On Representation and Learning of Context-Specific Independence

Timo Koski
KTH Royal Institute of Technology

TK

Machine Learning Summer School at University of Reykjavik
04.05.2014

KTH Matematik

Outline of the Lecture

- Context-specific independence (CSI), definition, examples

KTH Matematik

Outline of the Lecture

- Context-specific independence (CSI), definition, examples
- Labeled Directed Acyclic Graphs (LDAG), examples

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of the Lecture

- Context-specific independence (CSI), definition, examples
- Labeled Directed Acyclic Graphs (LDAG), examples
- Equivalence classes of LDAGs

KTH VETENSKAP OCH KONST
KTH Matematik

Outline of the Lecture

- Context-specific independence (CSI), definition, examples
- Labeled Directed Acyclic Graphs (LDAG), examples
- Equivalence classes of LDAGs
- Bayesian Learning of LDAGs

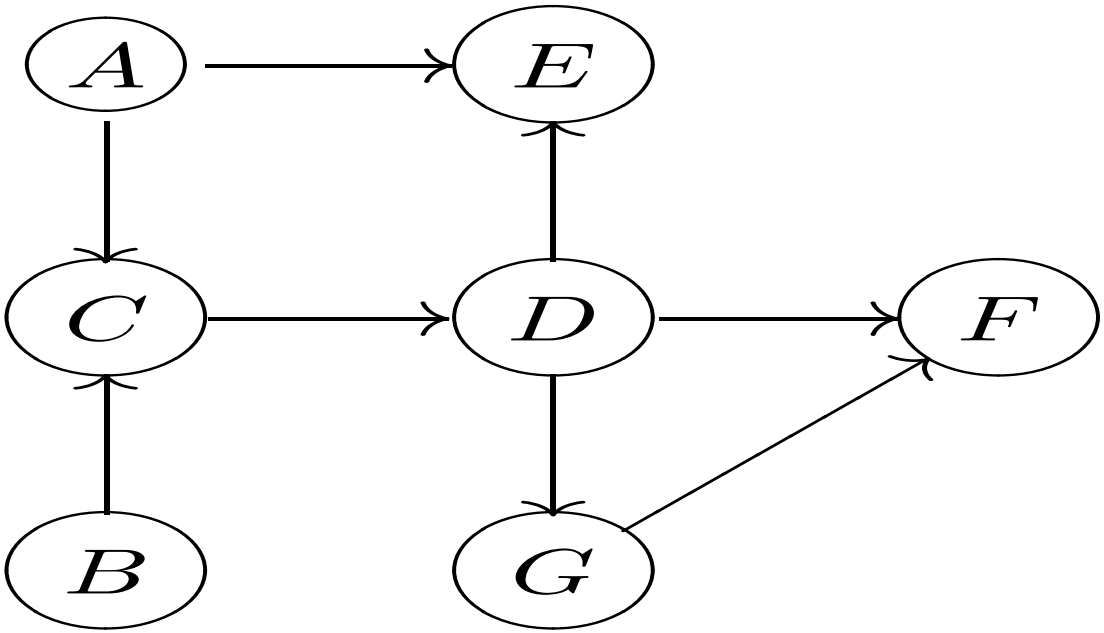
KTH VETENSKAP OCH KONST
KTH Matematik

This is joint work with Jukka Corander (Univ. of Helsinki), Johan Pensar and Henrik Nyman (both at Åbo Academy University).

KTH Matematik

Background

Directed acyclic graphs (DAGs) have gained widespread popularity as representations of complex multivariate systems.



KTH Matematik

Background

Despite their advantageous properties for representing dependencies among variables in a modular fashion, several proposals for making them more flexible and parsimonious have been presented.

KTH VETENSKAP OCH KONST
KTH Matematik

Extending and bringing together of:

- Boutilier, C. et.al.: Context-specific independence in Bayesian networks, *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pp. 115−123, 1996.
- Geiger, D. and Heckerman, D.: Knowledge representation and inference in similarity networks and Bayesian multinets,*Artificial Intelligence*, 82, pp. 45−74, 1996.
- Corander, J.: Labelled graphical models, *Scandinavian journal of statistics*, 30, 3, pp. 493−508, 2003.

KTH VETENSKAP OCH KONST

KTH Matematik

Conditional Independence

A DAG encodes independence statements in the form of conditional independencies.

The textbook definition: Two random variables X and Y are conditionally independent given a third random variable Z if and only if they are independent in their conditional probability distribution given Z. That is, X and Y are conditionally independent given Z if and only if, given any value of Z, the probability distribution of X is the same for all values of Y and the probability distribution of Y is the same for all values of X.

KTH VETENSKAP OCH KONST
KTH Matematik

Conditional Independence (CI)

Two random variables X and Y are conditionally independent (CI) given a third random variable Z if and only if they are independent in their conditional probability distribution given Z. That is, X and Y are conditionally independent given Z if and only if, given any value of Z, the probability distribution of X is the same for all values of Y and the probability distribution of Y is the same for all values of X.

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z),$$

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

$\Leftrightarrow$

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

CI: irrelevance

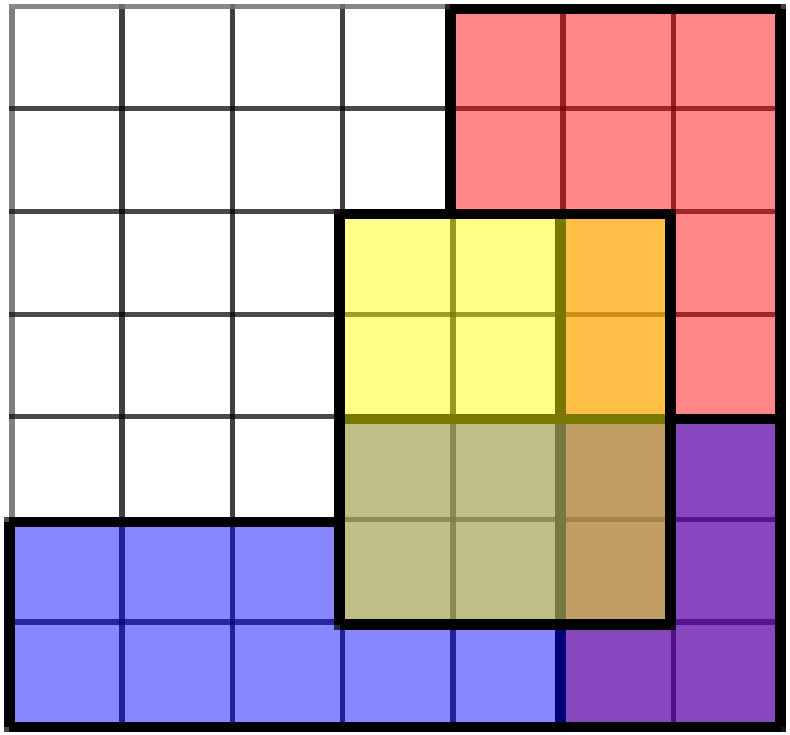
If we know Z, then Y is irrelevant for prediction of X

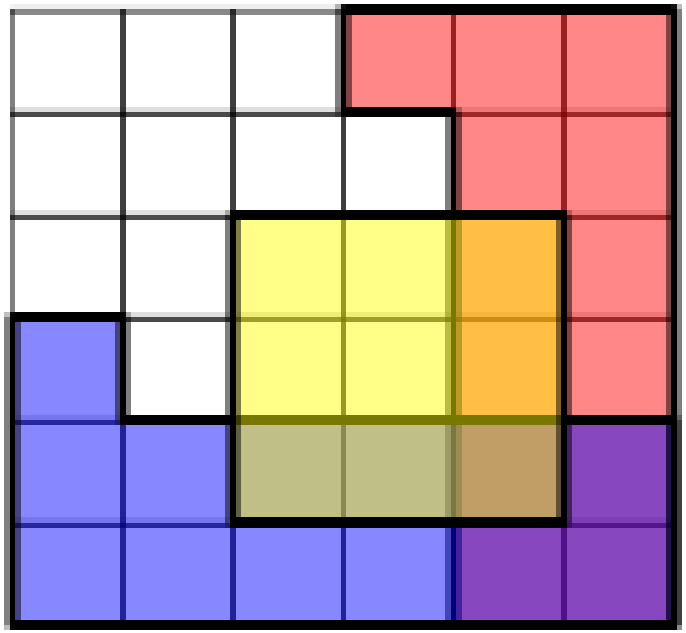
$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z),$$

If we know Z, then X is irrelevant for prediction of Y

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

KTH VETENSKAP OCH KONST
KTH Matematik





KTH Matematik

$\{X_n\}_{n\geq 1}$ is a time homogeneous Markov chain, n is "time'. An equivalent formulation of the Markov property is

$$P(X_1 = x, X_3 = y \mid X_2 = z) = P(X_1 = x \mid X_2 = z)P(X_3 = y \mid X_2 = z)$$

i.e., past and future are CI given the present.

KTH VETENSKAP OCH KONST

KTH Matematik

In general independence does not imply CI

Let $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and set $p_i = P(i) = \frac{1}{8}$ Set $X =$ the indicator function of $= \{1, 2, 3, 4\}$ and $Y =$ the indicator function of $= \{3, 4, 5, 6\}$ and $Z =$ the indicator function of $\{2, 3, 4, 5\}$. Then X and Y are independent but they are not conditionally independent given Z.

KTH VETENSKAP OCH KONST

KTH Matematik

Independence lost under conditioning

X and Y are independent Bernoulli random variables with $P(X = 1) = P(Y = 1) = p, 0 < p < 1$. Set $Z = X + Y$. Then $P(X = 1 \mid Z = 1) > 0$ and $P(Y = 1 \mid Z = 1) > 0$ but

$$P(X = 1, Y = 1 \mid Z = 1) = 0.$$

Berkson's phenomenon or selection bias

Another example showing that X independent of Y neither implies nor is implied by X and Y being CI given Z. Let Z denote the event that someone is admitted to a (US ?) college, which is made true if they are either brainy (X) or sporty (Y). Suppose in the general population, X and Y are independent (due to Kevin Murphy).
Now look at a population of college students, those for which Z is observed to be true. It will be found that being brainy makes you less likely to be sporty and vice versa, because either property alone is sufficient to explain the evidence on Z, i.e.,
$P(Y = 1|Z = 1, X = 1) \leq P(Y = 1|Z = 1))$.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayesian statistics and CI

Assume that $X_i \sim f(x \mid \theta)$ for $i = 1, \ldots, n.$, and that θ is an outcome of a random variable $\Theta \sim f_\Theta(\theta)$ (=prior density). Then Bayesian statistics evokes the model

$$X_i \perp X_j \mid \Theta, \quad i \neq j$$

X_i, $i = 1, \ldots, n$, are conditionally independent and identically distributed (I.I.D.)

KTH VETENSKAP OCH KONST
KTH Matematik

References

- B.L.S. Rao: Conditional independence, conditional mxing and conditional association. *Annals of the Institute of Statistical Mathematics*, 61, pp.441−460, 2009 .
- Schervish, M.J.: *Theory of statistics*, 1995, Springer.

KTH Matematik

Graphoids

A **dependency model** M over a finite set of variables U is is any set of triplets $X, Y; Z$ of disjoint subsets of U. The interpretation of M is that $(X, Y; Z)$ belongs to M if and only if X is independent of Y given Z, we write this as $I(X, Y; Z)$.

Example: Any probability distribution p gives a dependence model, which we write as M$= \mathcal{I}(p)$. $I(X, Y; Z)$ designates then CI.

KTH Matematik

Graphoids

A **graphoid** is any dependency model M for which the following set of axioms holds:

- Triviality: $I(X, \emptyset | Z) \in M$
- Symmetry: $I(X, Y; Z) \in M \Rightarrow I(Y, X; Z) \in M$.
- Decomposition: $I(X, Y \cup W; Z) \in M \Rightarrow I(Y, X; Z) \in M$.
- Weak union: $I(X, Y \cup W; Z) \in M \Rightarrow I(Y, X; Z \cup W) \in M$.
- Contraction: $I(X, Y; Z) \in M \& I(X, W; Y \cup Z) \in M \Rightarrow I(Y \cup W, X; Z) \in M$.

KTH Matematik

References (on graphoids)

This set of lectures will not be getting into the theory of (semi)graphoids. The point of the theory is that the axioms describe other notions of dependence than just probabilistic- a basic reference is:

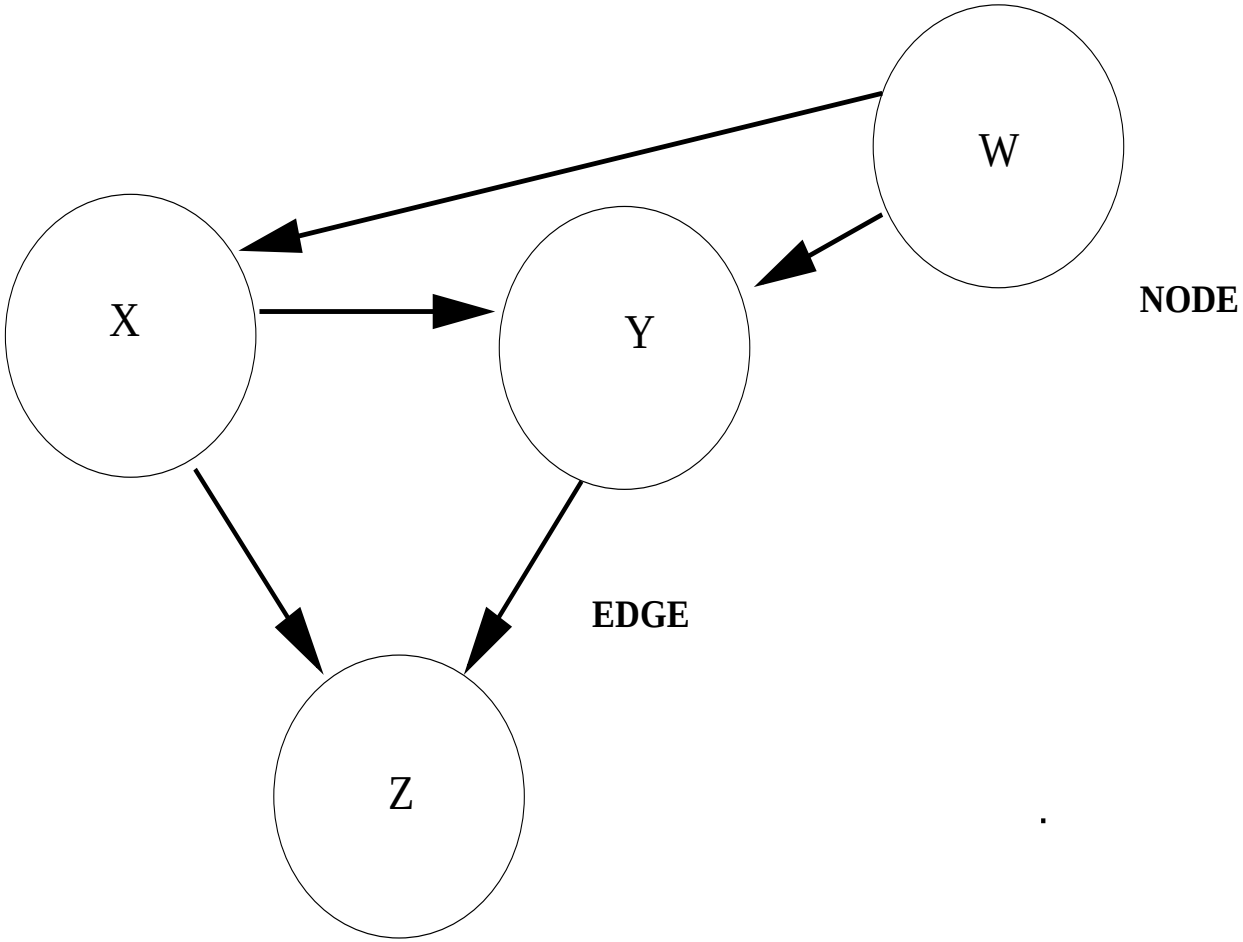
- M.Studený: *Probabilistic conditional independence structures*, 2005.

We have performed no check of the graphoid axioms for context-specific independence.

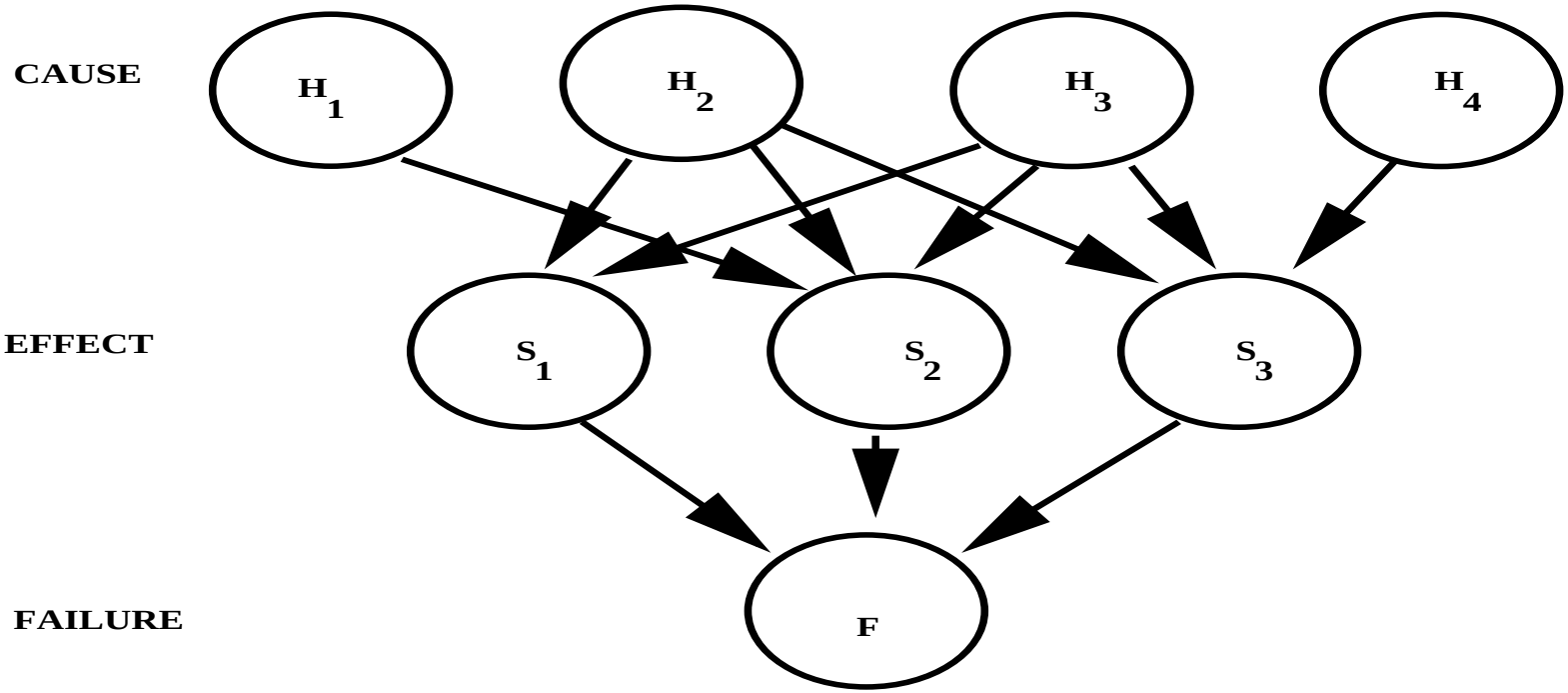
KTH VETENSKAP OCH KONST
KTH Matematik

DAG

A DAG is a graph that is built up by nodes and directed edges. The acyclic property ensures that no directed path starting from a node leads back to that particular node.



DAG FOR RCA



Notation

A DAG will be denoted by $G = (V, E)$ where $V = \{1, \ldots, d\}$ is the set of nodes and $E \subset V \times V$ is the set of edges such that if $(i, j) \in E$ then the graph contains a directed edge from node i to j.

KTH VETENSKAP OCH KONST

KTH Matematik

Notation:Parent set

Nodes, from which there is a directed edge to node j, are called parents of j and the set of all such nodes is denoted by $\Pi_j = \{i \in V : (i,j) \in E\}$. The nodes V give the indices of a set of stochastic variables $X = \{X_1, \ldots, X_d\}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Notation

Due to the close relationship between a node and its corresponding variable, the terms node and variable are used interchangeably. We use small letters x_j to denote a value taken by the corresponding variable.

KTH VETENSKAP OCH KONST

KTH Matematik

Notation

If $S \subseteq V$, then X_S denotes the corresponding set of variables. The outcome space of a variable X_j is denoted by $\mathcal{X}_j$ and the joint outcome space of a set of variables by the Cartesian product $\mathcal{X}_S = \times_{j \in S} \mathcal{X}_j$. The cardinality of the outcome space of X_S is denoted by $|\mathcal{X}_S|$.

KTH VETENSKAP OCH KONST
KTH Matematik

A DAG encodes independence statements in the form of conditional independencies.

Definition

Conditional Independence (CI)
Let $X = \{X_1, \ldots, X_d\}$ be a set of stochastic variables where $V = \{1, \ldots d\}$ and let A, B, S be three disjoint subsets of V. X_A is conditionally independent of X_B given X_S if

$$p(x_A \mid x_B, x_S) = p(x_A \mid x_S)$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_S) > 0$. This will be denoted by

$$X_A \perp X_B \mid X_S.$$

If we let $X_S = \varnothing$, then $X_A \perp X_B$ simply denotes marginal independence between the two sets of variables.

Local Directed Markov

A probability function p over the random vector $\underline{X} = (X_1, \ldots, X_d)$ satisfies the *local directed Markov condition* with respect to a DAG $\mathcal{G} = (V, E)$ or, equivalently, if and only if there is an ordering of the variables σ such that $\Pi_j^{(\sigma)} \in \{X_{\sigma(1)}, \ldots, X_{\sigma(j-1)}\}$ for each $j \in \{1, \ldots, d\}$ and such that $X_{\sigma(j)}$ is conditionally independent, given $\Pi_j^{(\sigma)}$ (the set of parents of $X_{\sigma(j)}$) of all the variables in the set $V \backslash (V_j^{(\sigma)} \cup \Pi_j^{(\sigma)})$, where $V_j^{(\sigma)}$ is the set of all *descendants* of $X_{\sigma(j)}$.

KTH VETENSKAP OCH KONST

KTH Matematik

Let p be a probability distribution over a set of variables $V = \{X_1, \ldots, X_d\}$. Then p satisfies the l.d.m.p. with respect to a graph $\mathcal{G} = (V, E)$ if and only if there is an ordering of the variables σ such that p factorises along $\mathcal{G}$, i.e.,

$$p(X_1, \ldots X_d) = \prod_{j=1}^{d} p(X_j \mid X_{\Pi_j}), \tag{1}$$

where the factors are **conditional probability table**s (CPTs) that correspond to local structures. By local structure, we refer to the node itself, its parents and the edges from the parents to the node.

Bayes Networks

The pair $(p, \mathcal{G})$ with

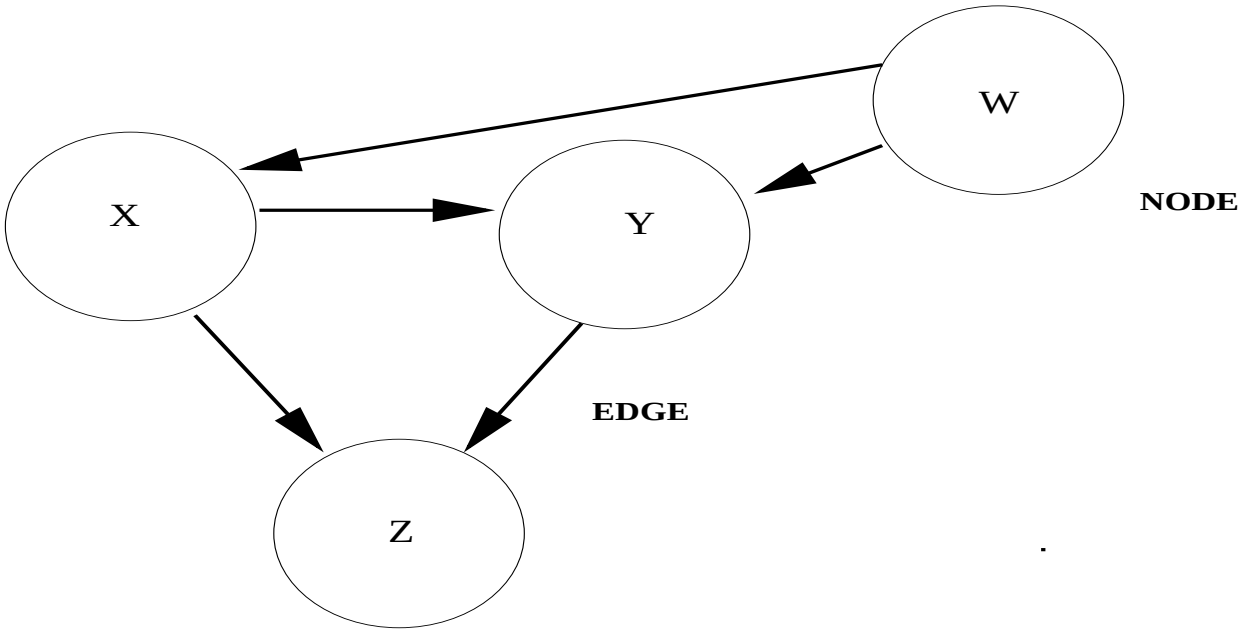
$$p(X_1, \ldots X_d) = \prod_{j=1}^{d} p(X_j \mid X_{\Pi_j}), \tag{2}$$

is a Bayes(ian) network.

KTH VETENSKAP OCH KONST

KTH Matematik

Bayes networks



$$p(Z, X, Y, W) = p(Z|X, Y)p(Y|X, W)p(X|W)p(W)$$

Conditional Probability Table (CPT)

Z, X, Y assume two values t and f in the DAG above. Textbook style CPT (=conditional probability table)

$$p_{Z|X,Y}(t|.,.) = \begin{array}{c|cc} X \backslash Y & t & f \\ \hline t & 0.5 & 0.99 \\ f & 0.85 & 0.0001. \end{array}$$

Background: Contex Specific Independence

The constraints imposed by the structure of a DAG alone have been recognized to be unnecessarily stringent under certain circumstances where context-specific or asymmetric independence can play a natural role.

KTH VETENSKAP OCH KONST
KTH Matematik

Background: Contex Specific Independence

In particular, one important notion is to allow the dependencies to have local structures, such that a node need not explicitly depend on all the combinations of values of its parents. This leads to context-specific independence (CSI) which can substantially reduce the parametric dimensionality of a network model and lead to a more expressive interpretation of the dependence structure.

KTH VETENSKAP OCH KONST
KTH Matematik

CPT with a regularity

Z, X, Y assume two values t and f in the DAG above. A CPT is

$$p_{Z|X,Y}(t|.,.) = \begin{array}{c|cc} X \backslash Y & t & f \\ \hline t & 0.6 & 0.6 \\ f & 0.85 & 0.0001. \end{array}$$

I.e., for $y \in \{t, f\}$ and $z \in \{t, f\}$,

$$p_Z(Z = z \mid X = t, Y = y) = p_Z(Z = z \mid X = t).$$

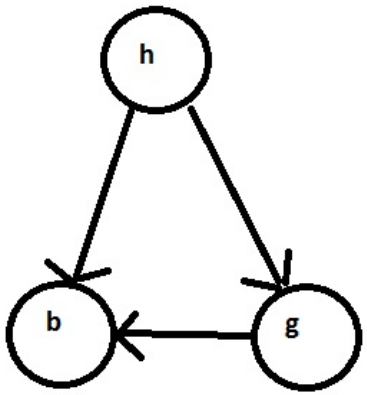
An Example due to Geiger and Heckerman. loc.cit.

A guard of a secured building expects three types of persons (h) to approach the building's entrance: workers in the buildings, approved visitors, and spies. As a person approaches the building, the guard can note its gender (g) and whether or not the person wears a badge (b). Spies are mostly men. Spies always wear badges in an attempt to fool the guard. Visitors don't wear badges because they don't have one. Female workers tend to wear badges more often than do male workers. The task of the guard is to identify the type of person approaching the building.

KTH VETENSKAP OCH KONST
KTH Matematik

the spy/visitor/worker-scenario

The topology of this graph, however, hides the fact that gender and badge wearing are conditionally independent, given that the person is a spy or visitor. The corresponding joint probability distribution is, as a result of this, overparameterized in the sense that it requires a total of 11 free parameters although some of these are identical.



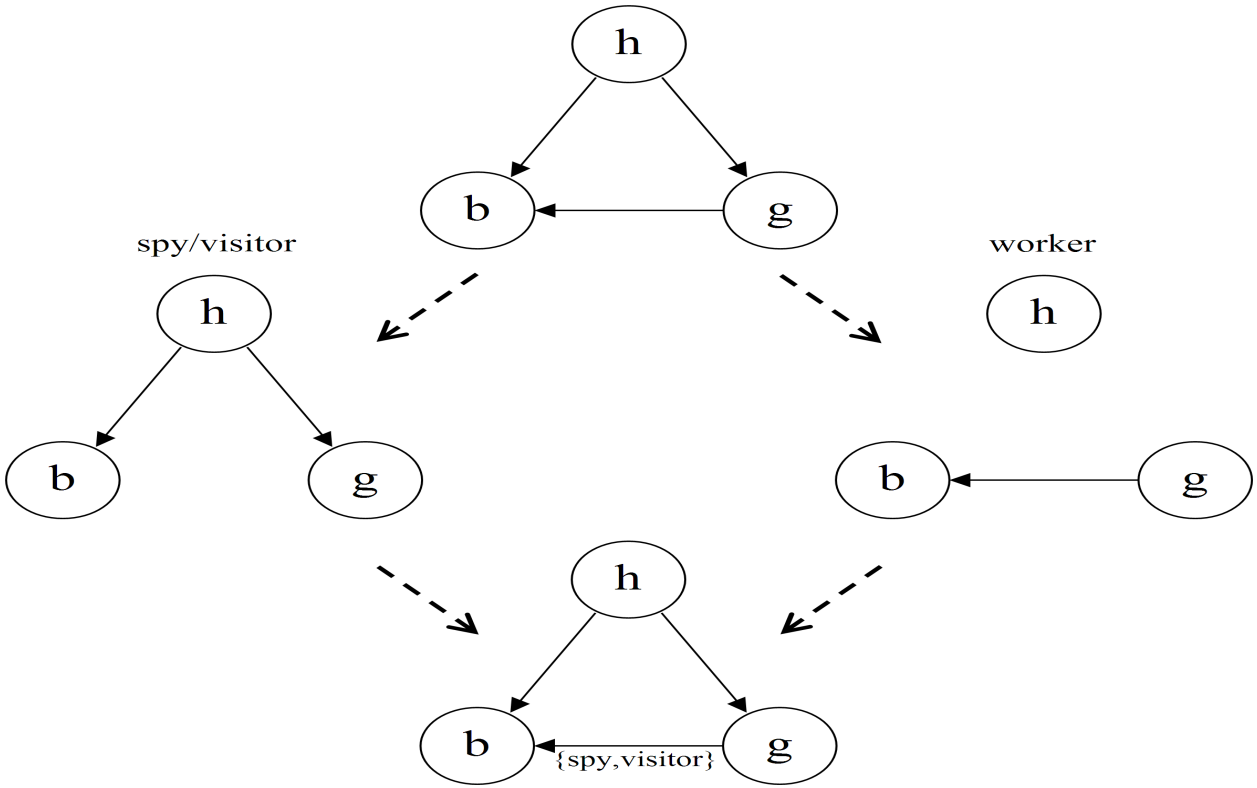
KTH Matematik

the spy/visitor/worker-scenario

Geiger and Heckerman noticed that this scenario is better represented by multiple graphs.
This representation of the scenario is made up of two context-specific graphs that together show that the dependence between gender and badge wearing only holds in the context of the person being a worker. The corresponding joint probability distribution now only requires 9 free parameters.

KTH VETENSKAP OCH KONST
KTH Matematik

Graph structures describing the spy/visitor/worker-scenario



Now consider the labeled DAG on the bottom in Figure. We have added the label $\{spy, visitor\}$ to the edge (g, b). This label implies that gender and badge are independent given that the person approaching the building is a spy or a visitor.

Background: Contex Specific Independence

A generalization of the CSI models: by allowing the independencies to be represented in terms of labels for the parental configurations of a node.
The approach here introduces a partition of the parental configurations into classes with invariant conditional probability distributions for the outcomes that are assigned to the same class.

KTH VETENSKAP OCH KONST
KTH Matematik

Background: Contex Specific Independence

It is shown that such a definition leads to a model class with a number of desirable properties, and we derive several properties of the models, including their identifiability and an LDAG version of the concept of a Markov equivalence class.

KTH VETENSKAP OCH KONST

KTH Matematik

Context-Specific Independence

The topology of a DAG restricts it to only encoding for independence relations that hold globally. However, as seen by the scenario above it is natural to consider independence relations that only hold in certain contexts.

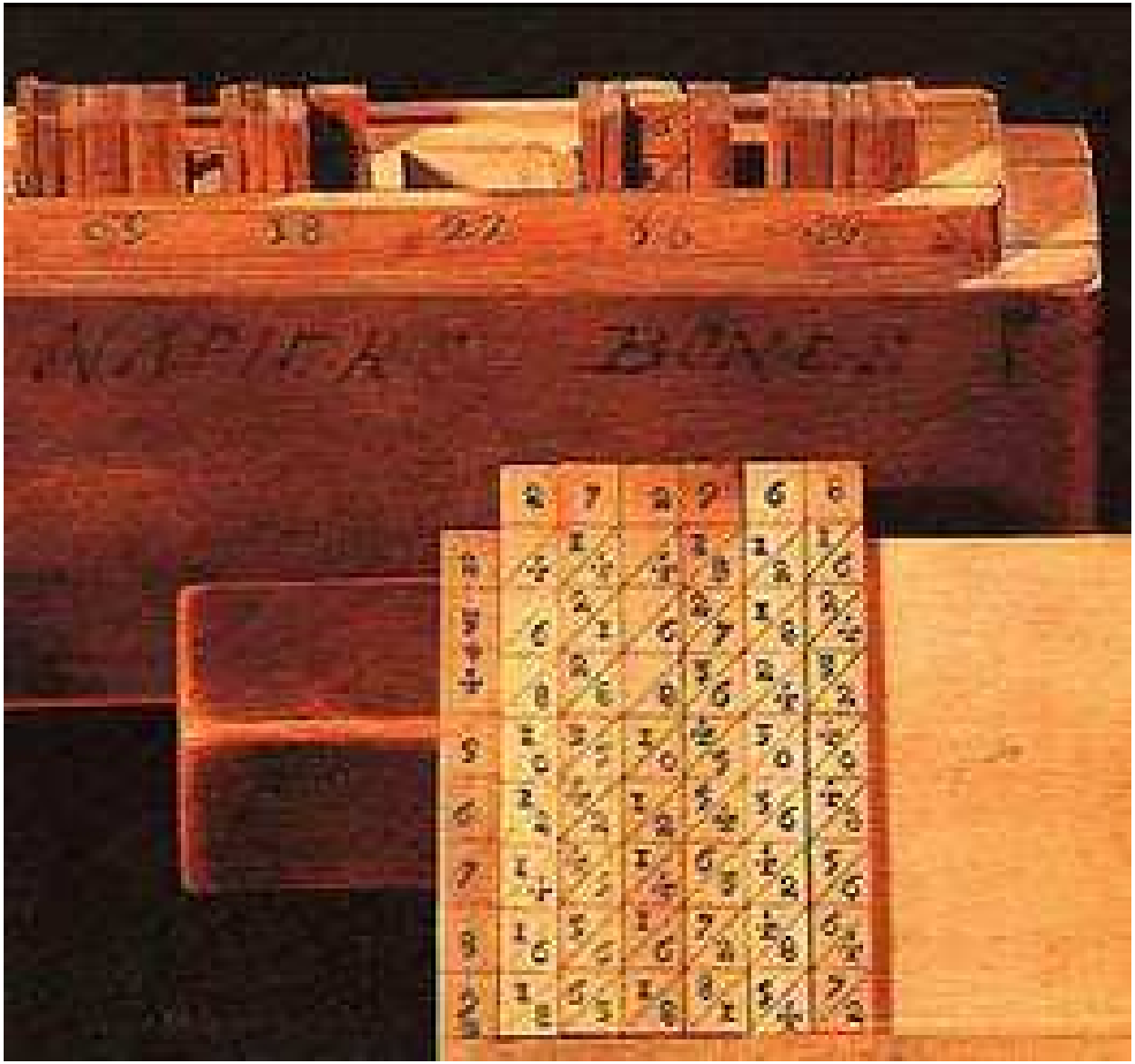
Definition

Context-Specific Independence (CSI)
Let $X = \{X_1, \ldots, X_d\}$ be a set of stochastic variables where $V = \{1, \ldots d\}$ and let A, B, C, S be four disjoint subsets of V. X_A is contextually independent of X_B given X_S and the context $X_C = x_C$ if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S)\,,$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_C, x_S) > 0$.

KTH Matematik


NAPIER BONES

Context-Specific Independence

Definition

Context-Specific Independence (CSI)
Let $X = \{X_1, \ldots, X_d\}$ be a set of stochastic variables where $V = \{1, \ldots d\}$ and let A, B, C, S be four disjoint subsets of V. X_A is contextually independent of X_B given X_S and the context $X_C = x_C$ if

$$p(x_A \mid x_B, x_C, x_S) = p(x_A \mid x_C, x_S)\,,$$

holds for all $(x_A, x_B, x_S) \in \mathcal{X}_A \times \mathcal{X}_B \times \mathcal{X}_S$ whenever $p(x_B, x_C, x_S) > 0$. This is denoted by

$$X_A \perp X_B \mid x_C, X_S.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Context-Specific Independence

The definition is valid for any probability distribution. Let $X_{[1}, X_2, X_3$ be binary r.v.'s and assume that $X_2 \perp X_3 \mid X_1 = 1$ and $X_1 \perp X_3 \mid X_2 = 1$, so that

$$P(X_1 = 1, X_2 = x_2, X_3 = x_3) = P(X_1 = 1)P(X_2 = x_2 \mid X_1 = 1)P(X_3 = x_3 \mid X_1 = 1)$$

for all outcomes $x_2 \in \{0, 1\}, x_3 \in \{0, 1\}$ and

$$P(X_1 = x_1, X_2 = 1, X_3 = x_3) = P(X_2 = 1)P(X_1 = x_1 \mid X_2 = 1)P(X_3 = x_3 \mid X_2 = 1)$$

for all outcomes $x_1 \in \{0, 1\}, x_3 \in \{0, 1\}$.
This pair of restrictions implies a simplification

$$P(X_3 = x_3 \mid X_1 = 1) = P(X_3 = x_3 \mid X_2 = 1) = P(X_3 = x_3 \mid X_1 = 1, X_2 = 1).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Local CSI

It has been discovered by numerous authors that certain CSIs can naturally be captured simply by further refining (2). We will refer to these statements as local CSIs as they are confined to the local structures.

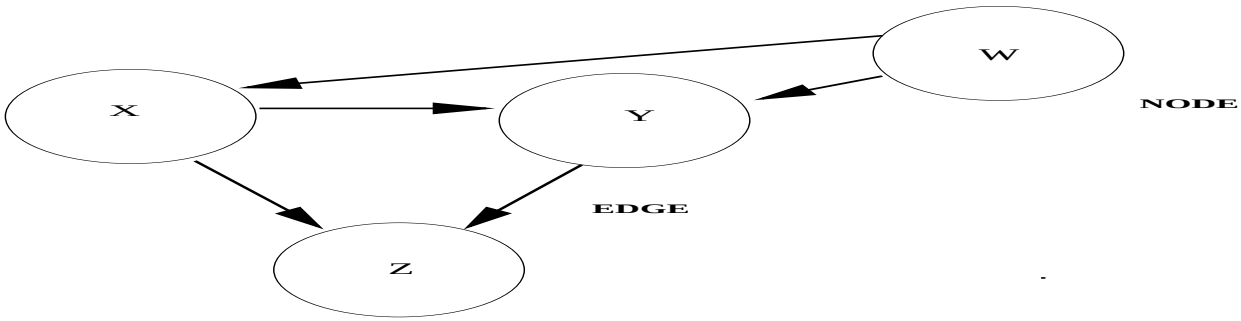
Definition

Local CSI in a DAG
A CSI in a DAG is local if it is of the form $X_j \perp X_B \mid x_C$, where B and C form a partition of the parents of node j.

In the CPD-based approaches to including CSI, the context-specific local structures cannot be read directly off the graph structure. This is the key to the usefulness of multinets. A multinet offers a natural representation of the dependence structure by explicitly showing the independencies in a graphical form.

KTH VETENSKAP OCH KONST
KTH Matematik

Local CSI



$$p_{Z|X,Y}(t|.,.) = \begin{array}{c|cc} X \backslash Y & t & f \\ \hline t & 0.6 & 0.6 \\ f & 0.85 & 0.0001. \end{array}$$

Local CSI

$$p_{Z|X,Y}(t|.,.) = \begin{array}{c|cc} X\backslash Y & t & f \\ \hline t & 0.6 & 0.6 \\ f & 0.85 & 0.0001. \end{array}$$

Let us note that in this example that the outcome space of the parent set (X, Y) is, as stated above, *partitioned* as $\{(t,t),(t,f)\} \cup \{(f,t)\} \cup \{f,f)\}$ where the set $\{(t,t),(t,f)\}$ corresponds to one and the same conditional distribution on Z.

KTH VETENSKAP OCH KONST

KTH Matematik

We visualize the local CSIs directly as a part of a single graph structure as done above; we add labels to the edges. This enables incorporation of local CSIs in a single graph as opposed to multinet -approaches, where one might need one graph for each distinct context. An LDAG is now formally defined as a DAG with labels representing local CSIs.

Definition

Labeled Directed Acyclic Graph (LDAG)
Let $G = (V, E)$ be a DAG over the stochastic variables $\{X_1, \ldots, X_d\}$. For all $(i, j) \in E$, let $L_{(i,j)} = \Pi_j \setminus \{i\}$. A label on an edge $(i, j) \in E$ is defined as the set

$$\mathcal{L}_{(i,j)} = \left\{ x_{L_{(i,j)}} \in \mathcal{X}_{L_{(i,j)}} : X_j \perp X_i \mid x_{L_{(i,j)}} \right\}.$$

Definition

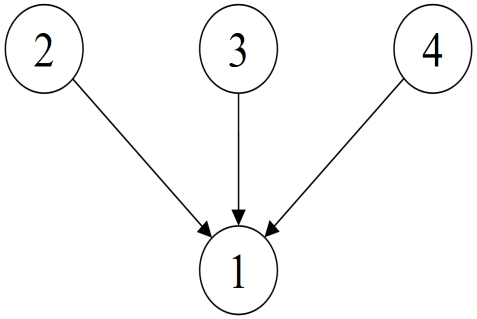
Labeled Directed Acyclic Graph (LDAG)
Let $G = (V, E)$ be a DAG over the stochastic variables $\{X_1, \ldots, X_d\}$. For all $(i, j) \in E$, let $L_{(i,j)} = \Pi_j \setminus \{i\}$. A label on an edge $(i, j) \in E$ is defined as the set

$$\mathcal{L}_{(i,j)} = \left\{ x_{L_{(i,j)}} \in \mathcal{X}_{L_{(i,j)}} : X_j \perp X_i \mid x_{L_{(i,j)}} \right\}.$$

An LDAG is a DAG to which the label set $\mathcal{L}_E = \{\mathcal{L}_{(i,j)} : \mathcal{L}_{(i,j)} \neq \varnothing\}_{(i,j)\in E}$ has been added, it is denoted by $G_L = (V, E, \mathcal{L}_E)$

KTH VETENSKAP OCH KONST
KTH Matematik

Local CSI-structure and the corresponding CPT



-5cm-

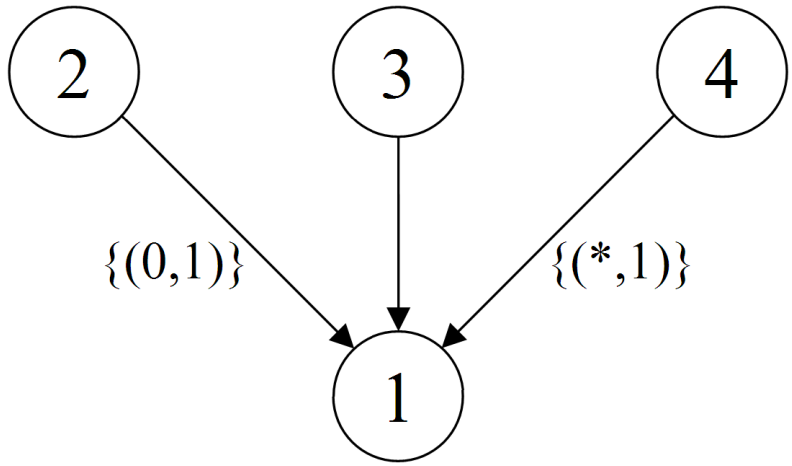
5cm

X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_3
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 0$	p_4
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 1$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_2
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_3
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 0$	p_5
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_5

KTH VETENSKAP OCH KONST

KTH Matematik

Local CSI-structure and the corresponding local CSIs



$$\mathcal{L}_{(2,1)} = (0,1) \qquad \Rightarrow X_1 \perp X_2 \mid (X_3, X_4) = (0,1)$$

$$\mathcal{L}_{(4,1)} = \mathcal{X}_2 \times \{1\} \qquad \Rightarrow X_1 \perp X_4 \mid X_2 \in \mathcal{X}_2, X_3 = 1$$

$$\Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1$$

KTH VETENSKAP OCH KONST

KTH Matematik

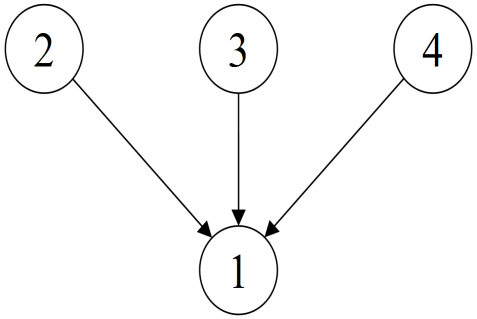
Representing CPT

We use complete AND-rules to represent the distinct parent configurations. A rule is complete if all parental variables are part of it.

KTH VETENSKAP OCH KONST

KTH Matematik

Local CSI-structure and the corresponding CPT



-5cm-

5cm

X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_3
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 0$	p_4
$X_2 = 0 \wedge X_3 = 1 \wedge X_4 = 1$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_2
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_3
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 0$	p_5
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_5

KTH VETENSKAP OCH KONST

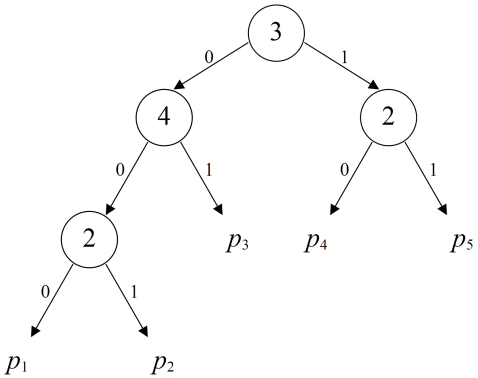
KTH Matematik

CPT trees

The regularities in the CPT above can be captured by the CPT-tree. Each path in the tree corresponds to a rule that can be described by the AND-operator. By simply traversing down each distinct path until we reach a terminal node or leaf, we can transform the CPT into its reduced counterpart on the right. All parent configurations satisfying a certain rule give rise to the same CPD. This implies that the rules in a reduced CPT must be mutually exclusive for the representation to be minimal.

KTH VETENSKAP OCH KONST
KTH Matematik

Reduced CPT



-5cm-

5cm

X_{Π_j}	$p(X_1 \mid X_{\Pi_j})$
$X_3 = 0 \wedge X_4 = 0 \wedge X_2 = 0$	p_1
$X_3 = 0 \wedge X_4 = 0 \wedge X_2 = 1$	p_2
$X_3 = 0 \wedge X_4 = 1$	p_3
$X_3 = 1 \wedge X_2 = 0$	p_4
$X_3 = 1 \wedge X_2 = 1$	p_5

KTH VETENSKAP OCH KONST

KTH Matematik

Reduced CPT

Each path in the tree corresponds to a rule that is described by the AND-operator. By simply traversing down each distinct path until we reach a terminal node or leaf, we can transform the CPT into its reduced counterpart on the right. All parent configurations satisfying a certain rule give rise to the same CPD. This implies that the rules in a reduced CPT must be mutually exclusive for the representation to be minimal.

KTH VETENSKAP OCH KONST
KTH Matematik

Reduced CPT

The rules corresponding to a tree are mutually exclusive as two distinct paths cannot lead to the same leaf. If a variable is not part of a path (or the corresponding AND-rule), it implies that the particular variable is contextually independent of the variable associated with the CPT given the context encoded by the path (or rule). Following this method we can read off the following local CSIs:

$$
\begin{aligned}
X_3 = 0 \wedge X_4 = 1 \quad &\Rightarrow \quad X_1 \perp X_2 \mid (X_3, X_4) = (0,1) \\
X_3 = 1 \wedge X_2 = 0 \quad &\Rightarrow \quad X_1 \perp X_4 \mid (X_2, X_3) = (0,1) \\
X_3 = 1 \wedge X_2 = 1 \quad &\Rightarrow \quad X_1 \perp X_4 \mid (X_2, X_3) = (1,1) \\
&\Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1
\end{aligned}
$$

KTH VETENSKAP OCH KONST

KTH Matematik

CPT trees

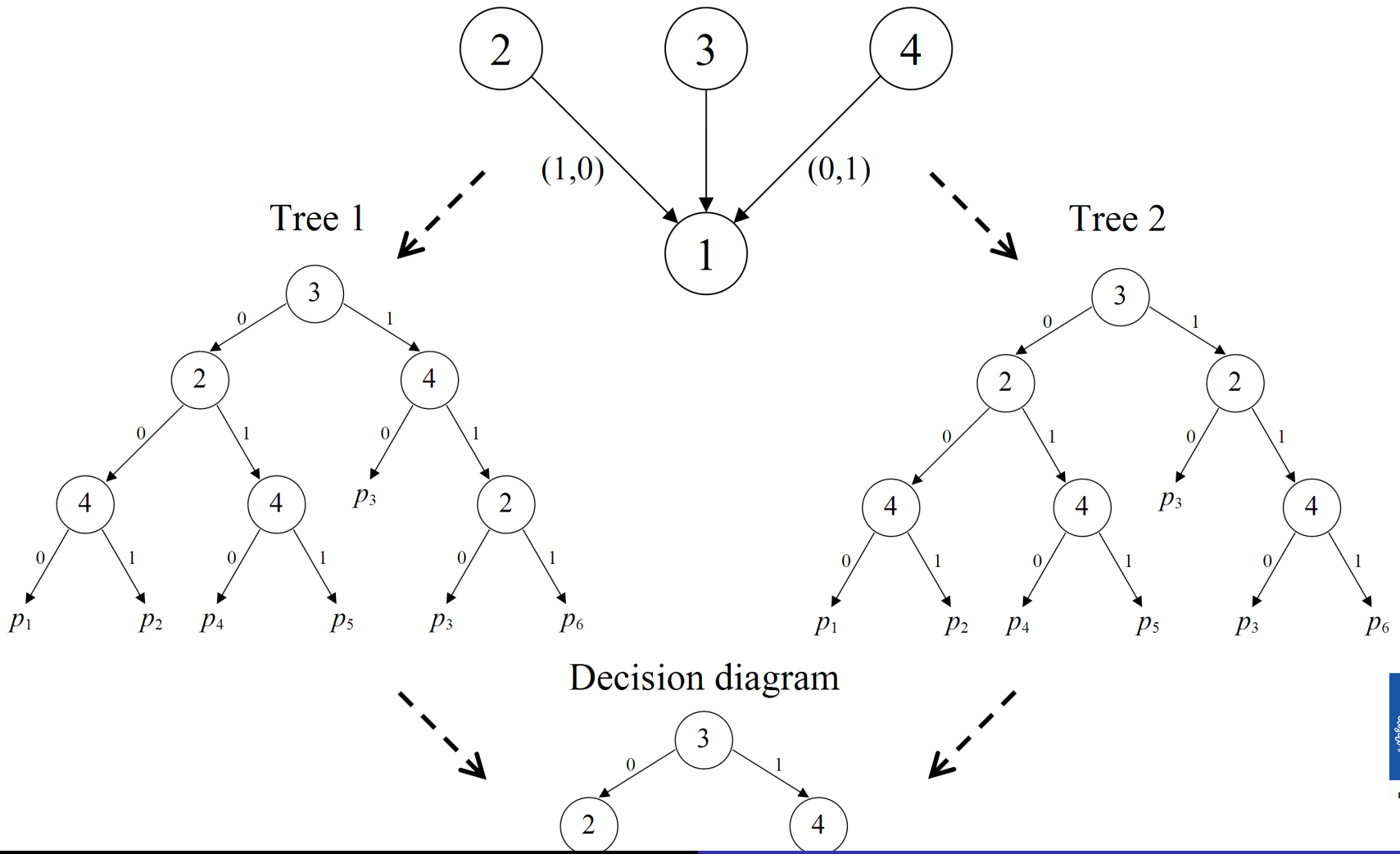
$$
\begin{aligned}
X_3 = 0 \wedge X_4 = 1 \quad &\Rightarrow \quad X_1 \perp X_2 \mid (X_3, X_4) = (0,1) \\
X_3 = 1 \wedge X_2 = 0 \quad &\Rightarrow \quad X_1 \perp X_4 \mid (X_2, X_3) = (0,1) \\
X_3 = 1 \wedge X_2 = 1 \quad &\Rightarrow \quad X_1 \perp X_4 \mid (X_2, X_3) = (1,1) \\
&\Leftrightarrow X_1 \perp X_4 \mid X_2, X_3 = 1
\end{aligned}
$$

The CSIs above coincide with the labels of this specific LDAG. More generally, any CPT-tree can be transformed into a reduced CPT by mutually exclusive AND-rules. Subsequently, incomplete rules can be turned into labels as illustrated in the above example.

KTH Matematik

LDAG

Consider the LDAG on the top in the Figure. Its associated minimal reduced CPT is shown later. This CSI-structure cannot be compactly represented by the structure of a CPT-tree.



To show how the minimal reduced CPT is recovered from the labels we proceed stepwise. Each of the labels corresponds to a single reduced AND-rule resulting in the table.

X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_2
$X_2 = 0 \wedge X_3 = 1$	p_3
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_5
$X_3 = 1 \wedge X_4 = 0$	p_3
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_6

KTH VETENSKAP OCH KONST
KTH Matematik

X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_2
$X_2 = 0 \wedge X_3 = 1$	p_3
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_5
$X_3 = 1 \wedge X_4 = 0$	p_3
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_6

The rules on row 3 and 6 are not mutually exclusive at this point as they both are satisfied by the common parent configuration $(X_2, X_3, X_4) = (0, 1, 0)$. This implies that any parent configuration satisfying any of these rules must give rise to the same CPD.

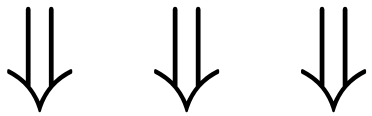
The AND-rules are therefore combined with the OR-operator resulting in the minimal reduced CPT shown next. More generally, each configuration in the labels of a local structure corresponds to an AND-rule. If any two rules overlap, they are combined with the OR-operator. The rules of a minimal reduced CPT created by this method will thus be mutually exclusive and exhaustive with respect to the outcome space of the parental variables.

LDAGs

X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_2
$X_2 = 0 \wedge X_3 = 1$	p_3
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_5
$X_3 = 1 \wedge X_4 = 0$	p_3
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_6

KTH VETENSKAP OCH KONST
KTH Matematik

CPT trees



X_{Π_1}	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 1$	p_2
$(X_2 = 0 \wedge X_3 = 1) \vee (X_3 = 1 \wedge X_4 = 0)$	p_3
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 1$	p_5
$X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1$	p_6

KTH VETENSKAP OCH KONST

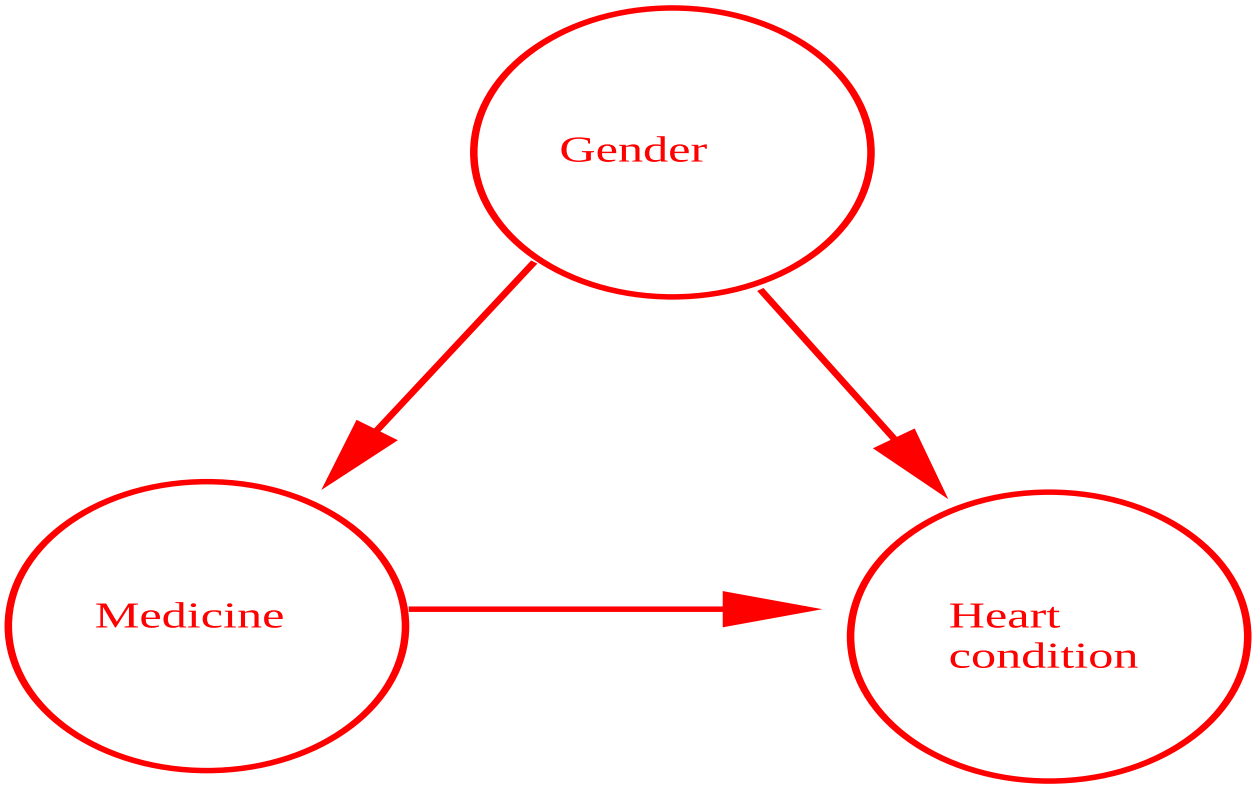
KTH Matematik

CSI-consistency

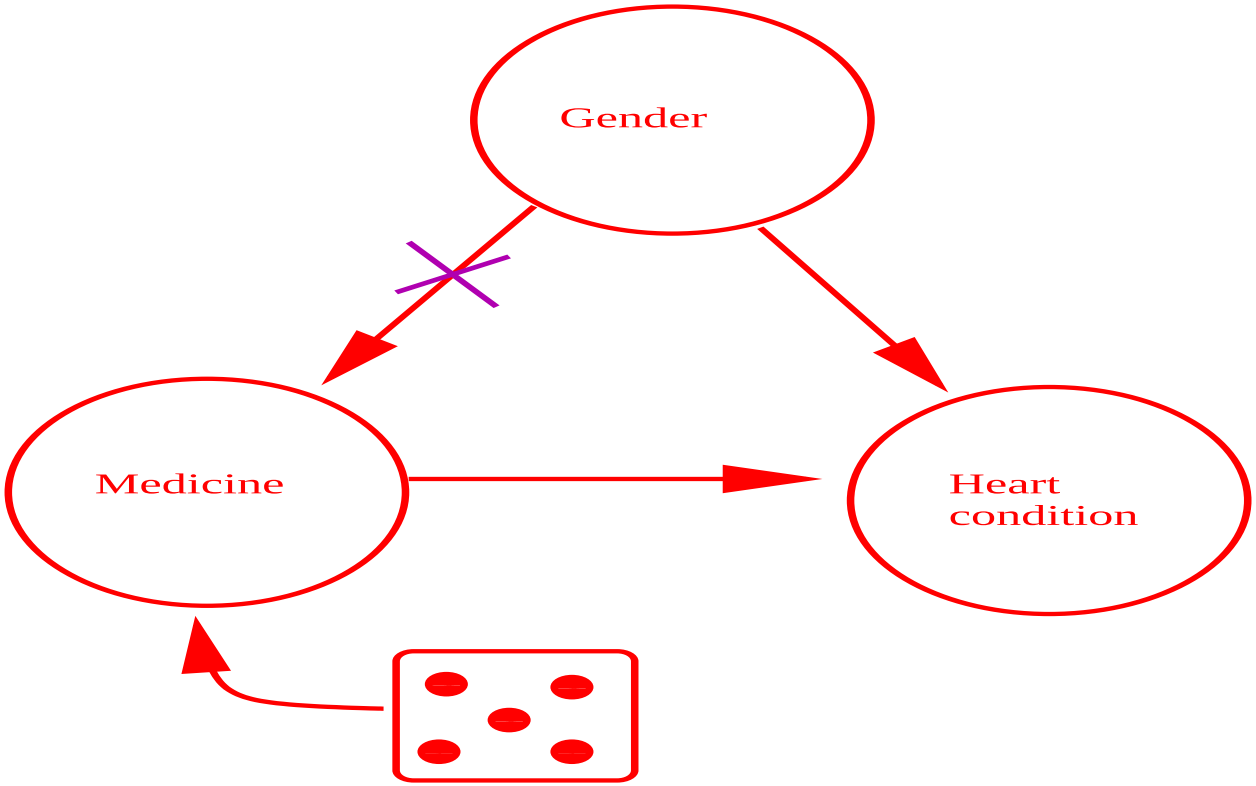
A CPT-representation may in fact be viewed as a function that given a parent configuration returns a CPD. The common factor among the different CPD-based representations is that they all induce partitions of the outcome space of the parental variables.

If the representation is based on the notion of CSI, the corresponding partition will be referred to as **CSI-consistent**.

KTH Matematik



KTH Matematik


Gender
Medicine
Heart
condition

The structure of LDAGs

- d-separation

The structure of LDAGs

- d-separation
- Equivalence classes of DAGs

KTH Matematik

The structure of LDAGs

- d-separation
- Equivalence classes of DAGs
- Maximal and regularity (extending Corander 2003)

KTH Matematik

The structure of LDAGs

- d-separation
- Equivalence classes of DAGs
- Maximal and regularity (extending Corander 2003)
- d-separation for LDAGs

KTH Matematik

The structure of LDAGs

- d-separation
- Equivalence classes of DAGs
- Maximal and regularity (extending Corander 2003)
- d-separation for LDAGs
- Equivalence classes of LDAGs

KTH VETENSKAP OCH KONST
KTH Matematik

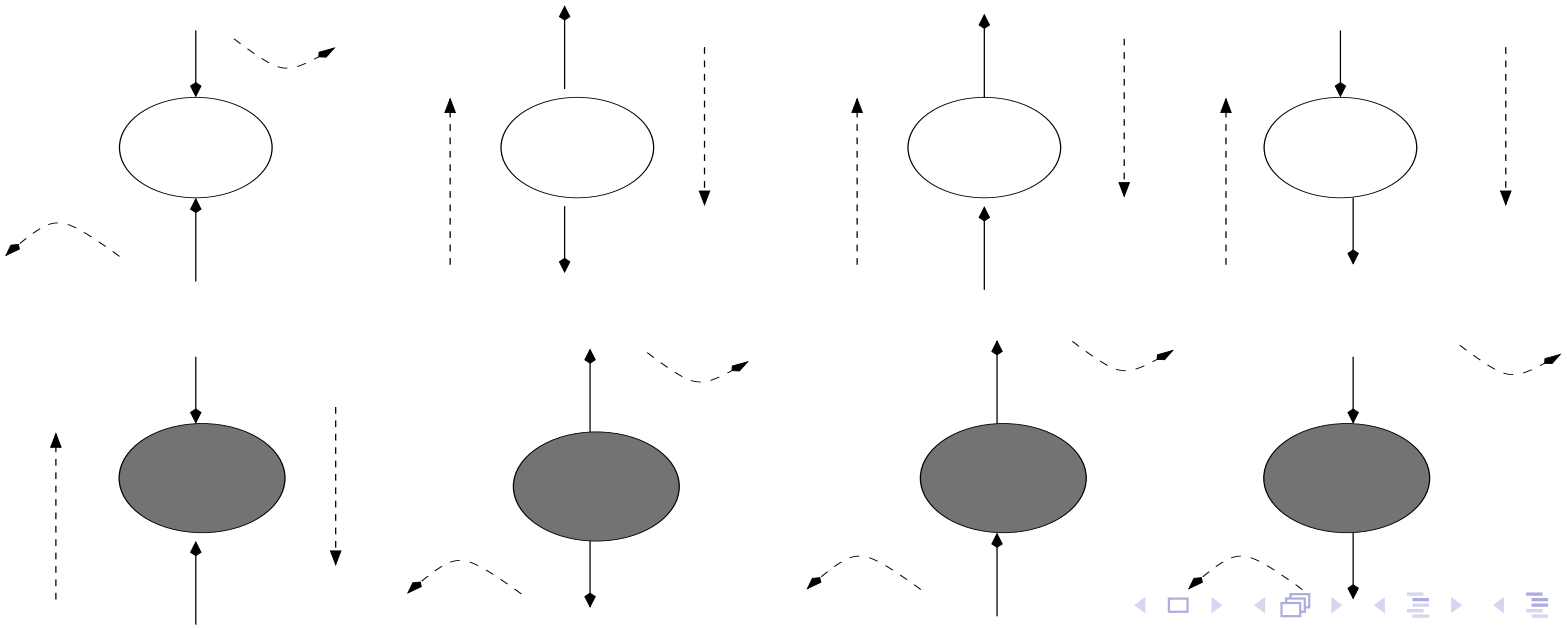
The structure of LDAGs

- d-separation: local directed Markov property gives local CI-statements. Further CI-statements that follow from these local statements are obtained by a graph theoretic criterion known as **d-separation**.

KTH VETENSKAP OCH KONST
KTH Matematik

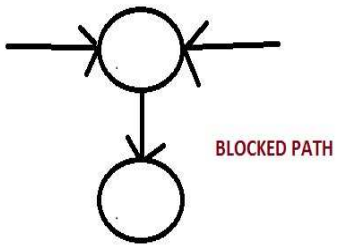
d-separation & Bayes Ball

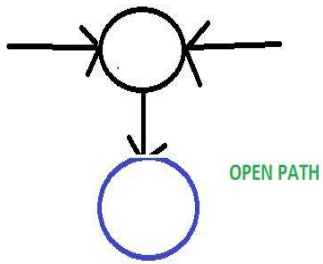
In general, the conditional independence relationships encoded by a DAG are best be explained by means of the "Bayes Ball" algorithm. Two (sets of) nodes A and B are conditionally independent (d-separated) given a set C if and only if there is no way for a ball to get from A to B in the graph, where the allowable movements of the ball are shown below. Hidden nodes are nodes whose values are not known, and are depicted as unshaded; the nodes we condition on are shaded. The dotted arcs indicate direction of flow of the ball.



KTH Matematik

d-separation & Bayes Ball

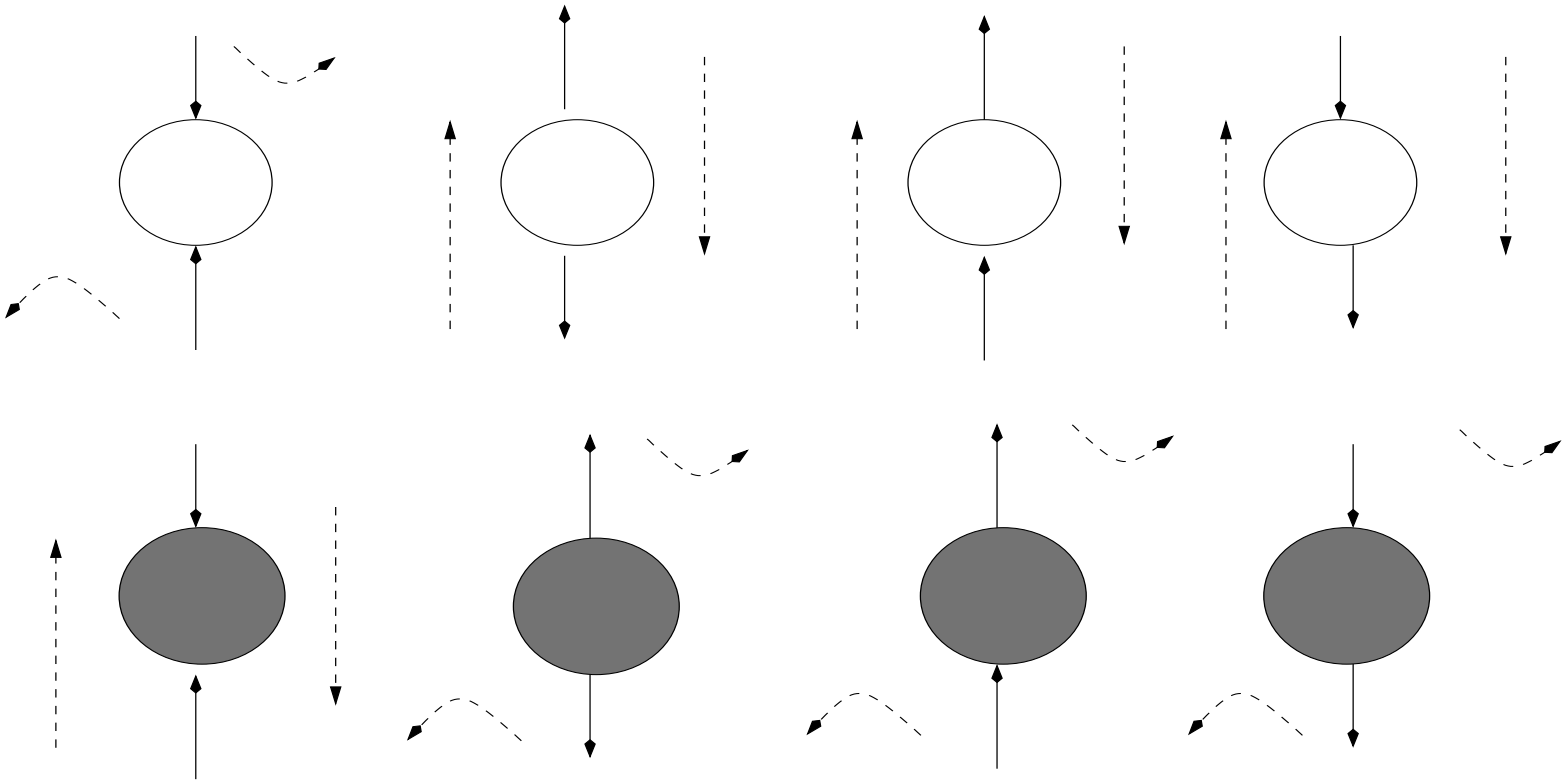




The blue node corresponds to conditioning.

KTH VETENSKAP OCH KONST
KTH Matematik

d-separation & Bayes Ball



$$X_A \perp X_B \,\|_{\mathcal{G}}\, X_S$$

KTH Matematik

d-separation

If a probability distribution factorises according to a directed acyclic graph, then any d-separation statement in the graph implies the corresponding conditional independence statement for the distribution. The converse does not hold in general.

$$X_A \perp X_B \,\|_{\mathcal{G}}\, X_S \Rightarrow X_A \perp X_B \mid X_S$$

KTH VETENSKAP OCH KONST
KTH Matematik

d-separation & Bayes Ball

Using the Bayes ball one now actually sees in the multinet of the gatekeeper scenario above that $b \perp g \mid h \in \{spy, visitor\}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Properties of LDAG

Definition

Maximal LDAG
An LDAG $G_L = (V, E, \mathcal{L}_E)$ is called maximal if there exists no configuration $x_{L_{(i,j)}}$ that can be added to the label $\mathcal{L}_{(i,j)}$ without inducing an additional local CSI.

Simple: in a maximal LDAG, all local CSIs are obtained directly from the graph.

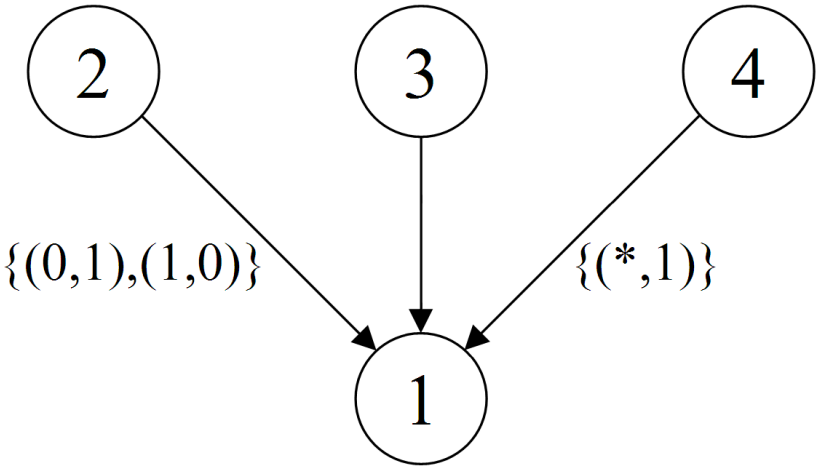
KTH VETENSKAP OCH KONST
KTH Matematik

Properties of LDAG

Theorem

Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E, \mathcal{L}_E^*)$ be two maximal LDAGs with the same underlying DAG $G = (V, E)$. Then G_L and G_L^* represent equivalent dependence structures if and only if $\mathcal{L}_E = \mathcal{L}_E^*$, i.e. $G_L = G_L^*$.

KTH VETENSKAP OCH KONST
KTH Matematik



$X_{\Pi_1} \in$	$p(X_1 \mid X_{\Pi_1})$
$X_2 = 0 \wedge X_3 = 0 \wedge X_4 = 0$	p_1
$X_3 = 0 \wedge X_4 = 1$	p_3
$X_3 = 1$	p_4
$X_2 = 1 \wedge X_3 = 0 \wedge X_4 = 0$	p_2

KTH VETENSKAP OCH KONST

KTH Matematik

Ilustration of Maximality of an LDAG

The local structure is similar to the previous example except that configuration $x_{L_{(2,1)}} = (1, 0)$ has been added to its label. The local structure is now not maximal since $(1, 1)$ can be added to $\mathcal{L}_{(2,1)}$ without resulting in an additional CSI.

$$\left.\begin{array}{rcl} x_{L_{(2,1)}} = (1,0) & \Rightarrow & X_3 = 1 \wedge X_4 = 0 \\ x_{L_{(4,1)}} = (0,1) & \Rightarrow & X_2 = 0 \wedge X_3 = 1 \\ x_{L_{(4,1)}} = (1,1) & \Rightarrow & X_2 = 1 \wedge X_3 = 1 \end{array}\right\} \Rightarrow X_3 = 1$$

KTH VETENSKAP OCH KONST

KTH Matematik

As $(0, 1, 1)$ and $(1, 1, 1)$ satisfy this rule, no further merging of rules is done when $x_{L_{(2,1)}} = (1, 1)$ is added to its label. This corresponds to the local CSI

$$X_1 \perp X_2 \mid X_3 = 1, X_4 = 1$$

implicitly being encoded by the other labels. This type of situation may arise when different label induced rules overlap and are combined with the OR-operator in order to achieve a minimal number of mutually exclusive rules.

KTH VETENSKAP OCH KONST
KTH Matematik

Regular Maximal LDAG

To ensure that the effect of an edge cannot completely vanish due to labels, we introduce the regularity condition for maximal LDAGs.

Definition

Regular maximal LDAG
A maximal LDAG $G_L = (V, E, \mathcal{L}_E)$ is regular if $\mathcal{L}_{(i,j)}$ is a strict subset of $\mathcal{X}_{L_{(i,j)}}$ for every label in G_L.

KTH VETENSKAP OCH KONST
KTH Matematik

Satisfied Label

We need to formulate the notion of separation for LDAGs. A couple of new definitions are needed.

Definition

Satisfied label

Let $G_L = (V, E, \mathcal{L}_E)$ be an LDAG and $X_C = x_C$ a context where $C \subseteq V$. In the context $X_C = x_C$, a label $\mathcal{L}_{(i,j)} \in \mathcal{L}_E$ is satisfied if $L_{(i,j)} \cap C \neq \varnothing$ and $\{x_{L_{(i,j)} \cap C}\} \times \mathcal{X}_{L_{(i,j)} \setminus C} \subseteq \mathcal{L}_{(i,j)}$.

KTH VETENSKAP OCH KONST

KTH Matematik

Context-specific LDAG

A context-specific LDAG is a reduced version of an LDAG where all satisfied edges are removed.

Definition

Let $G_L = (V, E, \mathcal{L}_E)$ be an LDAG. For the context $X_C = x_C$, where $C \subseteq V$, the context-specific LDAG is denoted by $G_L(x_C) = (V, E \setminus E', \mathcal{L}_{E \setminus E'})$ where $E' = \{(i,j) \in E : \mathcal{L}_{(i,j)} \textit{ is satisfied}\}$. The underlying DAG of the context-specific LDAG is denoted by $G(x_C) = (V, E \setminus E')$.

KTH VETENSKAP OCH KONST

KTH Matematik

CSI-separation

CSI-separation can now be defined.

Definition

CSI-separation in LDAGs
Let $G_L = (V, E, \mathcal{L}_E)$ be an LDAG and let A,B,S,C be four disjoint subsets of V. X_A is CSI-separated from X_B by X_S in the context $X_C = x_C$ in G_L, denoted by

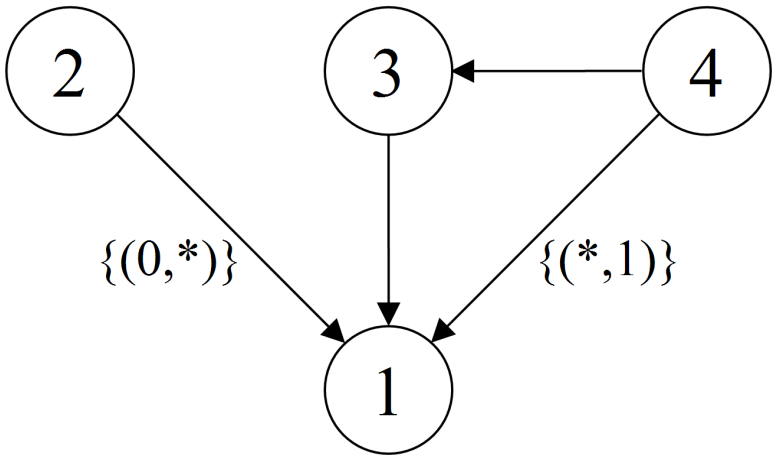
$$X_A \perp X_B \,||_{G_L}\, x_C, X_S,$$

if X_A is d-separated from X_B by $X_{S \cup C}$ in $G(x_C)$.

If $C = \varnothing$ in the above definition, the method describes the procedure of d-separation with respect to the underlying DAG.

KTH Matematik

LDAGs with CI inducing CSI-structures



When only considering the underlying DAG, it appears (Bayes ball) that

$$X_2 \not\perp X_4 \,||_G\, X_1, X_3 \Rightarrow X_2 \not\perp X_4 \mid X_1, X_3$$

LDAGs with CI inducing CSI-structures

However, through CSI-separation and reasoning by cases we recover

$$X_2 \perp X_4 \,||_{G_L}\, X_1, X_3 = 0 \;\; \Rightarrow X_2 \perp X_4 \mid X_1, X_3 = 0$$
$$X_2 \perp X_4 \,||_{G_L}\, X_1, X_3 = 1 \;\; \Rightarrow X_2 \perp X_4 \mid X_1, X_3 = 1$$

which eventually leads us to the conclusion that

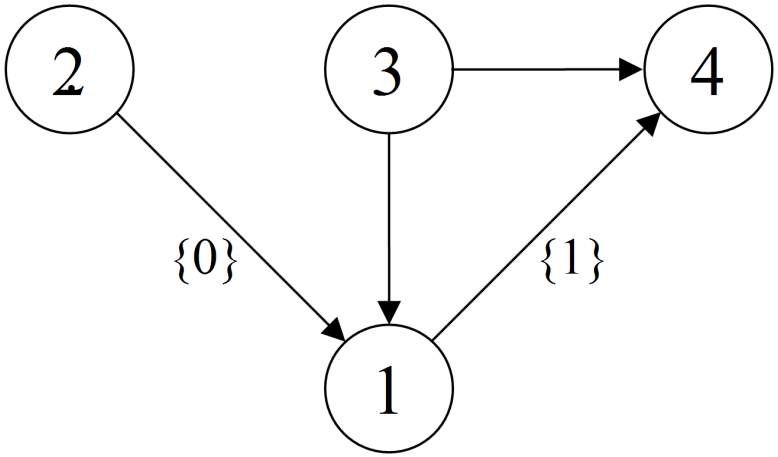
$$X_2 \perp X_4 \mid X_1, x_3 \;\; \forall x_3 \in \mathcal{X}_3 \Leftrightarrow X_2 \perp X_4 \mid X_1, X_3.$$

KTH VETENSKAP OCH KONST
KTH Matematik

d-separation is based on the notion of active trails, i.e. trails along which information can flow from one variable to another, and a lack of such trails will imply *d*-separation. Labels in an LDAG have the ability to cut off an active trail for a certain context by removing an edge in it and render the trail non-active or blocked in that context. The regularity condition prohibits this from occurring throughout the outcome space for a single edge but certain combinations of labels can still deactivate a trail that appears active when only considering the underlying DAG.

KTH VETENSKAP OCH KONST
KTH Matematik

CSI-separation



When considering the underlying DAG alone, it appears that

$$X_2 \not\perp X_4 \,||_G \, \varnothing \Rightarrow X_2 \not\perp X_4.$$

KTH VETENSKAP OCH KONST
KTH Matematik

CSI-separation

However, we can recover the following CSIs through CSI-separation:

$$\left.\begin{array}{l} X_2 \perp X_4 \,||_{G_L}\, X_3 = 0 \\ X_2 \perp X_4 \,||_{G_L}\, X_3 = 1 \end{array}\right\} \Rightarrow X_2 \perp X_4 \mid X_3$$

$$X_2 \perp X_3 \,||_G\, \varnothing \quad \Rightarrow X_2 \perp X_3$$

The first of the CIs must be discovered through reasoning by cases. The second is easily discovered from the underlying DAG.

KTH VETENSKAP OCH KONST

KTH Matematik

CSI-separation

Combining the CIs leads us (c.f., graphoid axioms) to the conclusion that

$$X_2 \perp X_4$$

indeed holds due to the structural properties of the LDAG. Several CSI-separation statements work together in order to achieve a non-local independence that is not easily discovered. However, both these situations are special cases that can only arise when the complete outcome space of a subset of variables is split up over several labels.

KTH VETENSKAP OCH KONST
KTH Matematik

CSI-separation

CSI-separation is proven to be a sound method for verifying CSIs, i.e.

$$X_A \perp X_B \,||_{G_L}\, x_C, X_S \Rightarrow X_A \perp X_B \mid x_C, X_S.$$

Unfortunately, it is not complete in the sense that there may arise situations where certain structure induced independencies cannot be discovered directly by the CSI-separation algorithm.

KTH VETENSKAP OCH KONST
KTH Matematik

KTH Matematik

Equivalence

We can restrict the model space to the subclass of regular maximal LDAGs without loss of generality. There still exist large classes of distinct LDAGs that encode equivalent dependence structures.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov Model

A distribution that has all the conditional independence statements corresponding to the entire set of d-separation statements for a DAG $\mathcal{G}$ is said to belong to the *Markov model* of $\mathcal{G}$. This is, of course, a dependency model as outlined above.

KTH VETENSKAP OCH KONST

KTH Matematik

Definition (Markov Model)

Let $V = \{X_1, \ldots, X_d\}$ denote a set of variables and let $\mathcal{G} = (V, D)$ be a directed acyclic graph. Let $\mathcal{V}$ denote the entire set of subsets of V. Let p be a probability function for the random vector $\underline{X} = (X_1, \ldots, X_d)$. Let

$$\mathcal{I}(p) = \{(X, Y, S) \in V \times V \times \mathcal{V} | X, Y \notin S, \qquad X \perp Y | S\}. \quad (3)$$

Note that $\phi \in \mathcal{V}$ and $X \perp Y | \phi$ means that $X \perp Y$.

Markov Model

The Markov Model $\mathcal{M}_{\mathcal{G}}$ *determined by a directed acyclic graph* $\mathcal{G} = (V, E)$ *is the set of conditional independence statements*

$$\mathcal{M}_{\mathcal{G}} = \{(X, Y, S) \in V \times V \times \mathcal{V} | X \perp Y \| _{\mathcal{G}} S\}. \tag{4}$$

That is, the Markov model *is the set of conditional independence relations satisfied by all distributions that are locally* $\mathcal{G}$*-Markovian. A distribution* p *is said to belong to the Markov Model of* $\mathcal{G}$*, written* $p \in \mathcal{M}_{\mathcal{G}}$*, if and only if* $\mathcal{M}_{\mathcal{G}} \subseteq \mathcal{I}(p)$*.*

KTH VETENSKAP OCH KONST

KTH Matematik

I-map

The collection of triples $\mathcal{M}_{\mathcal{G}}$ defined in equation (4) represents the entire set of conditional independence statements that it is possible to infer from the DAG, but this collection does not necessarily represent the complete set of independence statements that hold for a collection of variables under a given probability distribution. When it does, it is known as a *perfect I-map*.

KTH VETENSKAP OCH KONST
KTH Matematik

Perfect I-map

(Perfect I-Map, Faithful)

A DAG $\mathcal{G} = (V, E)$ over a set of variables V is known as a perfect I-map *for a probability function p over V if for any three disjoint subsets of variables A, B and S,*

$$X_A \perp X_B | X_S \Leftrightarrow X_A \perp X_B \parallel_{\mathcal{G}} X_S,$$

If $\mathcal{G}$ is a perfect I-map for p, then $\mathcal{G}$ is said to be faithful *to p.*

KTH VETENSKAP OCH KONST
KTH Matematik

An Aside on Faithfulness

Let Y_1, Y_2, Y_3 be three independent identically distributed binary variables, with probability function $p(0) = p(1) = \frac{1}{2}$. Let

$$X_1 = \begin{cases} 1 & Y_2 = Y_3 \\ 0 & Y_2 \neq Y_3 \end{cases}$$

$$X_2 = \begin{cases} 1 & Y_1 = Y_3 \\ 0 & Y_1 \neq Y_3 \end{cases}$$

$$X_3 = \begin{cases} 1 & Y_1 = Y_2 \\ 0 & Y_1 \neq Y_2 \end{cases}$$

KTH VETENSKAP OCH KONST

KTH Matematik

An Aside on Faithfulness

Then X_1, X_2, X_3 are pairwise independent, but not jointly independent.

$$p_{X_1,X_2,X_3}(1,1,1) = p(Y_1 = Y_2 = Y_3) = p_{Y_1,Y_2,Y_3}(1,1,1) + p_{Y_1,Y_2,Y_3}(0,0,0) = \frac{1}{4}$$

$$p_{X_1,X_2,X_3}(1,1,0) = p_{X_1,X_2,X_3}(1,0,1) = p_{X_1,X_2,X_3}(0,1,1) = p(Y_2 = Y_3 = Y_1, Y_1 \neq Y_2)$$

$$\begin{aligned} p_{X_1,X_2,X_3}(1,0,0) &= p_{X_1,X_2,X_3}(0,1,0) = p_{X_1,X_2,X_3}(0,0,1) \\ &= p(Y_2 = Y_3, Y_1 \neq Y_3, Y_1 \neq Y_2) \\ &= p(Y_1 = 1, Y_2 = Y_3 = 0) + p(Y_1 = 0, Y_2 = Y_3 = 1) = \frac{1}{4} \end{aligned}$$

$$p_{X_1,X_2,X_3}(0,0,0) = 0$$

KTH VETENSKAP OCH KONST

KTH Matematik

An Aside on Faithfulness

It follows that

$$p_{X_1,X_2}(1,1) = p_{X_1,X_2}(1,0) = p_{X_1,X_2}(0,1) = p_{X_1,X_2}(0,0) = \frac{1}{4}$$

so that $p_{X_1}(1) = p_{X_1}(0) = \frac{1}{2}$ and in all cases

$$p_{X_1,X_2} = p_{X_1} p_{X_2}.$$

But

$$\frac{1}{4} = p_{X_1,X_2,X_3}(1,1,1) \neq p_{X_1}(1)p_{X_2}(1)p_{X_3}(1) = \frac{1}{8}.$$

An Aside on Faithfulness

Since $X_1 \perp X_2$ but $X_3 \not\perp \{X_1, X_2\}$, $X_3 \not\perp X_1|X_2$ and $X_3 \not\perp X_2|X_1$, it follows that the factorisation obtained for the distribution p_{X_1,X_2,X_3} is

$$p_{X_1,X_2,X_3} = p_{X_1} p_{X_2} p_{X_3|X_1,X_2}.$$

In the corresponding DAG, $X_1 \not\perp X_3 \| _{\mathcal{G}} \phi$ and $X_2 \not\perp X_3 \| _{\mathcal{G}} \phi$, even though the independence statements hold.

An Aside on Faithfulness

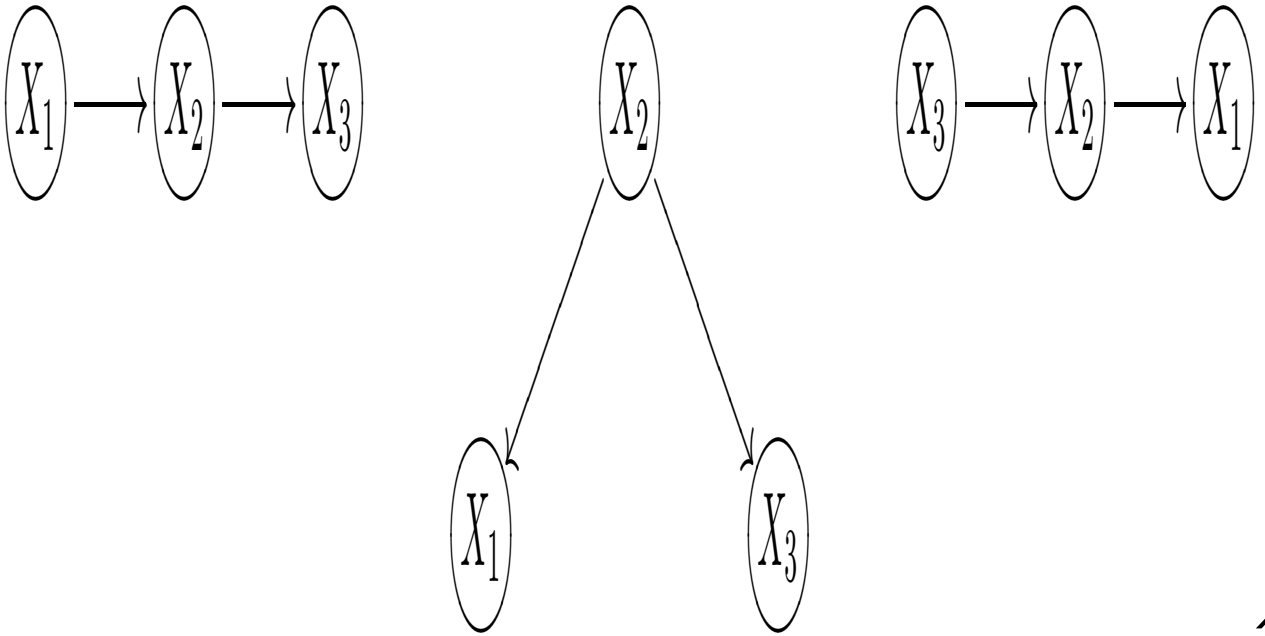
By considering other orderings of the variables, the other possible factorisations are

$$p_{X_1,X_2,X_3} = p_{X_1} p_{X_3} p_{X_2|X_1,X_3} = p_{X_2} p_{X_3} p_{X_1|X_2,X_3};$$

in none of the cases do the d-separation statements of the DAG represent all the conditional independence statements of the distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov Equivalence



$$X_1 \perp X_3 | X_2$$

I-sub-map, I-map, I-equivalence, Markov Equivalence

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two DAGs over the same variables. The DAG $\mathcal{G}_1$ is said to be an *I-sub-map* of $\mathcal{G}_2$ if any pair of variables d-separated by a set in $\mathcal{G}_1$ are also d-separated by the same set in $\mathcal{G}_2$. That is, the set of d-separation statements for $\mathcal{G}_1$ is a subset of the set of d-separation statements for $\mathcal{G}_2$. They are said to be *I-equivalent* if $\mathcal{G}_1$ is an I-sub-map of $\mathcal{G}_2$ and $\mathcal{G}_2$ is an I-sub-map of $\mathcal{G}_1$.
I-equivalence is also known as *Markov equivalence*.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov Equivalence

Theorem

Two DAGs are Markov equivalent if and only if they have the same skeleton *and the same* immoralities*.*

Andersson, S.A and Madigan, D. and Perlman, M.D. : A characterization of Markov equivalence classes for acyclic digraphs, *The Annals of Statistics*, 25, 505−541.

KTH Matematik

Markov Equivalence

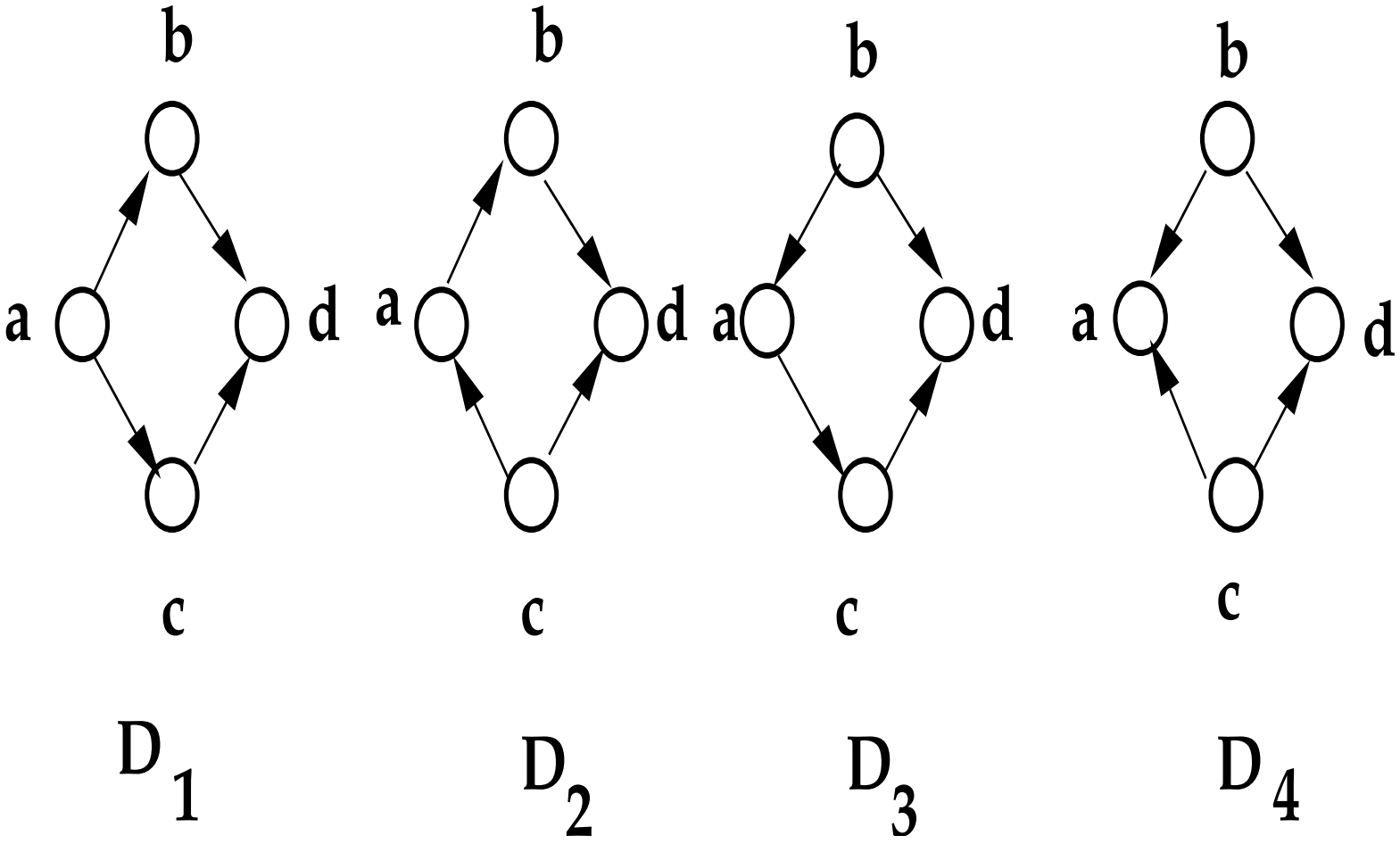
Definition (Immorality)

Let $\mathcal{G} = (V, E)$ be a DAG. An *immorality* in a graph is a triple of nodes (α, β, γ) such that $(\alpha, \beta) \in E$ and $(\gamma, \beta) \in E$, but $(\alpha, \gamma) \notin E$, $(\gamma, \alpha) \notin E$.

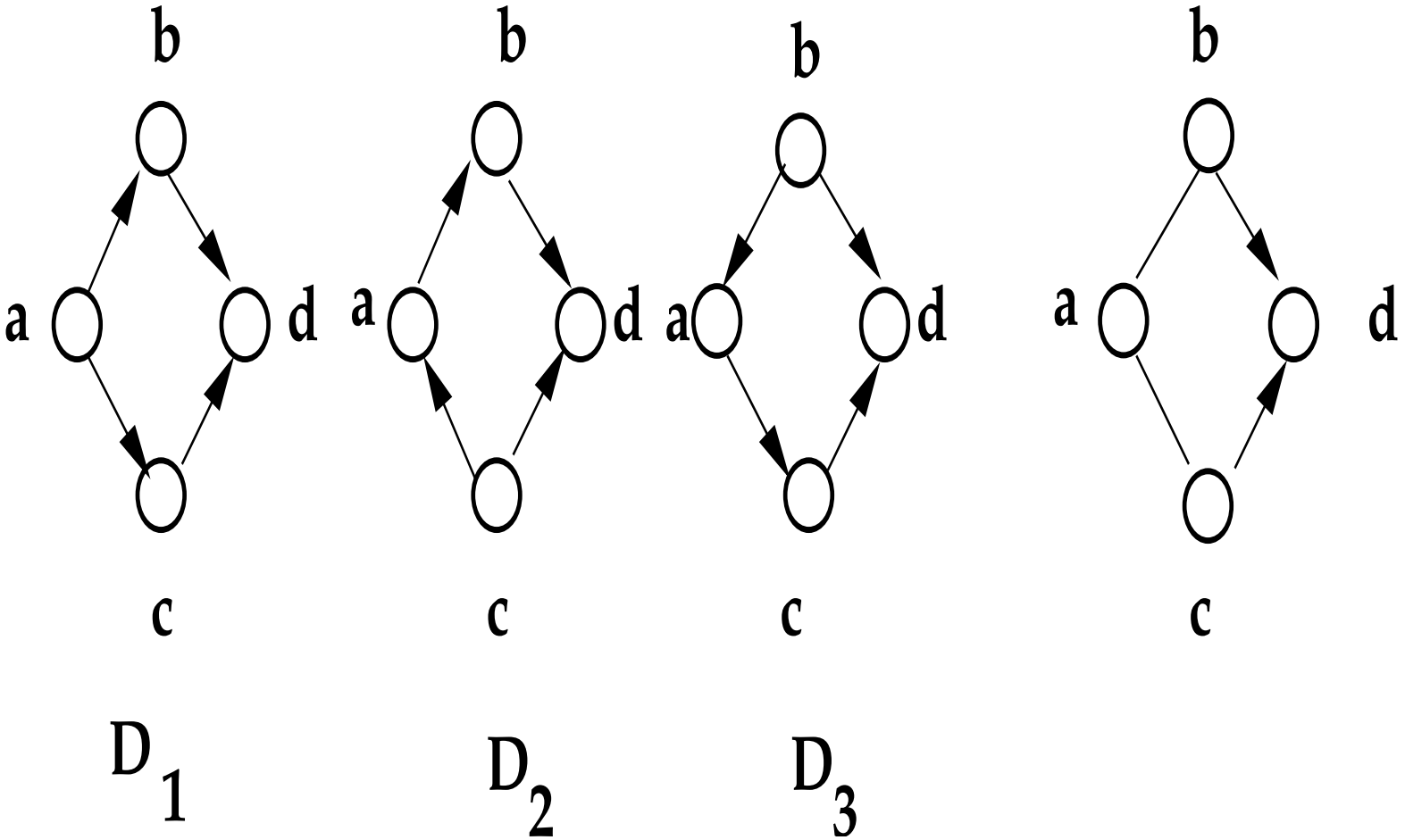
KTH VETENSKAP OCH KONST

KTH Matematik

Example



Example



Markov Equivalance

Each Markov-equivalence class is uniquely determined by a single PDAG (partial DAG=a chain graph), simultaneously equivalent to each graph in the class, hence the learning of structure should be based on these representatives. The question is, how to construct them.

KTH VETENSKAP OCH KONST
KTH Matematik

Essential Graph

Definition

Let $\mathcal{G}$ be a Directed Acyclic Graph. The *essential graph* $\mathcal{G}^*$ associated with $\mathcal{G}$ is the graph with the same skeleton as $\mathcal{G}$, but where an edge is directed in $\mathcal{G}^*$ if and only if it occurs as a directed edge with the same orientation in *every* DAG that is Markov equivalent to $\mathcal{G}$. The *directed* edges of $\mathcal{G}^*$ are the *essential* edges of $\mathcal{G}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Let P denote a distribution over the same set of variables as an LDAG G_L and let $\mathcal{I}(P)$ denote the set of CSIs satisfied by P. If P factorizes according to G_L, it must hold that $\mathcal{I}_{loc}(G_L) \subseteq \mathcal{I}(G_L) \subseteq \mathcal{I}(P)$ and G_L is called a CSI-map of P. There may, however, exist distribution-specific independences that hold in P even when they are not represented by the structure of G_L. A distribution P is said to be faithful to G_L if equality $\mathcal{I}(G_L) = \mathcal{I}(P)$ holds. The LDAG is then called a perfect CSI-map of P and can be considered a true representation in the sense that no artificial dependences are introduced.

KTH VETENSKAP OCH KONST
KTH Matematik

Equivalence

As for DAGs, the difference between two equivalent LDAGs can occur from reversing non-essential edges. It is worth noting that the criteria for an edge being essential will differ from DAGs. This observation is based on the fact that the direction of the edges determines which local CSIs may be included in an LDAG.

KTH VETENSKAP OCH KONST
KTH Matematik

CSI-Equivalence

Definition

CSI-equivalence for LDAGs
Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E^*, \mathcal{L}_E^*)$ be two distinct regular maximal LDAGs. The LDAGs are said to be CSI-equivalent if $\mathcal{I}(G_L) = \mathcal{I}(G_L^*)$. A set containing all CSI-equivalent LDAGs forms a CSI-equivalence class.

KTH VETENSKAP OCH KONST
KTH Matematik

Equivalence

We will discuss some structural properties that two distinct LDAGs must fulfill to belong to the same CSI-equivalence class. We begin by considering the underlying DAG.

Theorem

Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E^*, \mathcal{L}_E^*)$ be two regular maximal LDAGs belonging to the same CSI-equivalence class. Their underlying DAGs $G = (V, E)$ and $G^* = (V, E^*)$ must then have the same skeleton.

KTH VETENSKAP OCH KONST
KTH Matematik

Equivalence

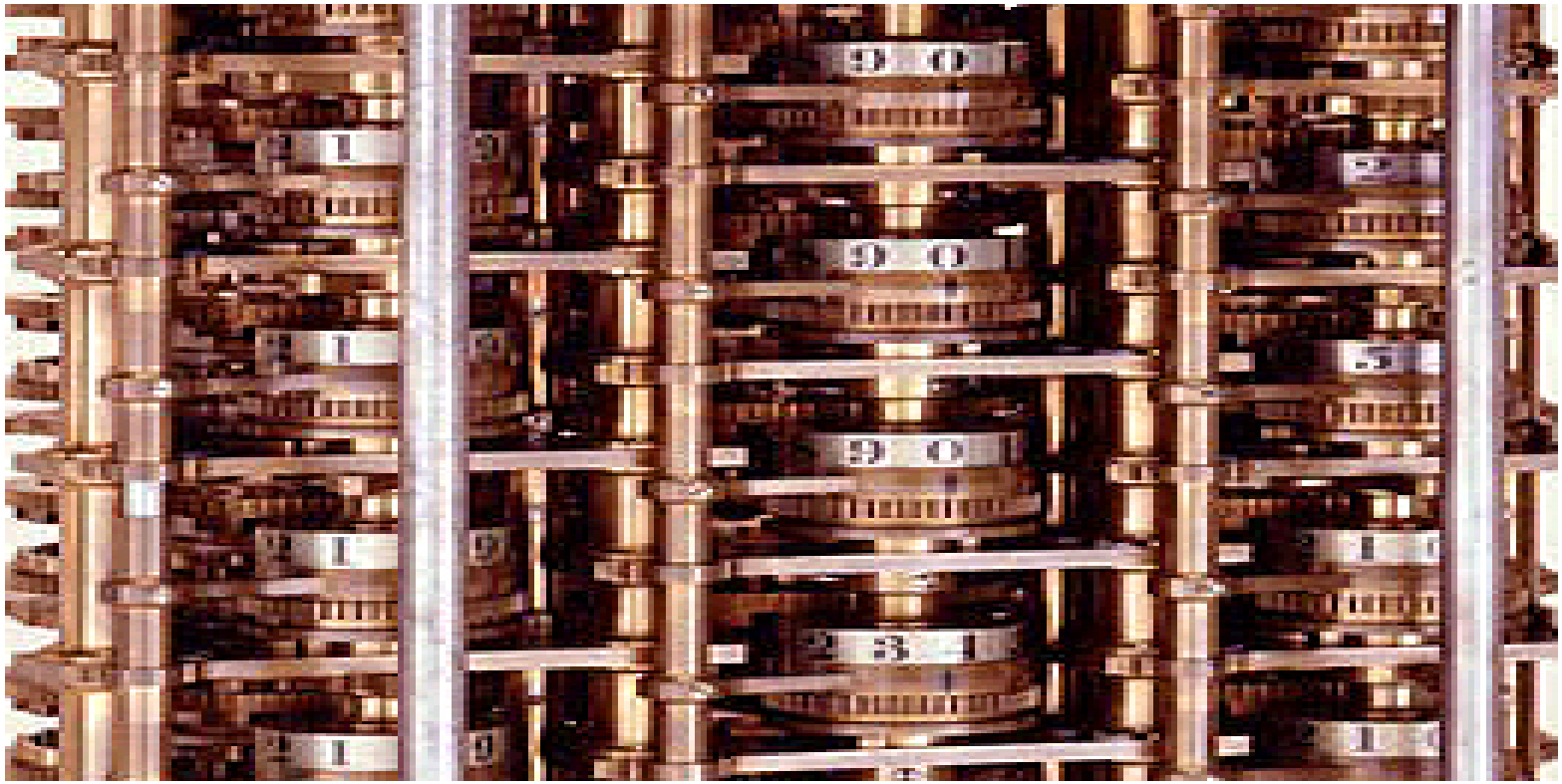
Next we introduce a criterion that ties together the concept of CSI-equivalence among LDAGs and the concept of Markov equivalence among DAGs.

Theorem

Let $G_L = (V, E, \mathcal{L}_E)$ and $G_L^* = (V, E^*, \mathcal{L}_E^*)$ be two maximal regular LDAGs for which there exists distributions P and P^* such that $\mathcal{I}(G_L) = \mathcal{I}(P)$ and $\mathcal{I}(G_L^*) = \mathcal{I}(P^*)$. G_L and G_L^* are CSI-equivalent if and only if their corresponding context-specific LDAGs $G_L(x_V) = G(x_V)$ and $G_L^*(x_V) = G^*(x_V)$ are Markov equivalent for all $x_V \in \mathcal{X}_V$.

KTH VETENSKAP OCH KONST

KTH Matematik



KTH Matematik

KTH Matematik

Bayesian learning of LDAGs

Learning the LDAG structure from a set of data poses some obvious problems due to the extremely vast model space as well as some additional not so obvious problems due to the flexibility of the models. We introduce a structural learning method that utilizes a non-reversible Markov Chain Monte Carlo (MCMC) method combined with greedy hill climbing. Such a combination of a stochastic and a deterministic algorithm provides solid performance with a reasonable time complexity.

KTH Matematik

Learning

A Bayesian score is used to evaluate the appropriateness of an LDAG given a set of observed data. In order to prevent overfitting, we impose a prior distribution that allows us to balance the ability of an LDAG to match the available learning data with its complexity.

KTH VETENSKAP OCH KONST

KTH Matematik

Learning

We begin with some additional notations.
Let $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$ denote a set of training data consisting of n observations $\mathbf{x}_i = (x_{i1}, \ldots x_{id})$ of the variables $\{X_1, \ldots, X_d\}$ such that $\mathbf{x}_i \in \mathcal{X}$. We assume that $\mathbf{X}$ is complete in the sense that it contains no missing values. We denote an LDAG by G_L and $\mathcal{G}_L$ denotes the set of all regular maximal LDAGs. We let Θ_{G_L} denote the parameter space induced by an LDAG and $dim(\Theta_{G_L})$ denotes the number of free parameters spanning the parameter space.

KTH Matematik

An instance $\theta \in \Theta_{G_L}$ corresponds to a specific joint distribution that factorizes according to the LDAG G_L. The CSI-consistent partition of the outcome space $\mathcal{X}_{\Pi_j}$ is denoted by $\mathcal{S}_{\Pi_j} = \{S_{j1}, \ldots, S_{jk_j}\}$ where $k_j = |\mathcal{S}_{\Pi_j}|$ is the number of outcome classes. We let $r_j = |\mathcal{X}_j|$ and $q_j = |\mathcal{X}_{\Pi_j}|$ denote the cardinality of the outcome space of variable X_j and its parents X_{Π_j}, respectively. Finally, we use $n(x_{ij} \times S_{jl})$ to denote the total count of the configurations $\{x_{ij}\} \times S_{jl}$ in $\mathbf{X}$.

KTH VETENSKAP OCH KONST
KTH Matematik

$p(\mathbf{X} \mid G_L)$ is the marginal probability of observing the data $\mathbf{X}$ (evidence) given a specific LDAG G_L and $p(G_L)$ denotes the prior probability of the LDAG.

$$\arg \max_{G_L \in \mathcal{G}_L} p(\mathbf{X} \mid G_L) \cdot p(G_L). \tag{5}$$

Marginal Data Distribution

The first issue is to compute the marginal data distribution

$$P(\mathbf{X} \mid \mathbf{G})$$

taking into account the equivalence classes. The solution due to

- M. Frydenberg: The Chain Graph Markov Property. *Scand. J. Stat.*, 17, 333−5353, 1990.
- S.A.Andersson, D. Madigan & M.D. Perlman: A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *The Annals of Statistics*, 25, 505−541, 1997

KTH VETENSKAP OCH KONST
KTH Matematik

To evaluate $p(\mathbf{X} \mid G_L)$, we need to consider all possible instances of the parameter vector satisfying the independencies encoded by the LDAG and weight them with respect to a prior according to

$$p(\mathbf{X} \mid G_L) = \int_{\theta \in \Theta_{G_L}} p(\mathbf{X} \mid G_L, \theta) \cdot f(\theta \mid G_L) d\theta, \tag{6}$$

where $p(\mathbf{X} \mid G_L, \theta)$ and $f(\theta \mid G_L)$ are the respective likelihood function and prior distribution over the parameters, given the graph G_L.

$$p(\mathbf{X} \mid G_L) = \prod_{j=1}^{d} \prod_{l=1}^{k_j} \frac{\Gamma\left(\sum_{i=1}^{r_j} \alpha_{ijl}\right)}{\Gamma\left(n(S_{jl}) + \sum_{i=1}^{r_j} \alpha_{ijl}\right)} \prod_{i=1}^{r_j} \frac{\Gamma\left(n(x_{ji} \times S_{jl}) + \alpha_{ijl}\right)}{\Gamma\left(\alpha_{ijl}\right)}, \tag{7}$$

where $n(\cdot)$ is the count defined earlier and the α_{ijl}:s are hyperparameters (also known as pseudocounts) defining a collection of local Dirichlet distributions. The hyperparameters characterize our prior belief about the CPDs and must be established to evaluate.

Appendix: Dirichlet

A Dirichlet distribution on the set

$$\{\underline{\theta} = (\theta_i)_{i=1}^d \mid \theta_i \geq 0, \sum_{i=1}^d \theta_i = 1\}$$

is defined by the probability density

$$p\left(\underline{\theta}\right) = \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d \theta_i^{\alpha_i - 1}$$

where $\alpha_i \geq 0$ is a hyperparameter and $\Gamma(z)$ is the Euler gamma function.

Learning: Prior

$$\alpha_{ijl} = \frac{N}{r_j \cdot q_j} \cdot |S_{jl}|, \tag{8}$$

where q_j is with respect to the underlying DAG and $|S_{jl}|$ denotes the number of configurations in that specific part. This is based on the thinking around Perks prior.

- W. Perks: Some observations on inverse probability including a new indifference rule,*Journal of the Institute of Actuaries*, pp. 285−334, 1947.

KTH VETENSKAP OCH KONST
KTH Matematik

The remaining issue at this point is to define the prior distribution over the set of LDAGs. This part is generally not given too much attention in Bayesian model learning but for LDAGs it plays a vital role.

$$p(G_L) \propto \kappa^{dim(\Theta_{G_L})} = \prod_{j=1}^{d} \kappa^{dim(\Theta_{G_L}(j))} \tag{9}$$

where $\kappa \in (0, 1]$.

The choice of model prior turns out to be an essential part of the Bayesian scoring function for LDAGs. We demonstrate below that the marginal likelihood alone has a tendency to overfit the dependence structure for limited sample sizes by favoring dense graphs with complex labelings. The number of free parameters associated with such a LDAG is low compared to the number of free parameters associated with its underlying DAG and the LDAG is said to have a high CSI-complexity.

The overfitting effect is thus reflected through a high CSI-complexity rather than an excessive number of free parameters. Although high CSI-complexity models may lead to high marginal likelihoods, they are more prone to contain false dependencies and thereby fail to capture the true global dependence structure. This has a direct negative effect on their out-of-sample predictive performance. Another drawback is that their high density will yield bulky CPDs. This basically counteracts the fundamental idea of modularity on which the concept of graphical models is based.

The overfitting phenomenon vanishes asymptotically when $n \to \infty$, since maximization of the marginal likelihood leads to a consistent estimator of the model structure. Consequently, we construct our prior such that it acts as a regularizer for smaller sample sizes and its effect will gradually vanish as the sample size is increased,

$$p(G_L) \propto \kappa^{dim(\Theta_G) - dim(\Theta_{G_L})} = \prod_{j=1}^{d} \kappa^{dim(\Theta_{G(j)}) - dim(\Theta_{G_L(j)})}, \qquad (10)$$

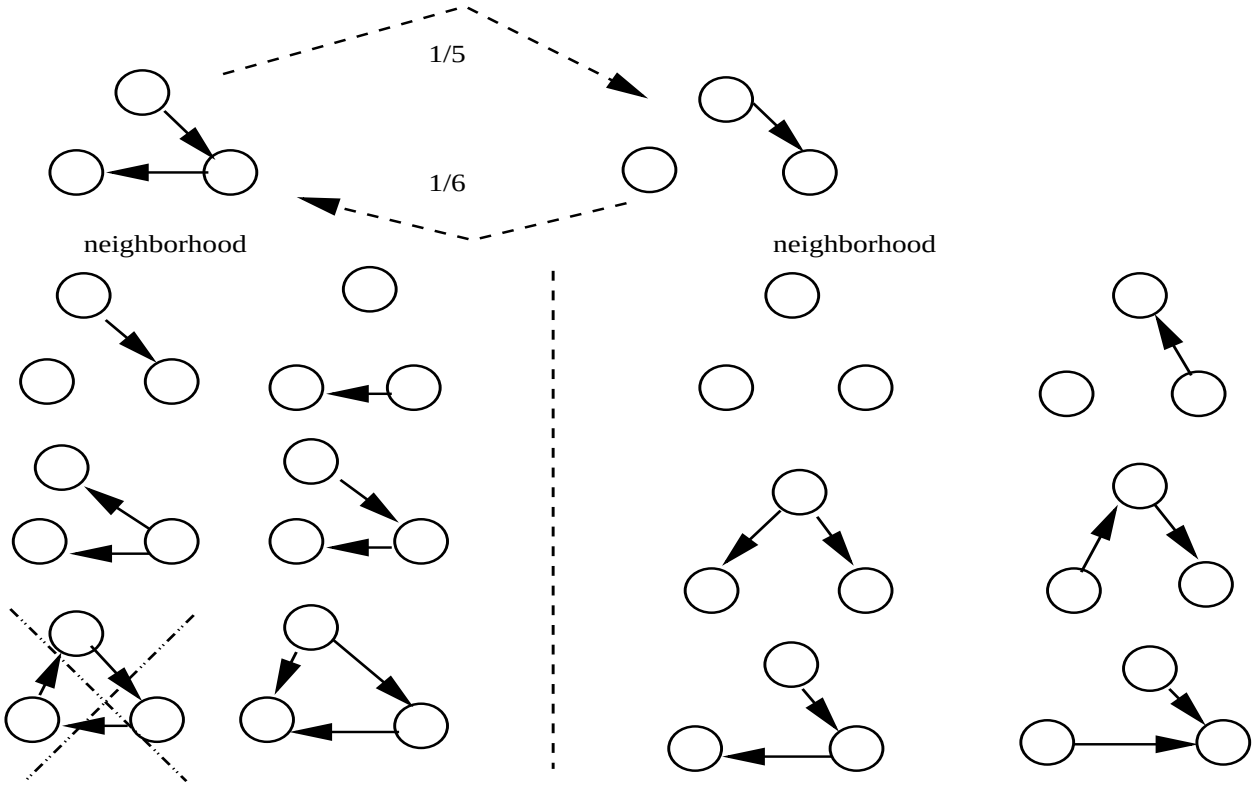
where $dim(\Theta_{G_L})$ and $dim(\Theta_G)$ are the number of free parameters associated with the LDAG and its underlying DAG, respectively. The parameter $\kappa \in (0, 1]$ can be considered a measure of how strongly a CSI inducing label configuration must be supported by the data in order for it to be included in the model.

For small values on κ, addition of a label configuration increases the score only if its associated CSI is firmly supported by the data while $\kappa = 1$ corresponds to a uniform prior.

Learning

For this purpose we introduce a search algorithm which utilizes a non-reversible MCMC method, combined with a direct form of optimization. The general idea is that the stochastic part of the algorithm jumps between neighbouring underlying DAGs, whose CSI structures are optimized by adding labels in a *greedy hill climbing*-manner. As our score decomposes variable-wise, instead of considering the whole DAG, we can optimize the local structure of one variable at a time.

KTH VETENSKAP OCH KONST
KTH Matematik


1/5
1/6
neighborhood
neighborhood

Procedure for optimizing the local CSI structure for X_j

X_j, //Variable whose local structure is optimized
X_{Π_j}, //Parental variables
$\mathbf{X}$, //A set of complete data over $X_{\Pi_j \cup \{j\}}$
)

1: $\mathcal{L}_j = \{\mathcal{L}_{(i,j)}\}_{i \in \Pi_j} \leftarrow \varnothing$
2: *keepClimb* $\leftarrow$ *True*
3: **while** *keepClimb*
4: $\quad \mathcal{L}_j^{top} \leftarrow \mathcal{L}_j$
5: $\quad$ **for** $x_{L_{(i,j)}} \notin \mathcal{L}_j : \{x_{L_{(i,j)}} \cup \mathcal{L}_{(i,j)}\} \subset \mathcal{X}_{L_{(i,j)}}$
6: $\quad\quad \mathcal{L}_j^{cand} \leftarrow \mathcal{L}_j \cup x_{L_{(i,j)}}$
7: $\quad\quad$ **if** $p(\mathbf{X}_j \mid \mathbf{X}_{\Pi_j}, \mathcal{L}_j^{cand}) > p(\mathbf{X}_j \mid \mathbf{X}_{\Pi_j}, \mathcal{L}_j^{top})$
8: $\quad\quad\quad \mathcal{L}_j^{top} \leftarrow \mathcal{L}_j^{cand}$
9: $\quad\quad$ **end**
10: $\quad$ **end**

KTH Matematik

We utilize a non-reversible version which has been shown to possess several advantageous properties. Let $q(\cdot|G_L)$ denote a generic proposal distribution over the model space $\mathcal{G}_L^{opt}$, given G_L for all $G_L \in \mathcal{G}_L^{opt}$. We let $G_L(t)$ denote the state of the chain at iteration t. At iteration $t = 1, 2, \ldots$ of the non-reversible chain, $q(\cdot|G_L(t))$ is used to generate the next candidate state G_L^* which is then accepted with probability

$$\min\left(1, \frac{p(G_L^*)p(\mathbf{X}|G_L^*)}{p(G_L(t))p(\mathbf{X}|G_L(t))}\right).$$

If G_L^* is accepted, we set $G_L(t+1) = G_L^*$ and otherwise $G_L(t+1) = G_L(t)$.

KTH VETENSKAP OCH KONST
KTH Matematik

The proposal probabilities need not to be explicitly calculated or even known as long as they remain unchanged over the iterations and the resulting chain is irreducible. The stationary distribution of such a chain does no longer follow the posterior distribution. However, our main objective is to identify only the maximum a posteriori model (5).

Non-reversible MCMC

The algorithm is based on m parallell Markov chains, $j = 1, 2, \ldots, m$ The chains draw at independent random times a value from $P(\mathbf{G}_j \mid \mathbf{X})$. This forces the chains to move to regions of maximal posterior probability.
Define the sequence of probabilities $\{\alpha_t, t = 2, 3, ...\}$ according to

$$\alpha_t = \frac{1}{q \log t},$$

where $q \geq 1$ can be chosen suitably, for instance $q \in [5, 10]$.
$Z_0 = 0$, and $P(Z_t = 1) = \alpha_t$, $P(Z_t = 0) = 1 - \alpha_t$, independently for $t = 1, 2, ...$.

KTH VETENSKAP OCH KONST
KTH Matematik

The Parallel Search

Find

$$\max P\left(\mathbf{G}_{tj} \mid \mathbf{X}\right)$$

over the space of the current states $\left\{\mathbf{G}_{t1}, \mathbf{G}_{t2}, ..., \mathbf{G}_{tm}\right\}$. For each $t = 0, 1, ...$ such that $Z_t = 1$, the transition to the next state is determined according to this distribution, such that the next state for each chain is sampled to the non-reversible proposal-acceptance formulae, independently for $j = 1, ..., m$.

KTH VETENSKAP OCH KONST

KTH Matematik

The Parallel Search

For each t, such that $Z_t = 0$, transition to the next state $G_{(t+1)j}$ is determined according to the non-reversible proposal-acceptance formulae above independently for $j = 1, ..., m$.

Learning

The approximate solution proposed by a search chain at iteration t is simply the one with the highest score visited thus far. Satisfying the conditions mentioned, the proposal distributions are defined as uniform distributions over the globally adjacent LDAGs that can be reached by adding, reversing or removing a single edge under the restriction that the resulting LDAG is acyclic.

KTH VETENSKAP OCH KONST
KTH Matematik

Learning

As the difference between two successive graphs may only differ for a single edge, at most two local structures are modified at each step of the chain. Since our score $p(\mathbf{X}, G_L)$ decomposes variable-wise, only the modified local structures must be re-evaluated as the score for the rest of the variables remains unchanged. This idea can be further exploited when optimizing the local CSI-structures. At each step of the optimization procedure, we need only to re-evaluate the score with respect to the parts of the partition that are modified. For our algorithm in particular, only a single new factor is created for each added label configuration.

KTH Matematik

Adding of labels yields a flexibility that facilitates the identification of "weaker" edges that might be deemed non-existing in the model space of DAGs. However, optimization of the CSI-structure cannot make up for unrealistic global independence assumptions made by an inferior underlying DAG structure. Hence, a prerequisite for learning a good LDAG structure is that it is based on a sensible underlying DAG. Getting stuck at regions with inferior underlying DAGs, will have a more severe negative effect on the learned LDAGs than not finding the optimal CSI-structure. This motivates the fact that the stochastic part of our method performs global changes whereas the optimization of the CSI-structures is done in a deterministic manner.

Learning κ

For choosing an appropriate value of κ, we propose a cross-validation scheme. First we partition the data $\mathbf{X}$ into a training set $\mathbf{Y}$ and a test set $\mathbf{Z}$. We then apply our search method on the training data under some prior (or κ) and identify the optimal model G_L^κ. We then evaluate the learned model's ability to predict the test data by calculating the posterior predictive probability of the test data given the training data,

$$p(\mathbf{Z} \mid \mathbf{Y}, G_L^\kappa) = \int_{\theta \in \Theta_{G_L^\kappa}} p(\mathbf{Z} \mid G_L^\kappa, \theta) \cdot f(\theta \mid \mathbf{Y}, G_L^\kappa) d\theta. \qquad (11)$$

Note that this applies the notion of conditionally independent and identically distributed data given the parameters, c.f. the introductory discussion of Bayesian statistics in the beginning of these lectures.

KTH Matematik

Under similar assumptions made earlier, (11) can be calculated analytically by

$$p(\mathbf{Z} \mid \mathbf{Y}, G_L^{\kappa}) = \prod_{j=1}^{d} \prod_{l=1}^{k_j} \frac{\Gamma\left(\sum_{i=1}^{r_j} (\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl}))\right)}{\Gamma\left(n_{\mathbf{Z}}(S_{jl}) + \sum_{i=1}^{r_j} (\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl}))\right)} \cdot \prod_{i=1}^{r_j} \frac{\Gamma\left(n_{\mathbf{Z}}(x_{ji} \times S_{jl}) + \alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl})\right)}{\Gamma\left(\alpha_{ijl} + n_{\mathbf{Y}}(x_{ji} \times S_{jl})\right)}, \tag{12}$$

where the bold case index indicates to which data set the outcome count refers.

Learning κ

To reduce the variability of the method, multiple partitions of $\mathbf{X}$ are created, $\{(\mathbf{Y}_1, \mathbf{Z}_1), (\mathbf{Y}_2, \mathbf{Z}_2), \ldots, (\mathbf{Y}_M, \mathbf{Z}_M)\}$, and the validation results are averaged according to

$$\rho_{pred}(\kappa) = \frac{1}{M} \sum_{m=1}^{M} \log p(\mathbf{Z}_m \mid \mathbf{Y}_m, G_L^{\kappa,m}). \tag{13}$$

The value on κ is finally chosen among the candidates as the one that maximizes (13).

KTH Matematik

KTH Matematik

Experimental results with real and simulated data sets

To illustrate the properties of LDAGs, we apply our search algorithm on both a real and two simulated data sets. First we consider a real data set that has been thoroughly investigated in earlier graphical modelling literature. After that we consider synthetic DAG- and LDAG-based models.

KTH VETENSKAP OCH KONST
KTH Matematik

Experimental results with real and simulated data sets

Our real data set contains 1841 cases composed of six binary risk factors for coronary heart disease.

KTH VETENSKAP OCH KONST

KTH Matematik

Table: Prognostic factors in coronary heart disease.

				X_2	yes		no	
X_6	X_5	X_4	X_3	X_1	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		> 140	no		35	12	80	33
			yes		109	67	7	9
	> 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		> 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		> 140	no		4	3	11	8
			yes		14	17	5	2
	> 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		> 140	no		4	0	13	11
			yes		5	14	4	4

Table: Explanations of the Labels in Table 1

Label	Meaning	Range
X_1	smoking	no,yes
X_2	strenuous mental work	no,yes
X_3	strenuous physical work	no,yes
X_4	systolic blood pressure	$< 140, > 140$
X_5	ratio of β and α lipoproteins	$< 3, > 3$
X_6	family anamnesis[1] of coronary heart disease	no,yes

[1] information concerning a medical patient and his/her background for use in analysis of her/his condition

KTH VETENSKAP OCH KONST
KTH Matematik

Here we get an indication of how the CSI-complexity increases with higher values on κ. The bold font indicate which κ was chosen as optimal by the cross-validation procedure. The LDAG identified for $\kappa = 0.001$ contains no labels and is thereby equal to its underlying DAG. The improvement, that an added label configuration induces to the marginal likelihood, is overshadowed by the simultaneous lowering of the prior probability mass.

KTH VETENSKAP OCH KONST

KTH Matematik

Experimental results with a real data set

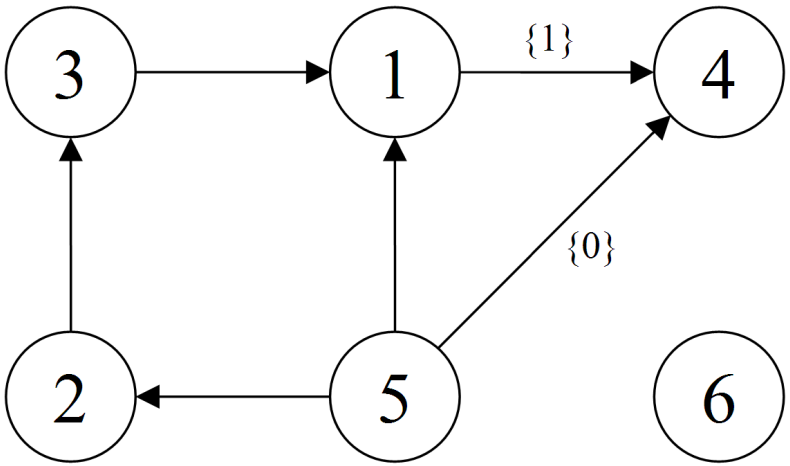
Consequently, when $\kappa \to 0$ the direct optimization will map the set of DAGs onto itself and the learning procedure is reduced to a search among ordinary DAGs.

κ	$\log p(\mathbf{X}, G_L)$	$\lvert E \rvert$	$dim(\Theta_G)$	$dim(\Theta_{G_L})$	ρ_{pred}
0.001	-6731.82	5	12	12	-671.30
0.1	-6729.69	6	14	12	-670.88
0.3	**-6727.50**	**6**	**14**	**12**	**-670.51**
0.5	-6724.68	7	18	11	-670.89

Table: Properties of identified LDAGs for coronary heart disease data.

KTH VETENSKAP OCH KONST
KTH Matematik

Experimental results with a real data set



Experimental results with a real data set

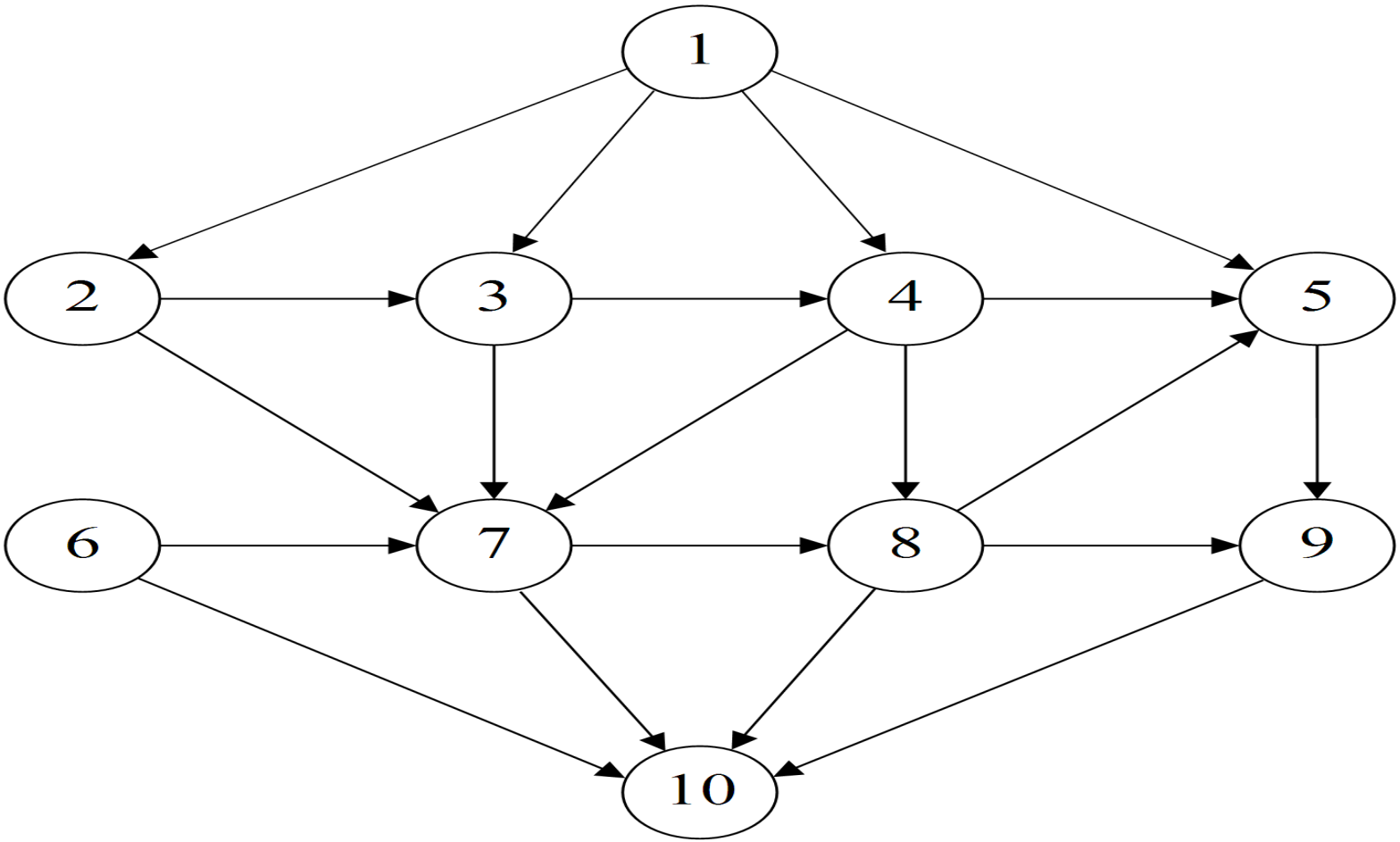
We now consider synthetic models from which data are generated to systematically compare models identified for different prior distributions and sample sizes. Since we know the generating model, we investigate how well the identified models approximate the true distribution. The CPDs of the models are estimated by the consistent mean a posteriori estimator as the expected value of the local posterior Dirichlet distributions. To compare the distributions, we utilize the concept of Kullback-Leibler (KL) divergence.

$$D_{KL}(p \parallel p^*) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{p^*(x)}.$$

KTH VETENSKAP OCH KONST

KTH Matematik

DAG and labels according to which the synthetic data sets were generated



KTH Matematik

DAG and labels according to which the synthetic data sets were generated

$$
\begin{aligned}
\mathcal{L}_3: \quad & \mathcal{L}_{(2,3)} = \{0\} \\
\mathcal{L}_4: \quad & \mathcal{L}_{(1,4)} = \{1\} \\
\mathcal{L}_5: \quad & \mathcal{L}_{(4,5)} = \{(0,*)\} \\
& \mathcal{L}_{(8,5)} = \{(0,*)\} \\
\mathcal{L}_7: \quad & \mathcal{L}_{(2,7)} = \{(1,1,0)\} \\
& \mathcal{L}_{(3,7)} = \{(0,1,1),(1,*,1)\} \\
& \mathcal{L}_{(4,7)} = \{(1,1,*)\} \\
& \mathcal{L}_{(6,7)} = \{(1,1,*)\}
\end{aligned}
$$

KTH VETENSKAP OCH KONST

KTH Matematik

More labels

$$
\begin{aligned}
\mathcal{L}_9: \quad & \mathcal{L}_{(5,9)} = \{1\} \\
\mathcal{L}_{10}: \quad & \mathcal{L}_{(7,10)} = \{(1,*,*)\} \\
& \mathcal{L}_{(8,10)} = \{(1,*,*)\} \\
& \mathcal{L}_{(9,10)} = \{(1,*,*)\}
\end{aligned}
$$

KTH VETENSKAP OCH KONST

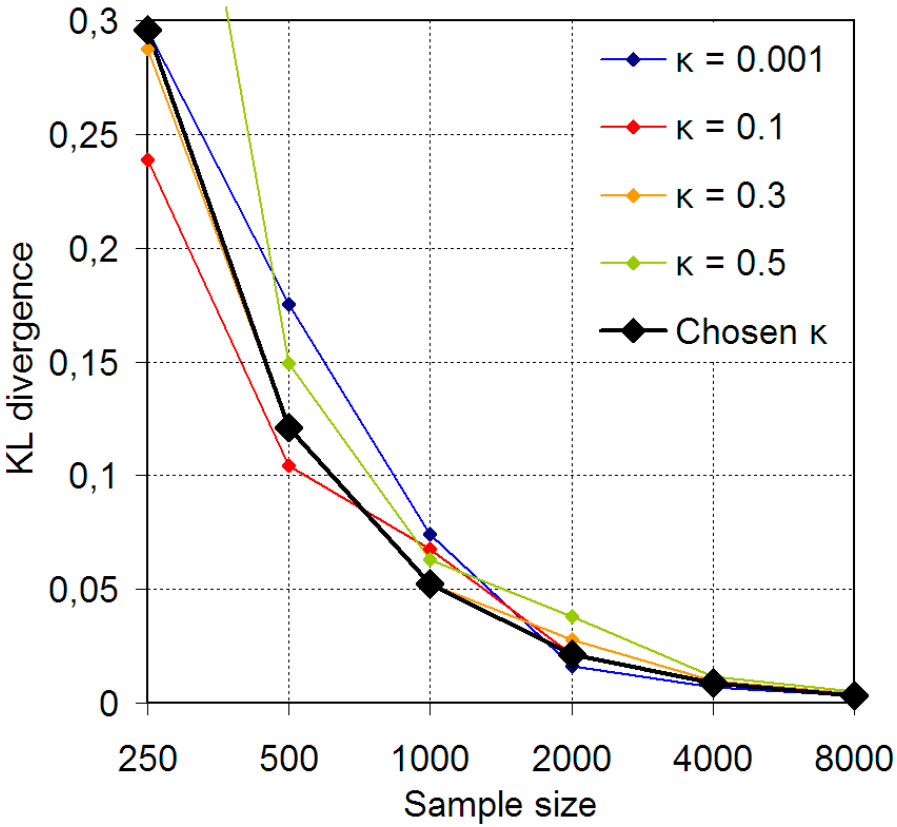
KTH Matematik

Experimental results with simulated data sets

As expected, the model distributions approach the true distribution when the sample size increases. This results in a steady improvement of the KL divergence as illustrated in Figure **??**. The decrease is evident for all values on κ but our results indicate that different prior distributions are to be preferred depending on the sample size. It also clear how the quality of most of the models begin to suffer under $\kappa = 0.5$ as a result of overfitting.

KTH Matematik

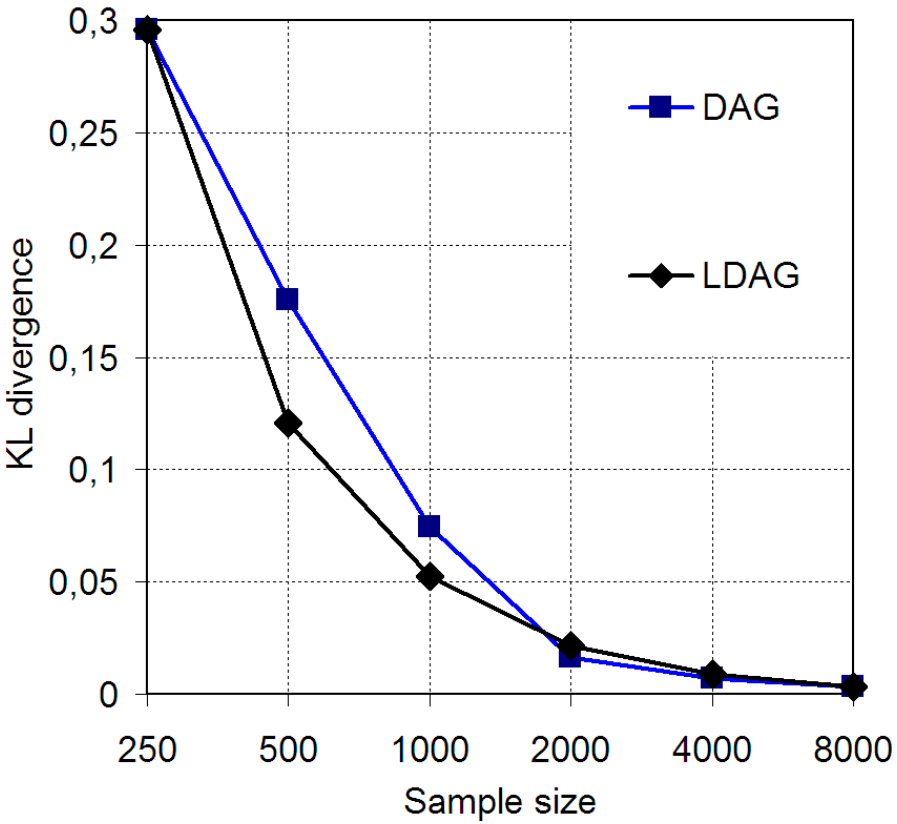
KL divergence for different sample sizes under different priors



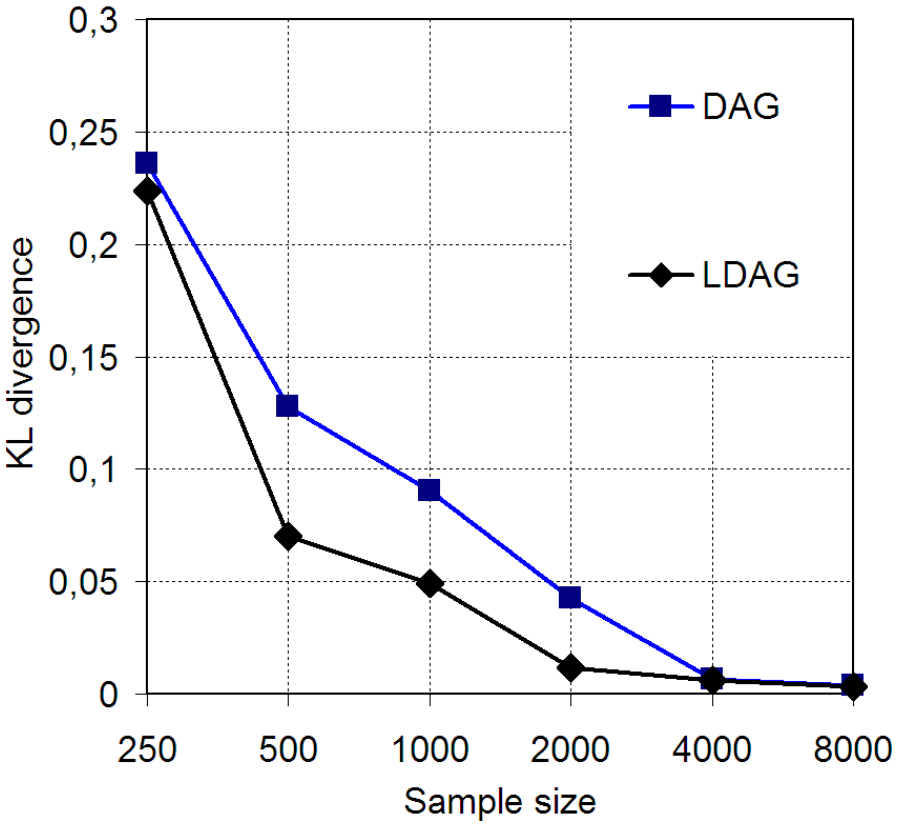
KTH Matematik

Comments:DAG-based generating model

All models identified under $\kappa = 0.001$ are without labels since $dim(\Theta_G) - dim(\Theta_{G_L}) = 0$. We can thus use this prior as a reference point for investigating how well LDAGs perform compared to traditional DAGs.



Comments:LDAG-based generating model



KTH Matematik

Comparison of DAGs and LDAGs for different sample sizes.

The difference in KL divergence between the true distribution and the approximate distributions induced by the models. The DAG curve in the figure corresponds to the 0.001-curve from Figure **??** and the LDAG curve corresponds to the thick black curve where the models where chosen by the initial cross-validation method. Note that the method in some cases picks the 0.001-prior which results in a converging of the curves. We see that the LDAGs mostly outperform traditional DAGs by inducing distributions that better approximate the true distribution.

KTH Matematik

Thank you !

KTH Matematik