Bayesian networks

An overview

P.Antal

antal@mit.bme.hu

BME

Historical overview I.

1920 The investigation of graphical models for probabilistic causal models goes back to 1920 in the work of Wright on path diagrams **?**.

1970 The first (medical) applications of a special class of Bayesian networks as a probabilistic expert system, including knowledge elicitation and learning appeared in 1970 **?**.

1988 A later large scale and commercial application was reported in 1988 **?**.

1979 The axiomatic investigation of the structure of independencies in a probability distribution was reported in 1979 **?**

1988 The issue of representability with DAGs in 1988 **?**.

1982 The decomposition of a probability distribution using annotated DAGs was reported in 1982 (for a general treatment of graph based decomposition see **?**).

1989 The decomposed representation of Bayesian networks has appeared in 1989 **??**, though first related to representing contextual independencies. Later extensions related to knowledge engineering and attempts to first-order probabilistic logical extension were reported from 1997 **????**.

Historical overview II.

1985< The causal interpretation of Bayesian networks and the related causal research is present from the proposal of the representation **???**, though first seen as auxiliary human constructs and in the probabilistic causation research the goals were to understand the limits of the learnability from observational data and the identifyability of causal effect **???**.

1995< 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 **??**.

1989 A generally applicable inference method (the so called join tree algorithm) in 1989 **?** (An efficient inference method for polytrees has appeared in 1983 **?**).

1989 The Bayesian approach to the parameters using Dirichlet priors was reported in 1989 **?**, a related evaluation methodology based on the prequential framework in 1993 **?**.

1995 The Bayesian approach to parameters was axiomatized in 1995 **?**.

1991 The Bayesian approach to the structure of the model was proposed in 1991 for models that compatible with a fixed causal order of the domain vairables**?**, the general treatment and practical learning was reported in 1992 **?**.

1995 A full-fledged Bayesian approach to perform Bayesian inference over structural properties was reported in 1995 **?** and a large-scale application in 1999 **??**.

Observational (in)dependencies

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

$$I_p(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{Z}) \; iff \; (\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \; p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{y} | \boldsymbol{z}) \; whenever \; p(\boldsymbol{z}, \boldsymbol{x}) > 0) \tag{1}$$

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

$$I_p(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{c}) \; iff \; (\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \; p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{c}) \; whenever \; p(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{x}) > 0) \tag{2}$$

Another notation for $I_p(X; Y | Z)$ is $(X \perp\!\!\!\perp Y | Z)_P$ and dependency is denoted with $D_p(X; Y | Z)$ (or $(X \not\perp\!\!\!\perp Y | Z)_P$). A set of (in)dependence statements is called *(in)dependence model* . A standard measure for the strength of the dependence (association) between X and Y is the (conditional) *mutual information* $MI_p(X; Y | Z) = KL(p(X, Y | Z) | p(X | Z) p(Y | Z))$.

Axioms of independencies

1. Symmetry: The observational probabilistic conditional independence is symmetric.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; iff \; I_p(\boldsymbol{Y};\boldsymbol{X}|\boldsymbol{Z})$$

2. Decomposition: Any part of an irrelevant information is irrelevant.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; and \; I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z})$$

3. Weak union: Irrelevant information remains irrelevant after learning (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}\cup\boldsymbol{W})$$

4. Contraction: Irrelevant information remains irrelevant after forgetting (other) irrelevant information.

$$I_p(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \; and \; I_p(\boldsymbol{X};\boldsymbol{W}|\boldsymbol{Z}\cup\boldsymbol{Y}) \Rightarrow \; I_p(\boldsymbol{X};\boldsymbol{Y}\cup\boldsymbol{W}|\boldsymbol{Z})$$

The model of causal (in)dependencies

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

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

$$CI_p(\boldsymbol{X}; \boldsymbol{Y}|\boldsymbol{Z}) \; iff \; (\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \; p(\boldsymbol{y}|do(\boldsymbol{z}), do(\boldsymbol{x})) = p(\boldsymbol{y}|do(\boldsymbol{z}))) \tag{3}$$

Note that this does no meant to be an exhaustive definition of causation (e.g. the counterfactual aspects remains outside this definition). Note that despite the symmetry of the probabilistic independence relation, the causal independence relation is asymmetric. Another notation for $CI_p(X; Y|Z)$ is $(X \not\Rightarrow Y|Z)_P$. The negated independence proposition (i.e. causal relevance or dependency) is denoted with $CD_p(X; Y|Z)$ (or $(X \Rightarrow Y|Z)_P$). If $\boldsymbol{Z} = \{\boldsymbol{V} \setminus \{\boldsymbol{X}, \boldsymbol{Y}\}\}$, then the causal relevancy/dependency relation is called *direct causal dependency* and denoted with $DCD_p(X; Y|Z)$ (or $(X \rightarrow Y|Z)_P$). A set of causal (in)dependence statements is called *causal model* .

Causation in biomedical sciences

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 attributable risk (θ) and the odds ratio (Ψ).

$$\delta = p(y|do(x)) - p(y|do(\neg x) \tag{4}$$

$$\theta = \frac{p(y|do(x)) - p(y|do(\neg x)}{p(y|do(x))} \tag{5}$$

$$\Psi = \frac{p(y|do(x))/p(\neg y|do(x))}{p(y|do(\neg x))/p(\neg y|do(\neg x))} \tag{6}$$

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

$$\tilde{p}(y|x) = \sum_z p(y|x,z)p(z) \tag{7}$$

Principles of causality

Beside this probabilistic, interventional definition of "X is a cause of Y" based on the $P(.|do())$ semantics other standard conditions are the following (modified from the list of "*principles of causality*" suggested within epidemiology **?**:

1. strong association,
2. X precedes temporally Y,
3. plausible explanation without alternative explanations based on confounding,
4. necessity (generally: if cause is removed, effect is decreased or actually: y would not have been occurred with that much probability if x had not been present),
5. sufficiency (generally: if exposure to cause is increased, effect is increased or actually: y would have been occurred with larger probability if x had been present). The probabilistic definition of causation formalizes many, but for example not the counterfactual aspects.

Furthermore, as shown, 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 to the scientific understanding of "stable and transportable" mechanism.

Goals of the BN representation

- **P** representation for the joint distribution ,
- **I** sound and complete representation for the independency model,
- **P-I** 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 compatible with the list of ”principles of causality",
- **I-C** understanding relation between M and C, i.e. the relation between the observationally defined, symmetric (in)dependence relations and the interventionally defined asymmetric causal relation (particularly the learnability of a causal model from observation data,
- **P-C** understanding 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 $P(y|do(x'), z')$ (i.e. if x' is more appropriate for interventional studies),
- **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 occured if x had not been present conditioned on that x was present and y occurred.

Markov conditions I.

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

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Pa(X_i)), \tag{8}$$

where $Pa(X_i)$ denotes the parents of X_i in G.

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

$$\forall\, i = 1, \ldots, n : (X_{\pi(i)} \perp\!\!\!\perp \{X_{\pi(1)}, \ldots X_{\pi(i-1)}\} / Pa(X_{\pi(i)}) | Pa(X_{\pi(i)}))_P, \tag{9}$$

where $\pi()$ is some ancestral ordering w.r.t. G (i.e. compatible with arrows in G).

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

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp \mathrm{Nondescendants(X_i)} | \mathrm{Pa(X_i)})_\mathrm{P}, \tag{10}$$

where $\mathrm{Nondescendants(X_i)}$ denotes the nondescendants of X_i in G.

Markov conditions II.

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

$$\forall\, X, Y, Z \subseteq U \; (X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_P, \tag{11}$$

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 its descendants of n is in Z.*

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

Connections

1. Theorem. *Let $P(U)$ a probability distribution and G a DAG, then the conditions above (repeated below) 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*.*

Further implied properties by FOLG

8. Definition. *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 non-adjacent in G and $X_j \in$ Nondescendants(X_i), $(X_i \perp\!\!\!\perp X_j | \mathrm{Nondescendants(X_i)} \setminus \{\mathrm{X_j}\})_\mathrm{P}$ holds* **?**.

9. Definition. *A distribution $P(X_1, \ldots, X_n)$ obeys the boundary Markov condition w.r.t. DAG G, if*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp U \setminus Bd(X_i) | bd(X_i, G))_P, \tag{12}$$

*where $bd(X_i, G)$ denotes the set of parents, children and the children's other parents for X_i, i.e. parents with common child with X_i (***?***):*

$$bd(X_i, G) = \{Pa(X_i, G) \cup Ch(X_i, G) \cup Pa(Ch(X_i, G), G)\} \tag{13}$$

Markov blanket and boundary

10. Definition. *A set $MB_P(X_i)$ is called the Markov blanket of X_i w.r.t. the distribution $P(X_1, \ldots, X_n)$, if $(X_i \perp\!\!\!\perp U \setminus MBl(X_i) | MB(X_i))_P$ (the reference to P is usually omitted). A minimal Markov blanket is called Markov boundary .*

That is the (FOLG) conditions for P, G implies that the set $bd(X_i, G)$ is a Markov blanket for X_i, $MB_P(X_i)$ for each variable (they are not necessarily Markov boundaries as may be not minimal, because of G). So we will also refer to $bd(X_i, G)$ as the Markov blanket for X_i in G using the notation $MB(X_i, G)$. The induced (symmetric) pairwise relation $MBM(X_i, X_j, G)$ w.r.t. G between X_i and X_j

$$MBM(X_i, X_j, G) \leftrightarrow X_j \in bd(X_i, G) \tag{14}$$

is called *Markov blanket membership* **?**. In short, these are the unconditional and conditional direct pairwise relevancies (i.e. the non-blockable pairwise (observational) dependencies 1).

11. Definition. *A subgraph of G is called the Markov Blanket (sub)Graph or Mechanism Boundary (sub)Graph $MBG(X_i, G)$ of variable X_i if it includes the nodes in the Markov blanket defined by $bd(X_i, G)$ and the incoming edges into X_i and into its children $Ch(X_i, G)$ (for a full definition of the MBG feature, see Def.**??**).*

Bayesian network defs

12. Definition. *A directed acyclic graph (DAG)* G *is a* Bayesian network *of distribution* $P(U)$ *iff 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*).*

13. Definition. *A Bayesian network model* M *of domain with variables* U *consists of a structure* G *and parameters* $\boldsymbol{\theta}$*. The structure* G *is a DAG such that each node represents a variable and* local probabilistic models $p(X_i|pa(X_i))$ *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* $pa(X_i)$*. As the conditionals are frequently from a certain parametric family, the conditional for* X_i *is parameterized by* $\boldsymbol{\theta_i}$*, and* $\boldsymbol{\theta}$ *denotes the overall parameterezation of the model.*

A calculus for independencies

D-separation provides a sound and complete, computationally efficient algorithm to read off an (in)dependency model consisting the independencies that are valid in all distributions Markov relative to G, that is $\forall\ X, Y, Z \subseteq V$

$$(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow ((X \perp\!\!\!\perp Y|Z)_P \text{ in all P Markov relative to G}). \tag{15}$$

Stable distributions

1. Example.

Consider $p(X,Y,Z)$ with binary X,Z and ternary Y. The conditionals $p(Y|X)$ and $p(Z|Y)$ can be selected such that $p(z|x) = p(z|\neg x)$. That is $(X \not\perp\!\!\!\perp Y)$ and $(Y \not\perp\!\!\!\perp Z)$, but $(X \perp\!\!\!\perp Z)$, demonstrating that the "naturally" expected transitivity of dependency can be destroyed numerically.

Consider $P(X,Y,Z)$ with binary variables, where $p(x) = p(y) = 0.5$ and $p(Z|X,Y) = 1(Z = XOR(X,Y))$. That is $(X \perp\!\!\!\perp Z)$ and $(Y \perp\!\!\!\perp Z)$, but $(\{X,Y\} \not\perp\!\!\!\perp Z)$, demonstrating that pairwise independence does not imply total independence.

However, such numerically encoded independencies correspond to the solution of equation systems and/or to functional dependencies, that is they are not stable for numerical perturbations leading to the following definition.

14. Definition. *The distribution P is said to 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 perfectly represents its (in)dependencies.*

Equivalence classes of BNs

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

The implied equivalence classes may contain $n!$ number of DAGs (e.g. all the full networks representing no independencies) or just 1.

2. Theorem. *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).*

16. Definition. *The essential graph representing observationally equivalent DAGs is a partially oriented DAG (PDAG), that represents the identically oriented edges called compelled edges of the observationally equivalent DAGs (i.e. in the equivalence class), such a way that in the common skeleton only the compelled edges are directed (the others are undirected representing inconclusiveness).*

On the possibility of casual interpretation

The direction of the edges corresponds to the intuitive expectation, for example in an (unconfounded) v-structure $X \rightarrow Y \leftarrow Z$ 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.
Counter-arguments

1. All direct dependency among the constructed domain variables are causal.(?)
2. Stability guaranteeing that a corresponding Bayesian network exactly represents the (in)dependencies.
3. The ”Boolean" Ockham principle, namely that only the minimal, consistent models are relevant, so orientations in the essential graph are determined by the minimal models (DAGs).

The Causal Markov Condition I.

17. Definition. *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 $p(X_i|pa(X_i))$ for each node w.r.t. the structure G describing the causal stochastic dependency of variable X_i on its parents $pa(X_i)$. As the conditionals are frequently from a certain parametric family, the conditional for X_i is parameterized by $\boldsymbol{\theta_i}$, and $\boldsymbol{\theta}$ denotes the overall parameterezation, so a causal model consists of a structure G and parameters $\boldsymbol{\theta}$.*

18. Definition. *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). Consequently, the precondition of the Causal Markov condition for (P, G) that the set of variables V is *causally sufficient* for P, that is all the common causes for the pairs $X, Y \in V$ inside V.

Note, that hidden variables are allowed (as the local probabilistic models are nearly always high level abstractions), only variables that influences two or more variables in V are necessary for causal sufficiency.

(The causal Markov condition implies sufficiency and stability implies necessity of G).

The Causal Markov Condition II.

In the presence of potential common causes (confounders), that is if the Causal Markov Condition is violated, certain causal dependencies can still be identified as the following example shows.

2. Example. *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. 7) resulting in a v-structure. If potential confounders are not excluded a priori, we have to observe at least one more variable 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).*

Functional (causal) Bayesian network

The axiomatic foundation for the graph surgery semantics of the $P(.|do(.),.)$ notation.

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

1. *$p(\boldsymbol{V}|do(\boldsymbol{x})$ is Markov relative to G,*
2. *$\forall\ X_i \in \boldsymbol{X}\ p(x_i|do(\boldsymbol{x}) = 1$ if value x_i and $\boldsymbol{x}$ is compatible,*
3. *$\forall\ X_i \notin \boldsymbol{X}\ p(x_i|pa_i, do(\boldsymbol{x}) = p(x_i|pa_i)$ if value(s) pa_i and $\boldsymbol{x}$ is compatible.*

The relativity of the interpretations

Counter-arguments

1. The presence of unobserved (hidden) variables as potential confounders.
2. Selection bias can occur if the observation depends on the joint combination of otherwise independent events, inducing non-causal dependencies between them.
3. 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).
4. Global physical and semantic constraints between the variables.
5. Stability can be also questioned, because of deterministic dependencies, resulting in the lack of guarantee for the uniqueness and exactness of the representation.
6. The (in)dependencies are relative to the set of variables and specifically, also to the values of the variables

Parameter priors:independence

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

$$P(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i|G), \tag{16}$$

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

$$p(\boldsymbol{\theta}_i|G) = \prod_{j=1}^{q_i} p(\boldsymbol{\theta}_{ij}|G), \tag{17}$$

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

Parameter priors:likelihood equivalence

The concept of likelihood equivalence extends observational equivalence of the structure coherently to the parameters .

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

$$p(\boldsymbol{\theta}_{\boldsymbol{V}}|G_1) = p(\boldsymbol{\theta}_{\boldsymbol{V}}|G_2), \tag{18}$$

where $\boldsymbol{\theta}_{\boldsymbol{V}}$ 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.)

Parameter priors: Dirichlet priors

3. Theorem. *The assumption of positive densities, likelihood equivalence and parameter independence for complete structures G_c implies that $p(\boldsymbol{\theta}_U|\xi)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n}$.*

The $p(\boldsymbol{\theta_i}|G_i) = J_{G_i} p(\boldsymbol{\theta_V}|\xi)$, where J_{G_i} is the Jacobian of the transformation from $\boldsymbol{\theta_V}$ to $\boldsymbol{\theta_{G_i}}$. Remarkable, that a structure level acausal constraint (i.e. likelihood equivalence of structures with multinomial local dependency models) implies a strong parameter-level contraint (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/virtual sample size* and $p^{prior} x_1, \ldots, x_n = E[\theta_{x_1,\ldots,x_n}] = N_{x_1,\ldots,x_n}/N'$. Furthermore, we need the following concept.

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

$$p(\boldsymbol{\theta_{ij}}|G_1) = p(\boldsymbol{\theta_{ij}}|G_2), \tag{19}$$

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

Parameter priors: Dirichlet priors II.

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

4. Theorem. *If $p(\boldsymbol{\theta}_{\boldsymbol{V}}|\xi)$ is a Dirichlet distribution with hyperparameters $N_{x_1,\ldots,x_n} = N'px_1,\ldots,x_n$ and the parameter modularity assumption holds and for all complete network 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*

$$p(\boldsymbol{\theta}|G) \quad = \quad \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta^{N'p^{prior}(X_i=k,pa(X_i,G)=pa_{ij})-1} \tag{20}$$

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

Parameter priors: Dirichlet priors III.

In case of a fixed structure G (or we shall see for a fixed ordering of the variables), the usage of Dirichlets with parameter independence can be attractive on its own right to specify a parameter distribution $p(\boldsymbol{\theta}|G)$ as follows

$$p(\boldsymbol{\theta}|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} Dir(\boldsymbol{\theta_{ij}}|\boldsymbol{N_{ij}}) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta^{N_{ijk}-1} \tag{21}$$

Structure priors

The global noninformative deviation prior **?** is derived from an a priori "reference" network structure G_0 by modeling each missing or extra edge e_{ij} independently with a uniform probability κ:

$$P(G) \propto \kappa^{\delta}, \text{where } \delta = \sum_{1 \leq i < j \leq n} I_{\{(e_{ij} \in G) \wedge (e_{ij} \notin G_0) \vee (e_{ij} \notin G) \wedge (e_{ij} \in G_0)\}}.$$

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 $F_i(G) = f_i \; i = 1, \ldots K$

$$P(G) = c \prod_{i=1}^{K} p(F_i(G)), \tag{22}$$

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

$$P(G) \propto \prod_{i=1}^{n} p(pa(X_i, G)). \tag{23}$$

Inference with BNs

1. Inference over domain values with observations
   - Fixed parameter and fixed structure
   - Bayesian parameter and fixed structure
   - Bayesian parameter and structure

$$p(\boldsymbol{y}|\boldsymbol{x}) = E_{p(G)}[E_{p(\boldsymbol{\theta}|G)}[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, G)]]. \tag{24}$$

2. Inference over domain values with interventionist data
3. Inference over model parameters
4. Inference over model structures

$$p(G|D_N) \propto p(G) \int p(D_N|\boldsymbol{\theta}, G)p(\boldsymbol{\theta}|G)d\boldsymbol{\theta} = p(G)p(D_N|G) \tag{25}$$

5. Inference over model properties and ABN-propositions

$$p(\alpha(G)|ABN - KB) = \sum_{\alpha(G)\ is\ true} p(G) \tag{26}$$

Inference over model parameters

Assuming parameter independence and a complete observation.

$$p(\boldsymbol{\theta}|x) = \prod_{i=1}^{n} p(x_i|pa_i(x), \boldsymbol{\theta}_{ij_0}) p(\theta_{ij_0}) \prod_{j \neq j_0} p(\theta_{ij}) / p(x) \tag{27}$$

$$\propto \prod_{i=1}^{n} p(x_i|pa_i(x), \boldsymbol{\theta}_{i_0 j_0}) p(\theta_{i_0 j_0}) \prod_{j \neq j_0} p(\theta_{ij}) \tag{28}$$

$$\propto \prod_{i=1}^{n} \theta_{ij_0 x_i} Dir(\theta_{ij_0}|\boldsymbol{\alpha}_{ij_0}) \prod_{j \neq j_0} Dir(\theta_{ij}|\boldsymbol{\alpha}_{ij}) \tag{29}$$

$$\propto \prod_{i=1}^{n} Dir(\theta_{ij_0}|\alpha_{ij_0 1}, \ldots, \alpha_{ij_0 x_i} + 1, \ldots, \alpha_{ij_0 r_i}) \prod_{j \neq j_0} Dir(\theta_{ij}|\boldsymbol{\alpha}_{ij}) \tag{30}$$

Inference over model structures I.

By assuming N complete observations, i.i.d. multinomial sampling, Bayesian network model with parameter independence and Dirichlet parameter priors, the observation of a complete case results in a local standard Bayesian updating of the hyperparameters of the appropriate Dirichlets and retains the parameter independence. The maintained parameter independence allows a standard parental decomposition w.r.t. the Bayesian network G for each observation, which allows the following rearrangement

$$p(\boldsymbol{C_1}, \ldots, \boldsymbol{C_N}|G) = \prod_{l=1}^{N} \prod_{i=1}^{n} p_l(x_i^{(l)}|pa_i^{(l)}) \quad (31)$$

$$= \prod_{i=1}^{n} \prod_{l=1}^{N} p_l(x_i^{(l)}|pa_i^{(l)}) \quad (32)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{l=1}^{N} p_l(x_i^{(l)}|pa_{ij})^{1(pa_{ij}=pa_i^{(l)})} \quad (33)$$

where $pa_i^{(l)}$ denotes the value(s) of parental set of X_i in case l.

Inference over model structures II.

This can be combined with the earlier result of the marginal probability of the data for a single Dirichlet prior and multinomial sampling (see Eq. **??**).That is for each variable X_{i_0} and parental configurations j_0 independently

$$\prod_{l=1}^{N} p_l(x_{i_0}^{(l)}|pa_{i_0 j_0}, G)^{1(pa_{i_0 j_0}=pa_{i_0}^{(l)})} = \frac{\prod_{k=1}^{r_{i_0}} \alpha_{i_0 j_0 k}..(\alpha_{i_0 j_0 k}+n_k)}{\alpha_{i_0 j_0 +}\ldots(\alpha_{i_0 j_0 +}+n)} \quad (34)$$

$$= \frac{\Gamma(\alpha_{i_0 j_0 +})}{\Gamma(\alpha_{i_0 j_0 +}+n_{i_0 j_0 +})} \prod_{k=1}^{r_{i_0}} \frac{\Gamma(\alpha_{i_0 j_0 k}+n_{i_0 j_0 k})}{\Gamma(\alpha_{i_0 j_0 k})}$$

where 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. The sign $+$ denotes the appropriate marginals.

Inference over model structures III.

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

$$p(G|D_N) \propto \prod_{i=1}^{n} p(Pa(X_i, G))S(X_i, Pa(X_i, G), D_N) \text{ where} \tag{35}$$

$$S(X_i, Pa(X_i, G), D_N) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})}. \tag{36}$$

E-science era, data and Bayesianism

1. *E*lectronic domain knowledge vs. printed and expert The availability of the semantic web with electronic domain literature and knowledge bases contrary to the earlier case relying exclusively on experts and on printed domain literature.
2. *S*tatistical data vs. test cases The availability of significant amount of statistical data for automated theory refinement contrary to earlier anecdotical cases, test cases.
3. *B*ayesianism vs. understandable, fixed model The availability of Bayesian methods by increased computational power offers a principled method for prior incorporation and theory refinement. It also allows to work with significant number of models contrary to the earlier goal to formalize an understandable consistent knowledge representation of the domain.

Consequences for knowledge engineering

New goals: specify *(*1) indirectly over the electronic resources *(*2) a Bayesian prior knowledge model and *(*3) compute the posterior of complex, semantic statements.

1. *R*eferential, interfacing and supportive knowledge representation for informal knowledge collection. The knowledge engineering process has to provide methods for exploring and collecting the electronic domain knowledge. In other words, the knowledge representation is becoming ”meta" (indirect/referential) is a sense, that it specifies possibly through lengthy computation the construction of the real (prior) knowledge model from electronic resources.
2. *C*onstruction of priors for Bayesian methods. A purpose of knowledge engineering is to represent possibly exhaustively the consistent alternative combinations with beliefs eligible for Bayesian update with the available data. Put it simply, the goal of knowledge representation is to formalize prior(s) and the ”final" knowledge model is provided by the posterior of the Bayesian update.
3. *I*nterpretation and evaluation of posteriors from Bayesian methods. A purpose of knowledge engineering is to provide ”gold standards" for evaluating the posteriors and to provide a semantic context for formulating complex, semantic statements with posteriors for evaluation and interpretation. That is the evaluation of the knowledge representation is performed partly by the data as part of the prior sensitivity analysis supported by complex, semantic probabilistic propositions.

Learning Bayesian networks

A closed-form for the posterior of a structure G,

$$p(G, D_N) = p(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})} \tag{37}$$

termed as *Bayesian Dirichlet* metric. The corresponding score functions are defined as $BD(G; D_N) = \log(p(G, D_N))$.
Another family of non-Bayesian score functions can be derived within the likelihood framework. Assuming a complete, discrete values, i.i.d. data set, let define a maximum likelihood score as follows

$$ML(G; D_N) = max_{\boldsymbol{\theta}} p(D_N | G, \boldsymbol{\theta}) \tag{38}$$

The ML learning

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 $Pa(X_i)$ (N_{ij+} is the appropriate sum) **??**. By substituting this maximum likelihood parameter selection back, we get

$$ML(G; D_N) = p(D_N|G, \boldsymbol{\theta}^*) = \prod_{l=1}^{N} \prod_{i=1}^{n} p(x_i^{(l)}|pa_i^{(l)}) \tag{39}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}}^{N_{ijk}} \tag{40}$$

by taking logarithm, rearranging and expanding with N

$$\log(ML(G; D_N)) = N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \frac{N_{ij+}}{N} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N_{ij+}} \log(N_{ijk}/N_{ij+}) \tag{41}$$

The information theoretic approach

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 $I(Y;X) = H(Y) - H(Y|X)$ **?**, it can be rewritten as

$$\log(ML(G;D_N)) = -N \sum_{i=1}^{n} H(X_i|Pa(X_i,G)) \quad (42)$$

$$= -NH(X_1,\ldots,X_n) \quad (43)$$

$$= N \sum_{i=1}^{n} I(X_i;Pa(X_i,G)) - N \sum_{i=1}^{n} H(X_i) \quad (44)$$

(45)

This shows that the maximization of the maximum likelihood score is equivalent with finding a Bayesian network parameterized with the observed frequencies that has minimum entropy or in other words the implicated coding of the observed cases is minimal (see 43). Another interpretation is that we are finding a Bayesian network parameterized with the observed frequencies that has maximum mutual information between its children and their parents (44, the terms not depending on the structure can be neglected). Note the close connection of this reading to the concept that causal ordering is related to the (maximal) determination of each variable by the earlier variables.

Complexity regularization

Because of the monotonicity of mutual information — if $Pa(X_i) \subset Pa(X_i)'$, then $I(X_i; Pa(X_i) \leq I(X_i; Pa'(X_i)$ **?** — so 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 shall use only the BIC-score defined as follows (for overviews of other score functions and for the derivation of the BIC-score see **?????**)

$$BIC(G; D_N) = \log(ML(G; D_N)) - \frac{1}{2} dim(G) \log(N) \tag{46}$$

where $dim(G)$ denotes the number of free parameters.

Score equivalence

24. Definition. *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.*

5. Theorem. *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 hypotheses are the equivalence classes or the prior is equal for such G_1, G_2, then the BD_e scoring metric is score equivalent* **?**.

Consequently, with BD metrics only the structure prior can incorporate causal information, which means an asymptotically vanishing term w.r.t. the acausal likelihood term (the differentiation within an equivalence class by a none likelihood equivalent BD score can not be used semantically).
The score equivalence of the BIC score is the direct consequence of the result that the number of free parameters (that is the term $dim(G)$) are equal in observationally equivalent Bayesian networks (here again as throughout the thesis, we assume discrete variables and multinomial local dependendency models) **??**.

6. Theorem. *The $BIC(G; D_N)$ scoring metric is score equivalent* **?**.

Asymptotic consistency

7. Theorem. *Let $\boldsymbol{V}$ be a set of variables. Let the prior distribution $p(G)$ over Bayesian network structures be positive. Let $p(\boldsymbol{V})$ be a positive and stable distribution and G_0 is a corresponding perfect map (i.e. a Bayesian network representing exactly all the independencies in $p(\boldsymbol{V})$, see Def.* **??***). Now, let D_N is an i.i.d. data set generated from $p(\boldsymbol{V})$. Then, for any network structure G over $\boldsymbol{V}$ that is not a perfect map of $p(\boldsymbol{V})$ we have that*

$$\lim_{N\to\infty} BD_e(G_0; D_N) - BD_e(G; D_N) = -\infty \; and \; also \tag{47}$$

$$\lim_{N\to\infty} BIC_e(G_0; D_N) - BD_e(G; D_N) = -\infty \tag{48}$$

For further results about the asymptotic optimality of the scores for not stable distributions see **?**.

An asymptotic consistency result for the maximum likelihood based scores is derived in **?**. Furthermore, a rate of convergence result 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

$$p(D_N : KL(p_0|p_{BN}) > \epsilon) < \delta \tag{49}$$

Search algorithms: DAG space I.

The cardinality of the space of DAGs is given by the following recursion **?**

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-1)} f(n-i) \; with \; f(0) = 1. \tag{50}$$

This is bounded above with the number of the combinations of the edges between different nodes ($2^{n(n-1)}$), because of the exclusions by the DAG-constraint. But it is still super-exponential even with a bound k on the maximum number of parents (consider that the number of parental sets for a given ordering of the variables is in the order of n^{kn}, so $2^{\mathcal{O}(kn \log n)}$ **?**.

Search algorithms: DAG space II.

The number of orderings, DAGs and order-compatible DAGs with parental constraints. The columns shows respectively the number variables (nodes) (n), DAGs ($|DAG(n)|$), DAGs compatible with a given ordering ($|G_{\prec}|$), DAGs compatible with a given ordering and with maximum parental set size <=4 ($|G_{\prec}^{|\pi|\leq 4}|$) and <=2 ($|G_{\prec}^{|\pi|\leq 2}|$), the number of orderings (permutations) ($|\prec|$) and the total number of parental sets in an order-compatible DAG $|\pi^{\prec}|$ and in an order-compatible DAG with maximum parental set size <=4 ($||\pi^{\prec}| \leq 4|$) and <=2 ($||\pi^{\prec}| \leq 2|$).

n	$\|DAG(n)\|$	$\|G_{\prec}\|$	$\|G_{\prec}^{\|\pi\|\leq 4}\|$	$\|G_{\prec}^{\|\pi\|\leq 2}\|$	$\|\prec\|$	$\|\pi^{\prec}\|$	$\|\|\pi^{\prec}\| \leq 4\|$
5	2.9e+004	1e+003	1e+003	6.2e+002	1.2e+002	30	30
6	3.8e+006	3.3e+004	3.2e+004	9.9e+003	7.2e+002	62	61
7	1.1e+009	2.1e+006	1.8e+006	2.2e+005	5e+003	1.3e+002	1.2e+002
8	7.8e+011	2.7e+008	1.8e+008	6.3e+006	4e+004	2.5e+002	2.2e+002
9	1.2e+015	6.9e+010	2.9e+010	2.3e+008	3.6e+005	5.1e+002	3.8e+002
10	4.2e+018	3.5e+013	7.5e+012	1.1e+010	3.6e+006	1e+003	6.4e+002
15	2.4e+041	4.1e+031	2.1e+027	3.1e+019	1.3e+012	3.3e+004	4.9e+003
35	2.1e+213	1.3e+179	1.8e+109	8.5e+068	1e+040	3.4e+010	3.8e+005

The complexity of BN learning

The computational complexity of finding a Bayesian network structure best fitting to the observations is bounded by the following two theorems (assuming $P \neq NP$). The first shows the NP-hardness of finding a Bayesian network for the observations (as minimal representation of the observed independencies see Def. 12, which is I-map) **?**.

8. Theorem. *Let $\boldsymbol{V}$ be a set of variables with joint distribution $p(\boldsymbol{V})$. Assume that an oracle is available that reveals in $\mathcal{O}(1)$ time whether an independence statement holds in p (see Def.* **??***). Let $0 < k \leq |\boldsymbol{V}|$ 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 less or equal to s edges by consulting the oracle is NP-complete.*

The second theorem shows the NP-hardness of finding a best scoring Bayesian network (i.e. the NP-hardness of optimization over DAGs) **?**.

9. Theorem. *Let $\boldsymbol{V}$ be a set of variables, D_N is a complete data set, $S(G, D_N)$ is a score function and a real value c. Then, the problem of deciding whether or not there exist a Bayesian network structure G_0 defined over the variables $\boldsymbol{V}$, where each node in G_0 has at most $1 < k$ parents, such that $p \leq S(G_0, D_N)$ is NP-complete.*

Constraint-based BN learning: IC

The Inductive Causation algorithm (assuming a stable distribution P):

1. *Skeleton:* Construct an undirected graph (skeleton), such that variables $X, Y \in \boldsymbol{V}$ are connected with an edge iff $\forall S (X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$.
2. *v-structures:* Orient $X \rightarrow Z \leftarrow Y$ iff X, Y are nonadjacent, Z is a common neighbour and $\neg \exists S$ that $(X \perp\!\!\!\perp Y | S)_P$, where $S \subseteq \boldsymbol{V} \setminus \{X, Y\}$ and $Z \in S$.
3. *propagation:* Orient undirected edges without creating new v-structures and directed cycle.

10. Theorem. *The following four rules are necessary and sufficient.*

R_1 *if* $(a \not\rightarrow c) \wedge (a \rightarrow b) \wedge (b - c)$, *then* $b \rightarrow c$

R_2 *if* $(a \rightarrow c \rightarrow b) \wedge (a - b)$, *then* $a \rightarrow b$

R_3 *if* $(a - b) \wedge (a - c \rightarrow b) \wedge (a - d \rightarrow b) \wedge (c \not\rightarrow d)$, *then* $a \rightarrow b$

R_4 *if* $(a - b) \wedge (a - c \rightarrow d) \wedge (c \rightarrow d \rightarrow b) \wedge (c \not\rightarrow b) \wedge (a - d)$, *then* $a \rightarrow b$

The Bayesian learning:inference

model selection/optimization/search vs. model averaging

$$
\begin{aligned}
p(\alpha(G)|ABN-KB, D_N) &= \sum_G 1(\alpha(G)\ is\ true\ in\ ABN-KB) p(G|D_N) \\
L_{\hat{G}|D_N} &= E_{p(G|D_N)}[L(G,\hat{G})] = \sum_G L(G,\hat{G}) p(G|D_N), \\
p(\boldsymbol{y}|\boldsymbol{x}, D_N) &= E_{p(G|D_N)}[E_{p(\boldsymbol{\theta}|G,D_N)}[p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},G)]].
\end{aligned}
$$

$\Rightarrow$ Bayesian inference with Monte Carlo methods.