Likelihood Based Methods for Learning of Credal Networks

S. Moral
University of Granada - Spain

IMPRECISE PROBABILITIES WORKSHOP, 27th-29th of May 2015

Outline

- Bayesian Networks
- Parametric learning in Bayesian networks
- Credal networks. Learning the parameters
  - Imprecise Dirichlet Model (IDM)
  - Imprecise Sample Size Dirichlet Model (ISSDM)
  - Likelihood based inference for learning the parameters
- Structure learning of Bayesian networks.
- Credal networks. Learning the structure.

Bayesian Networks

Definition

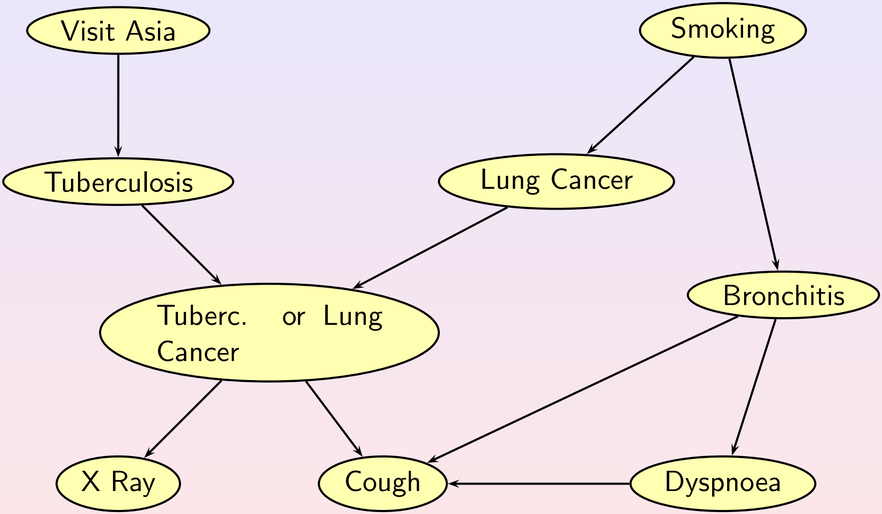
A Bayesian network for a set of variables $(X_1, \ldots, X_m)$ is a pair (G, Π) where G is a directed acyclic graph with a node for each variable X_i and Π is a list of conditional probability distributions $P(X_1|Pa_1), \ldots, P(X_m|Pa_m)$, one for each variable given its parents in G.

Meaning

- The graph G encodes a set of independent relationships: each variable X_i is independent of its non-descendent variables given its parents.
- The Bayesian network encodes the joint probability distribution:

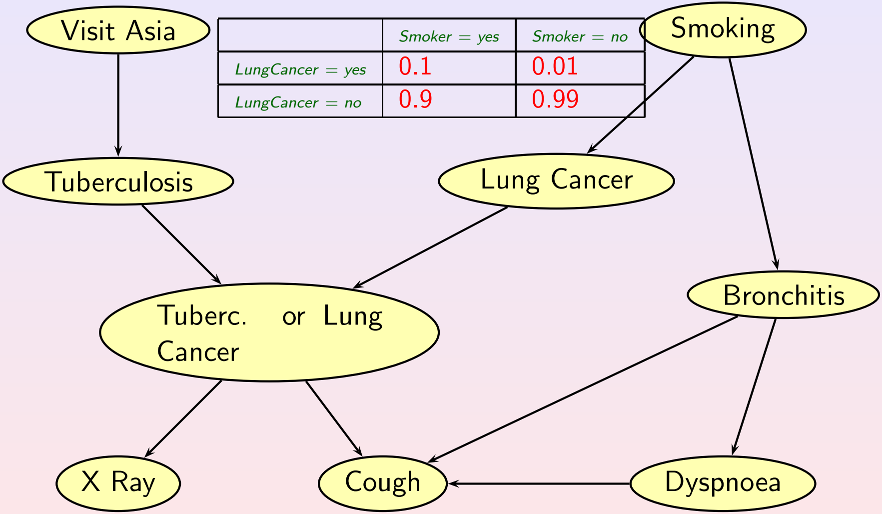
$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i|Pa_i)$$

Example


Visit Asia
Smoking
Tuberculosis
Lung Cancer
Bronchitis
Tuberc. or Lung Cancer
X Ray
Cough
Dyspnoea

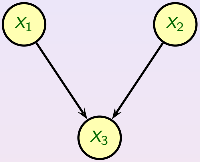
Example

	Smoker = *yes*	*Smoker* = *no*
LungCancer = *yes*	0.1	0.01
LungCancer = *no*	0.9	0.99



Conditional Distributions

$P(X_1 = 0)$	0.3
$P(X_1 = 1)$	0.7

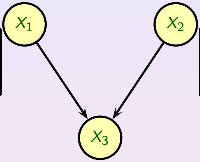


$P(X_2 = 0)$	0.6
$P(X_2 = 1)$	0.4

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	0.6	0.8	0.3	0.1
$P(X_3 = 1)$	0.4	0.2	0.7	0.9

Conditional Distributions: Parametrizations

$P(X_1 = 0)$	θ_{111}
$P(X_1 = 1)$	θ_{112}



$P(X_2 = 0)$	θ_{211}
$P(X_2 = 1)$	θ_{212}

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	θ_{311}	θ_{321}	θ_{331}	θ_{341}
$P(X_3 = 1)$	θ_{312}	θ_{322}	θ_{332}	θ_{342}

$$\theta_{ijk} \begin{cases} i & \text{variable} \\ j & \text{conditional distribution} \\ k & \text{case of variable} \end{cases}$$

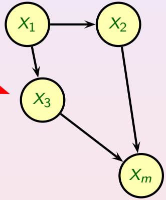
The $j-th$ conditional distribution for X_i

$$\theta_{ij} = (\theta_{ij1}, \theta_{ij2})$$

Learning Bayesian Networks

Learning in Bayesian networks can be defined as the process of inducing a model from a database.

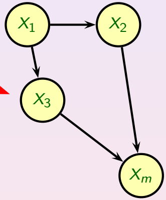
X_1	X_2	$\dots$	X_m
x_1^1	x_2^1	$\dots$	x_m^1
x_1^2	x_2^2	$\dots$	x_m^2
x_1^3	x_2^3	$\dots$	x_m^3
x_1^4	x_2^4	$\dots$	x_m^4



Learning Bayesian Networks

Learning in Bayesian networks can be defined as the process of inducing a model from a database.

X_1	X_2	$\ldots$	X_m
x_1^1	x_2^1	$\ldots$	x_m^1
x_1^2	x_2^2	$\ldots$	x_m^2
x_1^3	x_2^3	$\ldots$	x_m^3
x_1^4	x_2^4	$\ldots$	x_m^4

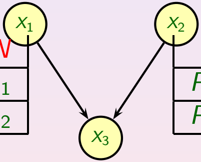


Learning = Inducing a graph + Estimating parameters

Learning Bayesian Networks: Parameter Estimation

- Usually a Bayesian approach is considered
- If prior distributions for the parameters of each conditional distribution θ_{ij} are independent and we do not have missing data, then the posterior is also independent, and we can decompose the problem in estimating each one of the conditional distributions θ_{ij}.

Assume the following network and a sample size N



	$N_{11} = N$
$P(X_1 = 0)$	θ_{111}
$P(X_1 = 1)$	θ_{112}

	$N_{21} = N$
$P(X_2 = 0)$	θ_{211}
$P(X_2 = 1)$	θ_{211}

$\sum_i N_{3i} = N$

	N_{31}	N_{32}	N_{33}	N_{34}
	$x_1 = 0$	$x_1 = 0$	$x_1 = 1$	$x_1 = 1$
	$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
$P(X_3 = 0)$	θ_{311}	θ_{321}	θ_{331}	θ_{341}
$P(X_3 = 1)$	θ_{312}	θ_{322}	θ_{332}	θ_{342}

- Without missing values, it can be decomposed in the estimation of a family of multinomial probabilities: one distribution for each conditional probability of a variable given a configuration or combination of values of its parents.
- It is important to notice that if the number of parents increase the sample size decreases (original sample splitted in an exponential number of subsamples).
- We shall now concentrate in the estimation of multinomial probabilities.

The Problem

- We have a random variable X taking values on a finite set $U = \{x_1, \ldots, x_k\}$
- Assume that $P(X = x_i) = \theta_i$
- $\theta = (\theta_1, \ldots, \theta_k)$
- We have N observations (iid) of this random variable: $D = (d_1, \ldots, d_N)$
- We want to estimate the parameters θ_i taking these observations as basis

Bayesian Models

Prior Density about the Parameter

Usually, a Dirichlet distribution $D(\alpha_1, \ldots, \alpha_k)$ with a density

$$f(\theta_1, \ldots, \theta_k) \propto \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

where $\alpha_i > 0$.

α_i: it is a weight for our prior belief in $P(X = x_i)$

Equivalent Sample Size

The value $s = \sum_{i=1}^{k} \alpha_i$ is called the equivalent sample size (relative importance of prior weights with respect to sample size)

Computational Advantages

Posterior Probability

If N_i is the number of occurrences of $X = x_i$ in D, then the posterior density $f(\theta|D)$ is also a Dirichlet density of parameters
$D(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$
where N_i is the number of observations of $X = x_i$ in the sample.

$$P(X = x_i|D) = \hat{\theta}_i = E[\theta_i|D] = \frac{N_i + \alpha_i}{N + s}$$

$X = x_1$	$X = x_2$	$X = x_3$
α_1	α_2	α_3

Computational Advantages

Posterior Probability

If N_i is the number of occurrences of $X = x_i$ in D, then the posterior density $f(\theta|D)$ is also a Dirichlet density of parameters
$D(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$
where N_i is the number of observations of $X = x_i$ in the sample.

$$P(X = x_i|D) = \hat{\theta}_i = E[\theta_i|D] = \frac{N_i + \alpha_i}{N + s}$$

$X = x_1$	$X = x_2$	$X = x_3$
α_1	α_2	α_3
N_1	N_2	N_3

Computational Advantages

Posterior Probability

If N_i is the number of occurrences of $X = x_i$ in D, then the posterior density $f(\theta|D)$ is also a Dirichlet density of parameters
$D(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$
where N_i is the number of observations of $X = x_i$ in the sample.

$$P(X = x_i|D) = \hat{\theta}_i = E[\theta_i|D] = \frac{N_i + \alpha_i}{N + s}$$

$X = x_1$	$X = x_2$	$X = x_3$
α_1	α_2	α_3
N_1	N_2	N_3
$N_1 + \alpha_1$	$N_2 + \alpha_2$	$N_3 + \alpha_3$

Computational Advantages

Posterior Probability

If N_i is the number of occurrences of $X = x_i$ in D, then the posterior density $f(\theta|D)$ is also a Dirichlet density of parameters $D(\alpha_1 + N_1, \ldots, \alpha_k + N_k)$
where N_i is the number of observations of $X = x_i$ in the sample.

$$P(X = x_i|D) = \hat{\theta}_i = E[\theta_i|D] = \frac{N_i + \alpha_i}{N + s}$$

$X = x_1$	$X = x_2$	$X = x_3$
α_1	α_2	α_3
N_1	N_2	N_3
$N_1 + \alpha_1$	$N_2 + \alpha_2$	$N_3 + \alpha_3$
$\frac{N_1+\alpha_1}{N+s}$	$\frac{N_2+\alpha_2}{N+s}$	$\frac{N_3+\alpha_3}{N+s}$

Prior ignorance: Symmetry. Objective Bayesian Models

Prior Ignorance: Symmetry Principle

Prior density is invariant under permutations:

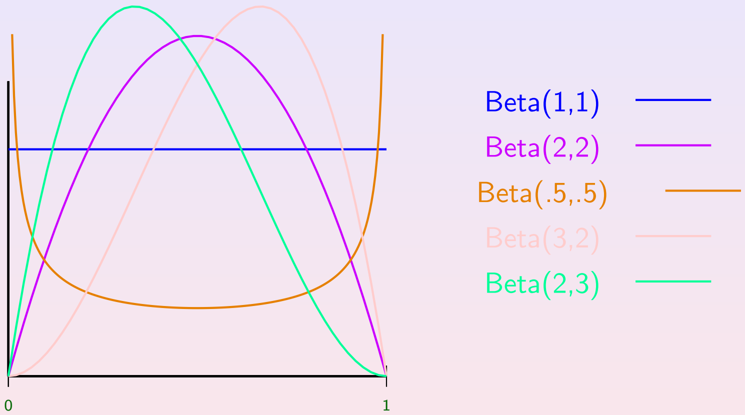
$$\alpha_1 = \cdots = \alpha_k = s/k; \quad P(X = a_i|D) = \hat{\theta}_i = E[\theta_i|D] = \frac{N_i + s/k}{N + s}$$

- Haldane (1948): $\alpha_i = 0$, $s = 0$ (maximum likelihood)
- Perks (1947): $\alpha_i = 1/k$, $s = 1$
- Jeffreys (1946,1961): $\alpha_i = 1/2$, $s = k/2$
- Bayes Laplace: $\alpha_i = 1$, $s = k$
- Berger-Bernardo: reference priors

Important Parameter

Equivalent Sample Size (s): *Relative importance of prior information with respect to the sample*

Beta shapes



Some Difficulties

Main Problem

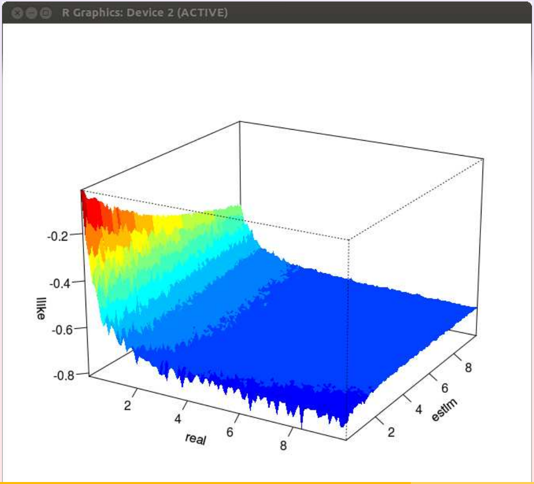
How to determine α_i and the equivalent sample size (s)?
These values assume that the parameters are generated according to some distributions.
If the real parameters have low density, the results can be poor.

Representation Invariance Principle (RIP)

Inferences should not depend on refinements or coarsenings of categories: if a category x_i is changed, the estimation of the probabilities of unchanged categories should be the same.

Experiment

We generate samples of size 10 according to a symmetric Beta (α real) and estimate with another Beta (α estimated). We compute the expected log of the estimated values with respect to the real parameter.



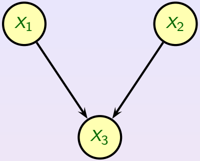
Going Back to Bayesian Networks

q_i number of conditional distributions of variable X_i
k_i number of values of variable X_i

- In a Bayesian network, the problem is more important, as the determination of s_{ij} for each distribution θ_{ij} can depend on the number of conditional distributions of X_i given its parents.
- We have to ways of selecting the parameters:
  - The local approach: Each α_{ijk} is selected with independence of q_i and k_i
  - The global approach: Each α_{ijk} depends of q_i and k_i

The Local Approach: An uniform α

$P(X_1 = 0)$	α
$P(X_1 = 1)$	α



$P(X_2 = 0)$	α
$P(X_2 = 1)$	α

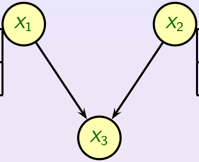
	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	α	α	α	α
$P(X_3 = 1)$	α	α	α	α

- It is the most usual approach in practice ($\alpha = 1$, Laplace correction)
- It is not considered correct, as equivalent networks (representing the same conditional independence relationships) give rise to different estimations.

The Global Approach: different α_{ijk}

It assumes a global Dirichlet distribution for all the variables

$P(X_1 = 0)$	α_{111}
$P(X_1 = 1)$	α_{112}



$P(X_2 = 0)$	α_{211}
$P(X_2 = 1)$	α_{212}

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	α_{311}	α_{321}	α_{331}	α_{341}
$P(X_3 = 1)$	α_{312}	α_{322}	α_{332}	α_{342}

- Some linear restrictions should be satisfied:

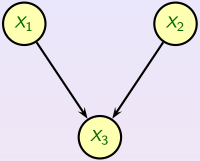
$$\sum_j \alpha_{31j} + \sum_j \alpha_{32j} = \alpha_{111}, \ \sum_j \alpha_{33j} + \sum_j \alpha_{34j} = \alpha_{112}$$

$$\sum_j \alpha_{31j} + \sum_j \alpha_{33j} = \alpha_{211}, \ \sum_j \alpha_{32j} + \sum_j \alpha_{34j} = \alpha_{212}$$

The Global Approach: different α_{ijk}

Under symmetry $\alpha_{ijk} = \alpha_{ijk'}$

$P(X_1 = 0)$	$s/2$
$P(X_1 = 1)$	$s/2$

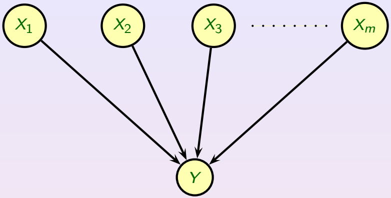


$P(X_2 = 0)$	$s/2$
$P(X_2 = 1)$	$s/2$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	$s/8$	$s/8$	$s/8$	$s/8$
$P(X_3 = 1)$	$s/8$	$s/8$	$s/8$	$s/8$

- s the global equivalent sample size
- $\alpha_{ijk} = \frac{s}{k_i q_i}$, where
  - k_i number of values of X_i
  - q_i number of configurations of parents of X_i (exponential in the number of parents)
  - $s_{ij} = \frac{s}{q_i}$ is the equivalent sample size for conditional distributions of X_i

Difficulty



Binary Variables

- If we consider an equivalent sample size of $S = 2$, then the marginal distributions about X_i is $D(1, 1)$ and the conditional distr. are $D(1/2^m, 1/2^m)$.
- It is possible that the sample with which we have estimated $P(y|\mathbf{x})$ is very short and the Dirichlet parameters are low too: very risky estimation.

The Imprecise Dirichlet Model

- Introduced by Walley(1996)
- Based on Imprecise Probability: it considers a set $\mathcal{P}$ of prior densities
- Updating is done by applying Bayes rule to each one of the densities in $\mathcal{P}$.
  $\mathcal{P}|D = \{f(.|D) \; : \; f \in \mathcal{P}\}$

Imprecise Dirichlet Model: Prior Information

s: Equivalent sample size.

$$\mathcal{P} = \{D(\alpha_1, \ldots, \alpha_k) \; : \; \sum_{i=1}^{k} \alpha_i = s, \; \alpha_i > 0\}$$

The Imprecise Dirichlet Model: Inferences

Imprecise Dirichlet Model: Prior Information

$$\mathcal{P} = \{D(\alpha_1, \ldots, \alpha_k) \; : \; \sum_{i=1}^{k} \alpha_i = s, \; \alpha_i > 0\}$$

Imprecise Dirichlet Model: Inferences

$$P(X = x_i|D) \in [\underline{P}(\theta_i|D), \overline{P}(\theta_i|D)] = \left[\frac{N_i + 0}{N + s}, \frac{N_i + s}{N + s}\right]$$

$s = 1$	x_1	x_2	x_3
N_i	3	4	0
Interv.	[3/8,4/8]	[4/8,5/8]	[0,1/8]

Example

Imagine that we have an urn with balls of different colors: red (R), blue (B), and green (G); but on an unknown quantity.

Assume that we picked up balls with replacement, with the following sequence: (B, B, R, R, B).

If we assume an imprecise Dirichlet 'a priori' distribution with $s = 3$, then the estimated intervals for red, blue, and green are:

	R	B	G
N_i	2	3	0
Int.	[2/8,5/8]	[3/8,6/8]	[0, 3/8]
Bayesian (Laplace)	3/8	4/8	1/8

Some properties

Properties

- If $N = 0$ the interval is $[0, 1]$
- The interval width is $s/(s + N)$ converging to 0 as N increases
- It satisfies the representation invariance principle

- When a category, for example B is divided between Dark Blue (B_1) and Light Blue (B_2), then the probability of red continues being $[2/8, 5/8]$. With Bayesian estimation if with 4 categories, we consider $D(0.75, 0.75, 0.75, 0.75)$, then the estimation of the probability of Red changes!

Credal Networks: Imprecise Probabilities in Bayesian Networks

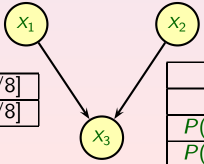
Credal Network, Cozman (2000)

It is a graph G and a set of probability distributions $\mathcal{P}$ such that each $P \in \mathcal{P}$ factorices according to G:

$$P(x) = \prod_i P_i(x_i|Pa_i)$$

Separately Specified Credal Network, Cozman (2000)

It is a graph G and a set of probability distributions for each variable X_i and each possible value of each of its parents



$P(X_1 = 0)$	[3/8,4/8]
$P(X_1 = 1)$	[4/8,3/8]

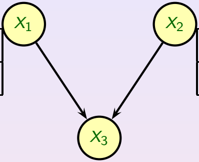
$P(X_2 = 0)$	[1/8,2/8]
$P(X_2 = 1)$	[6/8,7/8]

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	[1/2,1]	[1/3,2/3]	[0,1]	[1/5,2/5]
$P(X_3 = 1)$	[0,1/2]	[1/3,2/3]	[0,1]	[3/5,4/5]

An IDM for each conditional distribution:

$P(X_1 = 0)$	α_{111}
$P(X_1 = 1)$	α_{112}

$\alpha_{111} + \alpha_{112} = s$



$P(X_2 = 0)$	α_{211}
$P(X_2 = 1)$	α_{212}

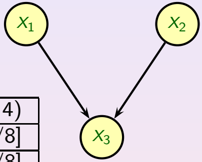
$\alpha_{211} + \alpha_{212} = s$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	α_{311}	α_{321}	α_{331}	α_{341}
$P(X_3 = 1)$	α_{312}	α_{322}	α_{332}	α_{342}

$\alpha_{3j1} + \alpha_{3j2} = s$

Proposed in Zaffalon (1999). We can estimate the intervals and then make a computation in a credal network. The results were too imprecise.

Learning Parameters: Applying the IDM for each conditional probability distribution



Freq.	(1,6)
$P(X_2 = 0)$	[1/8,2/8]
$P(X_2 = 1)$	[6/8,7/8]

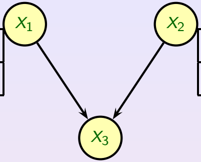
Freq.	(3,4)
$P(X_1 = 0)$	[3/8,4/8]
$P(X_1 = 1)$	[4/8,3/8]

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
Freq.	(1,0)	(1,1)	(0,0)	(1,3)
$P(X_3 = 0)$	[1/2,1]	[1/3,2/3]	[0,1]	[1/5,2/5]
$P(X_3 = 1)$	[0,1/2]	[1/3,2/3]	[0,1]	[3/5,2/5]

Intervals are wider if the number of parents increase.

The Global Approach: IDM for the joint distribution

$P(X_1 = 0)$	α_{111}
$P(X_1 = 1)$	α_{112}



$P(X_2 = 0)$	α_{211}
$P(X_2 = 1)$	α_{212}

$\alpha_{111} + \alpha_{112} = s$

$\alpha_{211} + \alpha_{212} = s$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	α_{311}	α_{321}	α_{331}	α_{341}
$P(X_3 = 1)$	α_{312}	α_{322}	α_{332}	α_{342}

- With the additional linear restrictions:

$$\sum_j \alpha_{31j} + \sum_j \alpha_{32j} = \alpha_{111}, \; \sum_j \alpha_{33j} + \sum_j \alpha_{34j} = \alpha_{112}$$

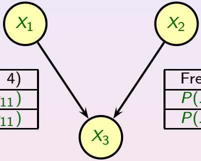
$$\sum_j \alpha_{31j} + \sum_j \alpha_{33j} = \alpha_{211}, \; \sum_j \alpha_{32j} + \sum_j \alpha_{34j} = \alpha_{212}$$

The Global IDM

- It is more restrictive: smaller intervals.
- It was proposed in Zaffalon (2002).
- It is more difficult from a computational point of view: you can not compute the intervals and forget the alphas. You have to optimize in the alphas.
- Locally it behaves as the local IDM: it is not necessary to divide the global sample size among the number of conditional distributions. The possible intervals for the local conditional distributions are the same than in the local model. The only difference is that there are restrictions between the probabilities of the different conditional distributions.

Global IDM

It applies the IDM assuming it for the set of parameters of the joint distribution. Given s, we have to compute all the probability distributions:



Freq.	(3, 4)
$P(X_1 = 0)$	$(3 + \alpha_{111})/(7 + s_{11})$
$P(X_1 = 1)$	$(4 + \alpha_{121})/(7 + s_{11})$

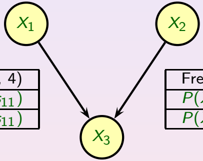
Freq.	(1, 6)
$P(X_2 = 0)$	$(1 + \alpha_{211})/(7 + s_{21})$
$P(X_2 = 1)$	$(6 + \alpha_{221})/(7 + s_{21})$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
Freq.	(1,0)	(1,1)	(0,0)	(1,3)
$P(X_3 = 0)$	$(1 + \alpha_{311})/(1 + s_{31})$	$(1 + \alpha_{321})/(2 + s_{32})$	$(1 + \alpha_{331})/(2 + s_{33})$	$(1 + \alpha_{341})/(2 + s_{34})$
$P(X_3 = 1)$	$(0 + \alpha_{311})/(1 + s_{31})$	$(1 + \alpha_{321})/(2 + s_{32})$	$(1 + \alpha_{331})/(2 + s_{33})$	$(1 + \alpha_{341})/(2 + s_{34})$

Constraints: $\sum_k \alpha_{ijk} = s_{ij}, \sum_j s_{ij} = s, \quad s_{31} + s_{32} = \alpha_{111}, \quad s_{33} + s_{34} = \alpha_{112}, \ldots$

Global IDM

It applies the IDM assuming it for the set of parameters of the joint distribution. Given s, we have to compute all the probability distributions:



Freq.	(3, 4)
$P(X_1 = 0)$	$(3+\alpha_{111})/(7+s_{11})$
$P(X_1 = 1)$	$(4+\alpha_{121})/(7+s_{11})$

Freq.	(1, 6)
$P(X_2 = 0)$	$(1+\alpha_{211})/(7+s_{21})$
$P(X_2 = 1)$	$(6+\alpha_{221})/(7+s_{21})$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
Freq.	(1,0)	(1,1)	(0,0)	(1,3)
$P(X_3 = 0)$	$(1+\alpha_{311})/(1+s_{31})$	$(1+\alpha_{321})/(2+s_{32})$	$(1+\alpha_{331})/(2+s_{33})$	$(1+\alpha_{341})/(2+s_{34})$
$P(X_3 = 1)$	$(0+\alpha_{311})/(1+s_{31})$	$(1+\alpha_{321})/(2+s_{32})$	$(1+\alpha_{331})/(2+s_{33})$	$(1+\alpha_{341})/(2+s_{34})$

Constraints: $\sum_k \alpha_{ijk} = s_{ij}, \sum_j s_{ij} = s, \quad s_{31} + s_{32} = \alpha_{111}, \quad s_{33} + s_{34} = \alpha_{112}, \ldots$

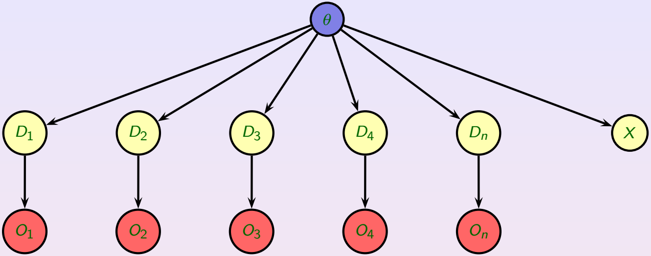
If we are interested only in one conditional: $P(X_3 = 0|0,0) \in [1/2, 1]$

The IDM: drawbacks

Piatti, Zaffalon, 2007: No learning from indirect observations

If $\mathbf{O}$ is a set of observations defining a strictly positive and continuous likelihood function on θ, $l(.|\mathbf{O})$, then for any prior model in θ defined by a credal set $\mathcal{P}$ for which the interval $[\underline{P}(x_i), \overline{P}(x_i)] = [0, 1]$ before the observations, we have that after the observations $[\underline{P}(x_i|\mathbf{o}), \overline{P}(x_i|\mathbf{o})] = [0, 1]$

No Learning



- If we do not observe $D_i = d_i$ but we have an indirect observation system:

$$P(O_i = o_i | D_i = d_i) = \begin{cases} 1-\epsilon, & \text{if } o_i = d_i \\ \epsilon/(k-1) & \text{otherwise} \end{cases}$$

- Even if we observe $\mathbf{o} = (o_1, \ldots, o_{1000})$ with $o_i = x_1$, $\forall i$, $[\underline{P}(X = x_i|\mathbf{o}), \overline{P}(X = x_i|\mathbf{o})] = [0, 1]$

The Learning Principle

Statement

An imprecise prior information $\mathcal{P}$ about a parameter set Θ satisfies the learning principle if and only if for any measurable set $A \subseteq \Theta$ with $|A| > 0$ and any sequence of likelihood functions $\{l_n\}$ such that

$$\frac{\inf\{l_n(\theta) : \theta \in A\}}{\sup\{l_n(\theta) : \theta \in \Theta \setminus A\}} \longrightarrow +\infty$$

then $\underline{P}(A|l_n) \longrightarrow 1$.

The Learning Principle

Statement

An imprecise prior information $\mathcal{P}$ about a parameter set Θ satisfies the learning principle if and only if for any measurable set $A \subseteq \Theta$ with $|A| > 0$ and any sequence of likelihood functions $\{l_n\}$ such that

$$\frac{\inf\{l_n(\theta) \,:\, \theta \in A\}}{\sup\{l_n(\theta) \,:\, \theta \in \Theta \setminus A\}} \longrightarrow +\infty$$

then $\underline{P}(A|l_n) \longrightarrow 1$.

Equivalence

An imprecise prior information $\mathcal{P}$ about a parameter set Θ satisfies the learning principle if and only if $A \subseteq \Theta$ with $|A| > 0$: $\underline{P}(A) > 0$ (under coherence conditioning).

In the binary case (IDM),$\underline{P}([a, b]) = 0$, except for the trivial interval $[0, 1]$.

The Bounded IDM

The IDM does not satisfies the learning principle. In fact $\underline{P}([a, b]) = 0$, except for the trivial interval $[0, 1]$.
The bounded IDM assumes the set of prior probabilities:

$$\mathcal{P} = \{D(\alpha_1, \ldots, \alpha_k) : \sum_{i=1}^{k} \alpha_i = s, \alpha_i > t\}$$

It satisfies the learning principle but fails to verify RIP.

Result

There is no coherent prior model $\mathcal{P}$ such that verifies RIP, symmetry and learning principles.
The learning principle implies that without observations $\underline{P}(X = a_i) \in [a, b]$ with $a > 0, b < 1$.

The Bounded IDM

The IDM does not satisfies the learning principle. In fact $\underline{P}([a,b]) = 0$, except for the trivial interval $[0,1]$.
The bounded IDM assumes the set of prior probabilities:

$$\mathcal{P} = \{D(\alpha_1,\ldots,\alpha_k) : \sum_{i=1}^{k} \alpha_i = s, \alpha_i > t\}$$

It satisfies the learning principle but fails to verify RIP.

Result

There is no coherent prior model $\mathcal{P}$ such that verifies RIP, symmetry and learning principles.
The learning principle implies that without observations $\underline{P}(X = a_i) \in [a,b]$ with $a > 0, b < 1$.

What is more important RIP or learning?
In Moral (2012) I give reasons in favor of learning.

Reasons for Learning against RIP

- The specification of the problem is relevant information.
- The number of values of a variable can also be learned: some are better than others (see for example discretization).
- We want to be vacuous in the predictions of next outcome under no observations, but we are not being vacuous about the parameter space.
- Betting interpretation: without observations, there is not an amount of money z such that we are ready to pay 1 to get z if $X = x$.

The Imprecise Sample Size Dirichlet Model (ISSDM)

$$\mathcal{P} = \{D(s/k, \ldots, s/k) \,:\, s_1 \leq s \leq s_2\}$$

Identical weights for all the cases, but imprecise equivalent sample size.

- Introduced by Walley (1990) in his book as an example (but without real practical interest)
- Imprecision orthogonal to the one in the IDM
- Sudied in Masegosa, Moral (2014)

ISSDM: Prior Information

$$P(X = x_i|D) \in [\underline{P}(\theta_i|D), \overline{P}(\theta_i|D)] = \begin{cases} \left[\frac{N_i+s_1/k}{N+s_1}, \frac{N_i+s_2/k}{N+s_2}\right] & \text{if } N_i/N < 1/k \\ \left[\frac{N_i+s_2/k}{N+s_2}, \frac{N_i+s_1/k}{N+s_1}\right] & \text{otherwise} \end{cases}$$

Comparison IDM - ISSDM

- Under no information $N = 0$, it produces precise (uniform) estimations of the probability: $P(X = a_i) = 1/k$
- Imprecision appears as deviations of the uniform distribution in relative frequencies.

Example

Comparison of results, ISSDM $(s_1 = 1, s_2 = 2)$ IDM $(s = 1)$ binary variable and approximate intervals

Imagine that we observe 20% of cases for a_1 against 80% for a_2.

Interval probabilities:

N	5	10	100	1000
IDM a_1	[0.17,0.33]	[0.18,0.27]	[0.198,0.208]	[0.200,0.201]
IDSSM a_1	[0.25,0.28]	[0.23, 0.25]	[0.203,0.208]	[0.200,0.201]

If we observe 50% of the cases a_1

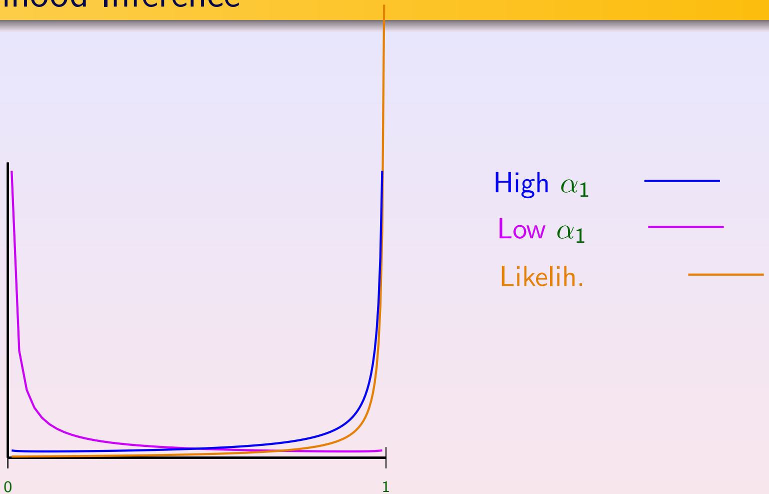
N	5	10	100	1000
IDM a_1	[0.42,0.58]	[0.45,0.56]	[0.495,0.505]	[0.499,0.501]
IDSSM a_1	0.5	0.5	0.5	0.5

The problem of IDM (bivariate case):

- In the indirected observation problem, if we observe $\mathbf{o} = (o_1, \ldots, o_{1000})$ with $o_i = x_1$, $\forall i$, $[\underline{P}(X = x_i|\mathbf{o}), \overline{P}(X = x_i|\mathbf{o})] = [0, 1]$
- The lower limit comes from prior densities $D(\alpha_1, \alpha_2)$ with very low α_1.
- The likelihood of parameters θ given the observations is:

$$L(\theta) = \prod_{i=1}^{1000} ((1-\epsilon)\theta + \epsilon(1-\theta))$$

- This likelihood is concentrated in high values of θ.
- With low α_1 the density is concentrated in low values of θ.
- Given a low value of α_1 the probability of the data is very small.



There is a likelihood associated to each density, which is small for small α_1 and large for large α_1. The 'non-learning' problem comes from not taking it into account.

Likelihood Based Inference

- Assume a multinomial variable X, and a set of densities on $(\theta_1, \ldots, \theta_k)$,

$$\mathcal{P} = \{f_r | r \in R\}$$

Each set of data D defines a likelihood in the set of densities:

$$L(r|D) = P(D|f_r) = \int_\theta f_r(\theta) P(D|\theta) d\theta$$

- How to use this density?

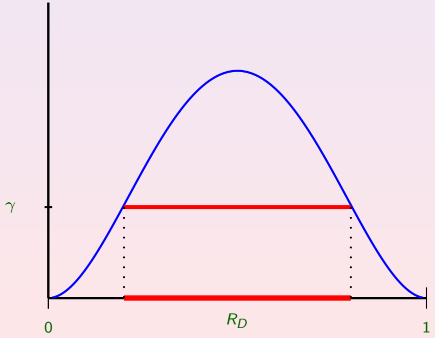
Alternatives

- Second order model: likelihood on R and probabilities given r (Cattaneo, 2012).
- To consider the likelihood as defining a possibility measure, Cano, Moral Verdegay-López (1991) Moral (1992) (probabilities defined on finite sets).
- Define upper and lower probabilities on R. P. Walley and S. Moral (1999) "Upper Probabilities Based Only in the Likelihood Function." (finite sets)
- To use pure likelihood based intervals (α-cut updating rule, Cattaneo, 2014) A thereshold γ is selected and after some data D we compute: $r_{max} = \arg\max_r L(r|D)$ and the conditional information is given by

$$\mathcal{P}_D = \{f_r(.|D) | r \in R, L(r|D) \geq \gamma L(r_{max}|D)\}$$

Cattaneo (2014)

- It is the only continuous updating rule.



Possible Prior Densities

- The IDM

$$\mathcal{P} = \{D(\alpha_1, \dots, \alpha_k) \,:\, \sum_{i=1}^{k} \alpha_i = s,\, \alpha_i > 0\}$$

- The ISSDM

$$\mathcal{P} = \{D(s/k, \dots, s/k) \,:\, s_1 \leq s \leq s_2\}$$

- The degenerated model

$$\mathcal{P} = \{f_\theta \,:\, \theta \in \Theta\}$$

  where f_θ is the density degenerated in θ.
  Cattaneo (2014) and Antonucci, Cattaneo, Corani (2012) consider the last case

In all of them, the learning principle is satisfied (with the α-cut conditioning rule).

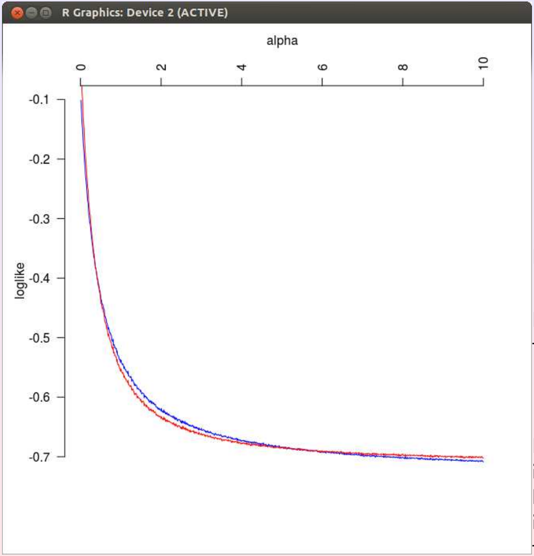
Experimental Comparison

- We have carried out some experiments in which we aim to compare the likelihood based results with Bayesian procedures.
- We select a parameter ($\theta \in \Theta$) (according a Dirichlet $D(\alpha_r, \alpha_r)$).
- 10000 samples of size 10 are obtained.
- For each one of them we make an estimation of the probability Q.
- The goodness of the approximation is measured with $E_\theta[\log(Q)]$.

Experimental Comparison (II)

- We have used the degenerated model.
- With this, we obtain a set of possible values for the parameters $\Theta_D \subseteq \Theta$.
- How to compare an imprecise procedure with an imprecise one? We select one of the parameters from Θ_D and compare this parameter with the Bayesian procedure.
- To select only one value Q from Θ_D we select the one giving rise to a probability with maximum entropy (it can be justified as a max-min decision rule with $\log(Q)$ as utility.

Results

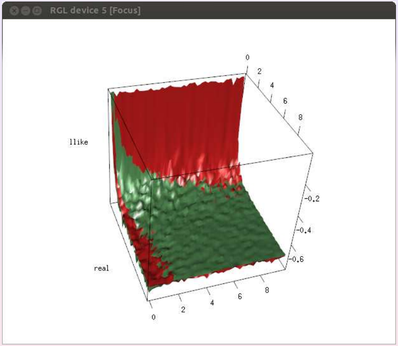


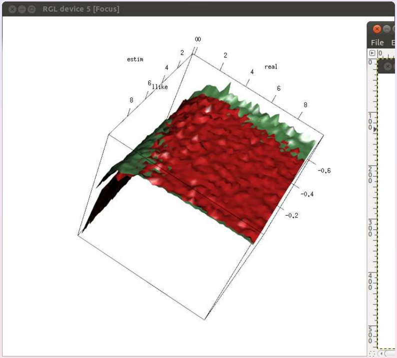
Blue is Laplace

Red is imprecise + maximum entropy

If the way of generating cases is similar to the hypothesis done by Laplace ($\alpha = 1$), then blue is better, but if we are far from this, red is better

Results


RGL device 5 [Focus]
llike
real
0
2
4
6
8
-0.2
-0.4
-0.6


RGL device 5 [Focus]
estim
real
llike
2
4
6
8
-0.6
-0.4
-0.2
File

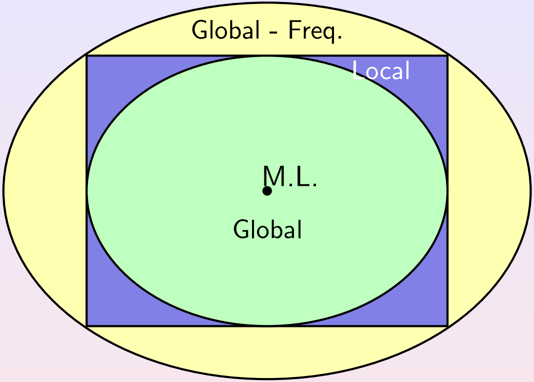
Applying Likelihood Inference to Learn Credal Networks

- We have to apply it to each one of the conditional distributions.
- We can use global and local models.
- Let R the global set of parameters and R_{ij} the set of parameters for the conditional distribution of X_i given that $Pa_i = \pi_j$.
- If we have a local model: R is given through R_{ij}.
- If data are complete, we have separation, in the following sense

$$L(R|D) = \prod_{ij} L(R_{ij}|D)$$

- When this happens there are 3 possible approaches for α-cut conditioning:
  1. Global: to apply α-cut conditioning to the global likelihood distribution
  2. Local: to apply α-cut conditioning to each one of the parameters, considering that the set of possible parameters is the Cartesian product.
  3. To apply α-cut conditioning to the global likelihood distribution, but descreasing the thereshold γ to γ^l where l is the number of parameters (Pawitan, "In all likelihood") (Akaike Information criterion calibration).

With 2 parameters:



The global one is the most informative. I believe that it is the one that could provide sensible results without producing too wide intervals as result of inference. Resulting credal networks are not separately specified.

Learning: Structure

Score + Search procedures

Search for the graph maximizing a metric or score measuring how good is a graph for the data.

Bayesian Score

$$P(G|D) = \frac{P(D|G).P(G)}{P(D)}$$

Under certain conditions (uniform prior on the graphs, prior Dirichlet densities about the parameters, independence on the parameters) there is a closed expression to compute the score from the data D.

$$P(G|D) \propto \prod_{i=1}^{m} \prod_{j=1}^{q_i} \frac{\Gamma(k_i \alpha_{ij})}{\Gamma(N_{ij} + k_i \alpha_{ij})} \prod_{k=1}^{k_i} \frac{\Gamma(\alpha_{ij} + N_{ijk})}{\Gamma(\alpha_{ij})}$$

Bayesian Score: Local vs Global Application

Local Score. K2 Score: $D(1, \ldots, 1)$

$$P(G|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(k_i)}{\Gamma(N_{ij} + k_i)} \prod_{k=1}^{k_i} \frac{\Gamma(1 + N_{ijk})}{\Gamma(1)}$$

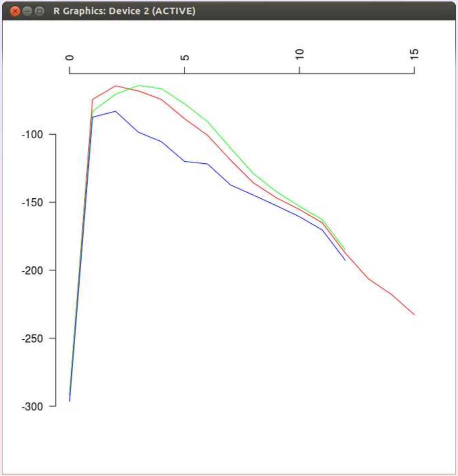
Global Equivalent Sample Size: $D(s/(q_i.k_i), \ldots, s/(q_i.k_i))$

$$P(G|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(s/q_i)}{\Gamma(N_{ij} + s/q_i)} \prod_{k=1}^{k_i} \frac{\Gamma(s/(q_i.k_i) + N_{ijk})}{\Gamma(s/(q_i.k_i))}$$

The global model gives rise to the BDEu the most used criterion for learning Bayesian networks.

Some Properties

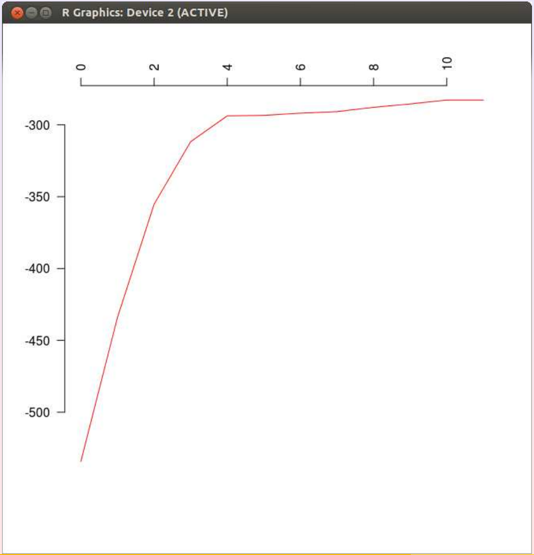
Evolution of the score as a function of the number of parents:



Evolution of the score as a function of the number of parents and different s values: $s = 0.01$ (blue), $s = 2$ (red), $s = 20$ (green)

Bad Behaviour

If we have deterministic distributions, we can have examples of wrong behaviour:



Imprecise Methods to Learn the Structure (old approach)

Generalized Credal Networks

A set of graphs with imprecise probabilities each one of them.

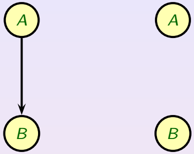
- A precise Bayesian procedure is based on a function F that assigns to each graph G a family of prior distributions for each variable given its parents: $F(G)_{ij}$ is the prior probability of X_i conditioned to the j-th value of its parents. Example $F_s(G)_{ij} = D(s/(q_i.k_i), \ldots, s/(q_i.k_i))$.
- An imprecise procedure can be based on a family $\mathcal{F}$ of assignment functions. Example $\mathcal{F} = \{F_s \mid s \in \{s_1, \ldots, s_n\}\}$
- The set of learned graphs would be the set of graphs that are optimal for the different functions $F \in \mathcal{F}$ (E-admissibility).

IDM Applied to Learn Structure

- We have as family of assignments $\mathcal{F}$, where each $F \in \mathcal{F}$ is determined for each one of the prior global Dirichlet distributions of the IDM: $F(G)_{ij} = D(\alpha_{ij1}, \ldots, \alpha_{ijk_i})$ a set of linear restrictions as $\sum_{jk} \alpha_{ijk} = s$ and other linear restrictions.
- It is difficult from a computational point of view.
- It is not useful

IDM Applied to Learn Structure

Only applied to a very simple case:



Result

if all the cases have non null frequency: independence can not be dominated by dependence.

The same reason for which the IDM does not satisfy the learning principle has as consequence that IDM is not good for deciding about dependence-independence.

Compromise

Assume a minimum value α_{ijk} (bounded IDM). Abellán and Moral (2005) obtain good results to estimate the joint probability.

ISSDM Applied to Learning the Structure

$\mathcal{F} = \{F_s \,|\, s \in \{s_1, \ldots, s_n\}\}$

Experimental Fact

Moral (2004,2005) T. Silander, P. Kontkanen, P. Myllymäki (2007): the parameter s determines how dense is the learned network: Small s values produce networks with a low number of arrows and large values of s networks with more links

An Imprecise Search Approach with ISSDM

- Build a network with the small bound s_1 with a search procedure and build the mimimal graph G_{s_1}
- Start using the score with $s = s_2$ and apply a search taking G_{s_1} as basis (none of the links of this graph can be removed, and a link of G_{s_1} can be inverted if its inversion in G_{s_1} produces an equivalent graph). Then we build G_{s_2} as a supergraph of G_{s_1}
- The links in G_{s_1} are necessary and the links in $G_{s_2} \setminus G_{s_1}$ are possible

Experiments: learning structure

- Alarm, Boblo, Boelarge, Hailfinder, Insurance
- The imprecise approach

Missing Edges s_2	S-[1,4.0]	S-[0.5,8]	S-[0.25,16]	S-2
100	16.7	15.08	13.26	18.18
500	9.94	9.24	8.48	11
1000	8.32	7.56	7.14	8.8
5000	5.26	5.02	4.64	5.44

Extra Edges s_1	S-[1,4.0]	S-[0.5,8]	S-[0.25,16]	S-2
100	12.44	11.16	10.7	16.08
500	6.02	4.94	4.92	8.12
1000	4.72	3.76	3.26	6.2
5000	1.68	1.28	1.2	2.84

Experiments: learning structure

Number of sure errors	S-[1,4.0]	S-[0.5,8]	S-[0.25,16]	S-2
100	29.14	26.24	23.96	34.26
500	15.96	14.18	13.4	19.12
1000	13.04	11.32	10.4	15
5000	6.94	6.3	5.84	8.28

Num. Links IMPRECISE	S-[1,4.0]	S-[0.5,8]	S-[0.25,16]
100	12.92	24.82	38.34
500	8.04	17.12	28.34
1000	5.52	14.06	24.18
5000	4.02	8.52	16.52

The Likelihood Approach

The Model

A set of graphs $\mathcal{G}$ (usually the set of all the directed acyclic graphs) and for each graph G a familiy of prior densities $\mathcal{P}(G)$ for the distributions of each variable given its parents.
Each element $f \in \mathcal{P}(G)$ assigns a prior distribution f_{ij} for the parameters θ_{ij} of each conditional variable X_i and the j-th combination of values of its parents $Pa_i = \pi_j$.

The Parameter Space

$$\mathcal{M} = \{(G, f) \,|\, G \in \mathcal{G},\, f \in \mathcal{P}(G)\}$$

Local Models

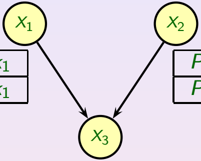
A model is local when $\mathcal{P}(G)$ is determined by families of densities $\mathcal{P}(G)_{ij}$ of each variable given its parents.

Examples

- A global approach: a set of equivalent sample sizes is selected $S = \{s_1, \ldots, s_l\}$ and $\mathcal{P}(G)$ is given by the families f^s such that f^s_{ij} is $D(s/(k_i q_i), \ldots, s/(k_i q_i))$ for $s \in S$.
- A local approach: a set of weights is selected $A = \{\alpha_1, \ldots, \alpha_l\}$ and $\mathcal{P}(G)_{ij}$ is given by the densities $D(\alpha, \ldots, \alpha)$, $\alpha \in A$.
- The local degenerated approach: $\mathcal{P}(G)_{ij}$ is the set of degenerated densities $f_{\theta_{ij}}$, $\theta_{ij} \in \Theta_{ij}$.
  In the local degenerated model the set of parameters is the set of all the Bayesian networks

Examples: a global model based on ISSDM

$S = \{s_1, \ldots, s_l\}$. Dirichlet distributions



$(X_1 = 0)$	$\alpha_{111} = s/k_1$
$(X_1 = 1)$	$\alpha_{112} = s/k_1$

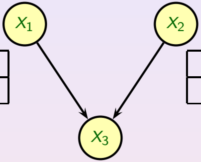
$P(X_2 = 0)$	$\alpha_{211} = s/k_2$
$P(X_2 = 1)$	$\alpha_{212} = s/k_2$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	$\alpha_{311} = s/(k_3 q_3)$	$\alpha_{321} = s/(k_3 q_3)$	$\alpha_{331} = s/(k_3 q_3)$	$\alpha_{341} = s/(k_3 q_3)$
$P(X_3 = 1)$	$\alpha_{312} = s/(k_3 q_3)$	$\alpha_{322} = s/(k_3 q_3)$	$\alpha_{332} = s/(k_3 q_3)$	$\alpha_{342} = s/(k_3 q_3)$

Particular case: $S = \{s\}$ (uncertain about the graphs).

Examples: a local model based on ISSDM

$A = \{\alpha_1, \dots, \alpha_l\}$. Dirichlet distributions



$P(X_1 = 0)$	$\alpha_{11} \in A$
$P(X_1 = 1)$	$\alpha_{11} \in A$

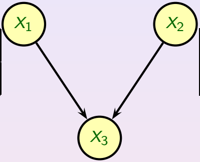
$P(X_2 = 0)$	$\alpha_{21} \in A$
$P(X_2 = 1)$	$\alpha_{21} \in A2$

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	$\alpha_{31} \in A$	$\alpha_{32} \in A$	$\alpha_{33} \in A$	$\alpha_{34} \in A$
$P(X_3 = 1)$	$\alpha_{31} \in A$	$\alpha_{32} \in A$	$\alpha_{33} \in A$	$\alpha_{34} \in A$

Examples: the local degenerated model

The set of graphs and parametrizations (G, Θ).

$P(X_1 = 0)$	θ_{111}
$P(X_1 = 1)$	θ_{112}



$P(X_2 = 0)$	θ_{211}
$P(X_2 = 1)$	θ_{211}

	$X_1 = 0$	$X_1 = 0$	$X_1 = 1$	$X_1 = 1$
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$P(X_3 = 0)$	θ_{311}	θ_{321}	θ_{331}	θ_{341}
$P(X_3 = 1)$	θ_{312}	θ_{322}	θ_{332}	θ_{342}

$\theta_{ijk} \in [0, 1]$

Learning from Data

A thereshold is selected: $\gamma \in [0, 1]$. Given $\mathcal{G}$ and the families $\mathcal{P}(G)$, the set of models is the set of graphs G_o and f_o such that

$$P(D|G_o, f_o) \geq \gamma \max_{G,f} P(D|G, f)$$

Learning the Structure

The set of possible networks is given by $\mathcal{G}_D$ of all the graphs G_o for which there is a f_o such that

$$P(D|G_o, f_o) \geq \gamma \max_{G,f} P(D|G, f)$$

Learning the Structure and parameters

It is the set of graphs $G_o \in \mathcal{G}_D$ with probabilities estimated with densities f_o such that:

$$P(D|G_o, f_o) \geq \gamma \max_{G,f} P(D|G, f)$$

The set of pairs (G_o, f_o) will be denoted as $\mathcal{M}_D$. Each pair (G_o, f_o) defines a Bayesian network with an associated joint probability P_{G_0,f_o}.

The Degenerated Model

- We consider all the Bayesian networks (G_0, Θ_0), such that

$$P(D|G_o, \Theta_o) \geq \gamma \max_{G,\Theta} P(D|G, \Theta)$$

- To compute $\max_{G,\Theta} P(D|G, \Theta)$ is quite simple (no missing data). If $D = \{d_1, \ldots, d_N\}$ is the set of vectors of observations
  $\max_{G,\Theta} P(D|G, \Theta) = \sum_{d_j} N_j \log(N_j/N)$
  where the sum is in the different vectors $d_j \in D$ and N_j is the frequency of d_j appears in D.
  If all the vectors are different $\max_{G,\Theta} P(D|G, \Theta) = -N \log N$.
- The main problem is that there are a high number of models in $\mathcal{M}_D$. In fact if $G \in \mathcal{G}_D$, then for any supergraph G' of G, we will have $G' \in \mathcal{G}_D$.

The Degenerated Model

- If we have to determine a single Bayesian network in $\mathcal{G}$, we should select a single model (G, f) with maximum entropy: this produces simple networks, as simpler networks represent more independence relationships and the entropy increases whith independence.
- If we have to select a subset of models $\mathcal{M}_D^*$ which is a good representation of $\mathcal{M}_D$, there is not a direct procedure as in the case of a single one.
- The problem could be formalized as determining a subset $\mathcal{M}_D^* \subseteq \mathcal{M}_D$, such that for any $(G, f) \in \mathcal{M}_D$ there is a $(G', f') \in \mathcal{M}_D$ such that $DKL(P_{G,f}, P_{G',f'}) \leq \epsilon$
- Simpler models should be preferred for $\mathcal{M}_D^*$ (no details of mathematical formulation)

The ISSDM global model

We could apply the ISSDM in this framework with a finite set of parameters $S = \{s_1, \ldots, s_k\}$.

Differences with previous formulation

- In previous cases, we considered all the graphs G_o such that there is $s' \in S$ such that

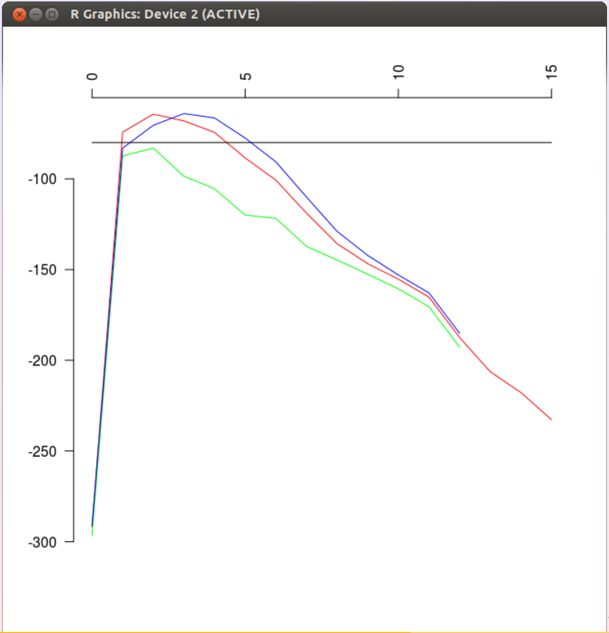
$$G_o = \arg\max_G P(D|G, f_{s'})$$

  where f_s is the model obtained by assigning prior distributions according to the BDEu model with s'.

- Now, we consider all the graphs G_o such that there is $s' \in S$ such that

$$P(D|G_o, f_{s'}) \geq \gamma \max_{G,s} P(D|G, f_s)$$

The ISSDM global model



Conclusions

- We have studied imprecise probability in learning credal networks
- Even if we have a single sample size s, Koller, Friedman (2000): "Model selection makes a somewhat arbitrary choice between models that explain the data reasonably well".
- Likelihood approaches and likelihood intervals are a promising approach to learn credal networks
- Discarding models with low likelihood solves some problems associated to the use of imprecise probabilities (too uniformative).
- Computational aspects should be studied in particular with the degenerated model.
- The problem of approximating a set of probabilities by a subset keeping the most relevant information is interesting and deserves more study.