1142, 1984.

[141] L. van der Gaag, S. Renooij, C. Witteman, B. Aleman, and B. Taal. How to elicit many probabilities. In Kathryn Blackmond Laskey and Henri Prade, editors, *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence (UAI-1999)*, pages 647–654. Morgan Kaufmann, 1999.

[142] T. Verma and J. Pearl. *Causal Networks: Semantics and Expressiveness*, volume 4, pages 69–76. Elsevier, 1988.

[143] T. Verma and J. Pearl. *Equivalence and synthesis of causal models*, volume 6, pages 255–68. Elsevier, 1990.

[144] G. Vita'nyi and K. Li. *An introduction to Kolmogorov complexity and it applicationes*. Springer-Verlag, New York, 1990.

[145] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.

[146] J. Williamson. *Foundations for Bayesian networks*, pages 11–140. Kluwer Academic Publ., 2001.

[147] J. Woodward. Scientific explanation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, 2003.

[148] M. Woodward. *Epidemiology: Study design and data analysis*. Chapman&Hall, 1999.

[149] S. Wright. Correlation and causation. *J. of Agricultural Research*, 20:557–585, 1921.

[150] Changwon Yoo and Gregory F. Cooper. An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *Artificial Intelligence in Medicine*, 31:169–182, 2004.

[151] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

Name: Péter Antal Date of birth: 07-08-1971

1995: M.Sc. in Computer Science (Informatics Engineer), 1995, Faculty of Electrical Engineering and Informatics, Technical University of Budapest

1995-1998: PH.D. studies on the informatics Ph.D. programme of Faculty of Electrical Engineering and Informatics at the Department of Measurement and Information Systems, Technical University of Budapest.

1998 - 1999: International scholar, Department of Electrical Engineering, Katholieke Universiteit Leuven.

2000-2002: Studying for Ph.D. at the Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium. Research topic: Combination of prior domain knowledge and data in statistical learning methods.