OSE SEMINAR 2012

A Bayesian score for LDAGs

Johan Pensar

CENTER OF EXCELLENCE IN
OPTIMIZATION AND SYSTEMS ENGINEERING
ÅBO AKADEMI UNIVERSITY

ÅBO, NOVEMBER 29 2012

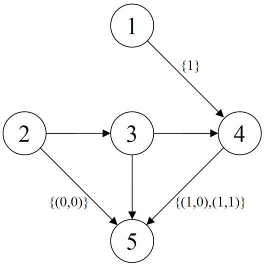
- Joint work with Henrik Nyman, Timo Koski and Jukka Corander.

- Structure of the presentation
  - Introduction
  - Deriving the score function
  - Example

Graphical model (GM)

- A GM is a probabilistic model for which a graph structure represents the dependence structure between a set of random variables.
- The nodes in the graph represent the variables and the edges represent direct dependencies among the variables.
- The absence of an edge represents statements of conditional independence (CI).
- In this talk we will only consider discrete variables.

Labeled Directed Acyclic Graph (LDAG)



- A directed acyclic graph for which certain labels have been added to edges.
- In an LDAG-based GM, the labels represent statements of context-specific independence (CSI).
- Consider the label on edge (4,5):

$$\mathcal{L}_{(4,5)} = \{(1,0),(1,1)\} \qquad \Rightarrow X_5 \perp X_4 \mid (X_2, X_3) \in \{(1,0),(1,1)\}$$
$$\Leftrightarrow X_5 \perp X_4 \mid X_2 = 1, X_3$$

Factorization of the joint distribution according to an LDAG

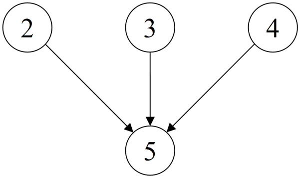
"Fundamental to the idea of a graphical model is the notion of modularity – a complex system is built by combining simpler parts."

- In a GM, the joint distribution is factorized by the graph into lower order distributions.
- Factorization according to an LDAG over $\{X_1, X_2, \ldots, X_d\}$:

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

- The result is a product of conditional probability distributions (CPD).

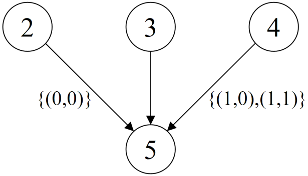
Conditional probability table (CPT)



$X_{Pa(5)}$	$p(X_5 = 1 \mid X_{Pa(5)})$
$(0,0,0)$	p_1
$(0,0,1)$	p_2
$(0,1,0)$	p_3
$(0,1,1)$	p_4
$(1,0,0)$	p_1
$(1,0,1)$	p_1
$(1,1,0)$	p_5
$(1,1,1)$	p_5

- Grows exponentially with the number of parents.
- Fails to capture any regularities among the CPDs.

Reduced conditional probability table



$X_{Pa(5)}$	$p(X_5 = 1 \mid X_{Pa(5)})$
$\{(0,0,0),(1,0,0),(1,0,1)\}$	p_1
$\{(0,0,1)\}$	p_2
$\{(0,1,0)\}$	p_3
$\{(0,1,1)\}$	p_4
$\{(1,1,0),(1,1,1)\}$	p_5

- $\mathcal{X}_{Pa(j)} \xrightarrow{\mathcal{L}_j} \mathcal{S}_{Pa(j)} = \{S_1, S_2, \ldots, S_{k_j}\}$ where $S_l \cap S_{l'} = \emptyset$ (for $l \neq l'$) and $\bigcup_{l=1}^{k_j} S_l = \mathcal{X}_{Pa(j)}$.

Learning of LDAGs

- In the learning process we want to find the optimal LDAG for a set of data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ consisting of n observations $\mathbf{x}_i = (x_{i1}, \dots x_{id})$ of the variables $\{X_1, \dots, X_d\}$ such that $x_{ij} \in \mathcal{X}_j$.
- This problem can be divided into two parts:
  1. To define a score that evaluates the appropriateness of the models.
  2. To develop a search algorithm that searches through parts of the model space in order to find the model with the highest score.

The Bayesian approach

- In the Bayesian approach to model learning, one is interested in the posterior distribution of the models given the data $\mathbf{X}$.
- The posterior probability of an LDAG (G_L) is

$$p(G_L \mid \mathbf{X}) = \frac{p(\mathbf{X}, G_L)}{p(\mathbf{X})} = \frac{p(\mathbf{X} \mid G_L) \cdot p(G_L)}{p(\mathbf{X})}.$$

- The denominator is a normalizing constant that does not depend on G_L and can therefore be ignored when comparing graphs.
- Our goal is thus to maximize

$$p(\mathbf{X}, G_L) = p(\mathbf{X} \mid G_L) \cdot p(G_L).$$

Marginal likelihood $p(\mathbf{X}, G_L) = p(\mathbf{X} \mid G_L) \cdot p(G_L)$

- $p(\mathbf{X} \mid G_L)$ is the marginal probability of observing the data $\mathbf{X}$ given a graph G_L.
- To evaluate $p(\mathbf{X} \mid G_L)$, we need to consider all possible instances of the parameter vector θ according to

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

  where Θ_{G_L} denotes the parameter space induced by the LDAG.
- $p(\mathbf{X} \mid G_L, \theta)$ and $f(\theta \mid G_L)$ are the respective likelihood function and prior distribution over the parameters.

Marginal likelihood $p(\mathbf{X}, G_L) = p(\mathbf{X} \mid G_L) \cdot p(G_L)$

- Under certain assumptions, the marginal likelihood can be calculated analytically

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

where α_{ijl} are hyperparameters and $n(S)$ is the number of times any of the elements in S occur in the data.

Prior over the LDAGs $p(X, G_L) = p(X \mid G_L) \cdot p(G_L)$

- Prior probability of the LDAG.
- Generally not given too much attention in model learning for ordinary DAGs (Uniform prior).
- Essential part of the score when evaluating LDAGs.
- We define our prior by

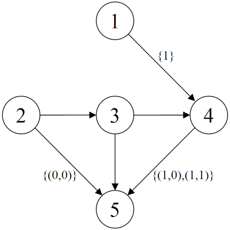
$$p(G_L) = c \cdot \kappa^{|\Theta_G| - |\Theta_{G_L}|} = c \cdot \prod_{j=1}^{d} \kappa^{(|\mathcal{X}_j|-1) \cdot (|\mathcal{X}_{Pa(j)}| - |\mathcal{S}_j|)},$$

where $\kappa \in (0, 1]$ can be considered a measure of how strongly a label is penalized when added to the graph.

Putting the pieces together: $p(\mathbf{X}, G_L) = p(\mathbf{X} \mid G_L) \cdot p(G_L)$

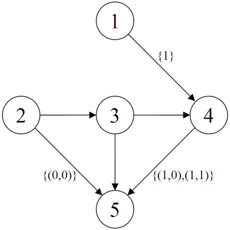
$$p(\mathbf{X}, G_L) = c \cdot \prod_{j=1}^{d} \kappa^{(|\mathcal{X}_j|-1)\cdot(|\mathcal{X}_{Pa(j)}|-|\mathcal{S}_j|)} \prod_{l=1}^{k_j} \frac{\Gamma\left(\sum_{i=1}^{r_j} \alpha_{ijl}\right)}{\Gamma\left(n(S_{jl}) + \sum_{i=1}^{r_j} \alpha_{ijl}\right)} \prod_{i=1}^{r_j} \frac{\Gamma(n(x_{ji} \times S_{jl}) + \alpha_{ijl})}{\Gamma(\alpha_{ijl})}$$

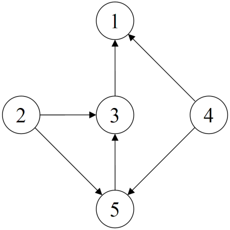
Example (n=500)



$\kappa = 0.001$ $\kappa = 0.25$ $\kappa = 0.5$ $\kappa = 1$

Example (n=500)





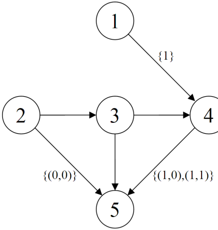
$\kappa = 0.001$
$|\Theta_{G_L}| = 14$ (14)

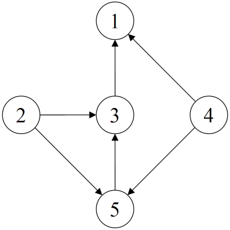
$\kappa = 0.25$

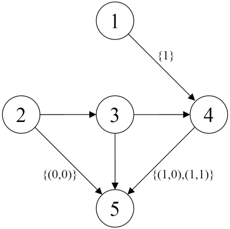
$\kappa = 0.5$

$\kappa = 1$

Example (n=500)

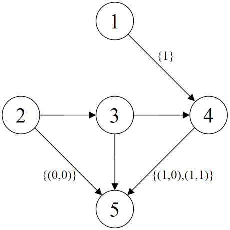


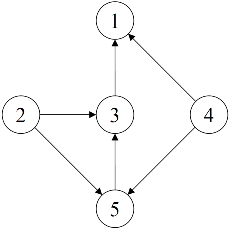




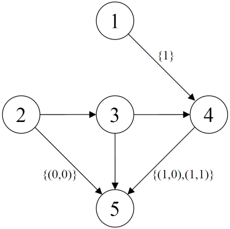
$\kappa = 0.001$	$\kappa = 0.25$	$\kappa = 0.5$	$\kappa = 1$
$\lvert\Theta_{G_L}\rvert = 14\,(14)$	$\lvert\Theta_{G_L}\rvert = 12\,(16)$		

Example (n=500)

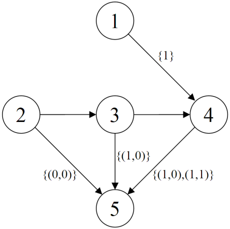




$\kappa = 0.001$
$|\Theta_{G_L}| = 14\ (14)$



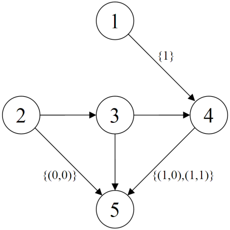
$\kappa = 0.25$
$|\Theta_{G_L}| = 12\ (16)$

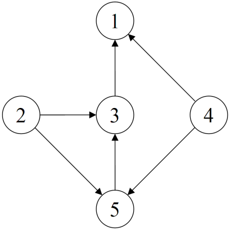


$\kappa = 0.5$
$|\Theta_{G_L}| = 11\ (16)$

$\kappa = 1$

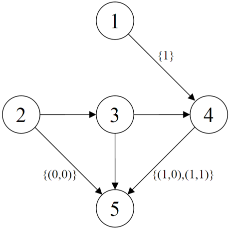
Example (n=500)





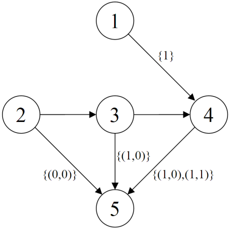
$\kappa = 0.001$

$|\Theta_{G_L}| = 14\ (14)$



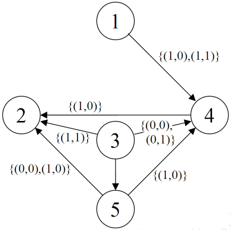
$\kappa = 0.25$

$|\Theta_{G_L}| = 12\ (16)$



$\kappa = 0.5$

$|\Theta_{G_L}| = 11\ (16)$



$\kappa = 1$

$|\Theta_{G_L}| = 11\ (20)$

Some references

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller.
Context-specific independence in bayesian networks.
In E. Horvitz and F.V. Jensen, editors, *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pages 115–123. Morgan Kaufmann, 1996.

J. Corander.
Labelled graphical models.
Scandinavian Journal of Statistics, 30:493–508, 2003.

N. Friedman and M. Goldszmidt.
Learning bayesian networks with local structure.
In E. Horvitz and F.V. Jensen, editors, *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, pages 252–262. Morgan Kaufmann, 1996.

J. Pensar, H. Nyman, T. Koski, and J. Corander.
Labeled directed acyclic graphs.
Submitted, 2012.

Thank you for listening!

Questions?