OSE SEMINAR 2013

Marginal pseudo-likelihood: A Bayesian approach for learning the graph structure of a Markov network

Johan Pensar

CENTER OF EXCELLENCE IN
OPTIMIZATION AND SYSTEMS ENGINEERING
ÅBO AKADEMI UNIVERSITY

ÅBO, NOVEMBER 15 2013

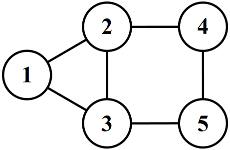
OSE OPTIMIZATION AND SYSTEMS ENGINEERING

Åbo Akademi University

- Joint work with Henrik Nyman and Jukka Corander.

- Structure of the presentation:
  - Introduction
  - Derivation of the score
  - Search algorithm

Markov network (MN)

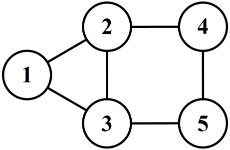


- A MN is a probabilistic graphical model over a set of discrete variables $(X_1, \ldots, X_d)$.
- The dependence structure over the variables is represented by an undirected graph $G = (V, E)$.
- The nodes in the graph, $V = \{1, \ldots, d\}$, represent the variables and the edges, $E \subseteq \{V \times V\}$, represent direct dependencies among the variables.
- Absence of edges represents statements of conditional independence, in particular

$$X_i \perp X_{V \setminus \{MB(i) \cup i\}} \mid X_{MB(i)}$$

where $MB(i) = \{j \in V : \{i, j\} \in E\}$ is the Markov blanket of node i.

Markov network (MN)



- A MN is a pair (G, θ_G) where θ_G is a parameterization of a joint distribution P_G over $(X_1, \ldots, X_d)$
- P_G must satisfy the restrictions imposed by G, in particular:

$$X_i \perp X_{V \setminus \{MB(i) \cup i\}} \mid X_{MB(i)} \Leftrightarrow P(X_i \mid X_{V \setminus i}) = P(X_i \mid X_{MB(i)})$$

- We assume that P_G is positive.
- The joint distribution factorizes according to its maximal cliques

$$P_G(X_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \phi_C(X_C)$$

where $\phi_C : \mathcal{X}_C \to \mathbb{R}_+$ is a clique factor and $Z = \sum_{x_V \in \mathcal{X}_V} P_G(x_V)$ is the partition function.

Structure learning

- We assume we have a data set $\mathbf{X}$ containing n complete i.i.d. joint observations $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,d})$ generated from θ_{G^*}.
- The aim is to discover the graph structure G^* from the set of all possible graph structures $\mathcal{G}$.
- Structure learning is basically model class learning.
- Reasons for structure learning:
  - Step in model learning - Learn distribution given the graph.
  - Knowledge discovery - The structure is a goal in itself.
- Structure learning methods can roughly be divided into two categories:
  - Constraint-based - Independence tests.
  - Score-based - Optimization problem.

The Bayesian approach

- We choose the graph with the highest posterior probability given the data:

$$p(G \mid \mathrm{X}) = \frac{p(\mathrm{X} \mid G) \cdot p(G)}{p(\mathrm{X})}$$

- Since $p(\mathrm{X})$ is a normalizing constant, the problem can be formulated as

$$\arg\max_{G \in \mathcal{G}} p(\mathrm{X} \mid G) \cdot p(G).$$

- The key term of the Bayesian score is the marginal likelihood which is evaluated according to

$$p(\mathrm{X} \mid G) = \int_{\theta \in \Theta_G} p(\mathrm{X} \mid \theta, G) \cdot f(\theta \mid G) d\theta.$$

- The marginal likelihood is hard to evaluate for MNs.

The pseudo-likelihood function

- The pseudo-likelihood (Besag, 1975) is given by

$$\hat{p}(\mathbf{X} \mid \theta) = \prod_{j=1}^{d} p(\mathbf{X_j} \mid \mathbf{X_{V \setminus j}}, \theta).$$

- Given a graph, the local Markov property allows us to simplify the pseudo-likelihood as

$$\hat{p}(\mathbf{X} \mid \theta, G) = \prod_{j=1}^{d} p(\mathbf{X_j} \mid \mathbf{X}_{MB(j)}, \theta, G).$$

- The marginal pseudo-likelihood (MPL) is evaluated according to

$$\hat{p}(\mathbf{X} \mid G) = \int_{\theta \in \Theta_G} \hat{p}(\mathbf{X} \mid \theta, G) \cdot f(\theta \mid G) d\theta.$$

Marginal pseudo-likelihood

- We assume global and local independence among the parameters similarly to the parameter independence assumption made for Bayesian networks (Heckerman et al., 1995).
- This allows us to factorize the parameter prior distribution and solve the MPL analytically:

$$\hat{p}(\mathrm{X} \mid G) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \frac{\Gamma(\alpha_{jl})}{\Gamma(n_{jl} + \alpha_{jl})} \prod_{i=1}^{r_j} \frac{\Gamma(n_{ijl} + \alpha_{ijl})}{\Gamma(\alpha_{ijl})}$$

- The MPL can in fact be considered the marginal likelihood for a bi-directional dependency network (Heckerman et al., 2001).

Number of possible graphs, $|\mathcal{G}|$

d	$\lvert\{V \times V\}\rvert = \binom{d}{2}$	$\lvert\mathcal{G}\rvert = 2^{\binom{d}{2}}$
2	1	2
4	6	64
8	28	268435456
16	120	$1.32\ldots \cdot 10^{36}$
32	496	$2.04\ldots \cdot 10^{149}$
$\vdots$	$\vdots$	$\vdots$

The direct approach

$$\arg\max_{G \in \mathcal{G}} \hat{p}(\mathrm{X} \mid G) \cdot p(G)$$

- We assume uniform prior $p(G) = 1/|\mathcal{G}|$.
- Two graphs G_1 and G_2 are compared by Bayes pseudo-factor

$$K(G_1; G_2) = \frac{\hat{p}(\mathrm{X} \mid G_1)}{\hat{p}(\mathrm{X} \mid G_2)}.$$

- If we assume a single edge difference $\{i, j\}$ between G_1 and G_2, then

$$K(G_1; G_2) = \frac{p(\mathrm{X}_i \mid \mathrm{X}_{MB_1(i)})}{p(\mathrm{X}_i \mid \mathrm{X}_{MB_2(i)})} \cdot \frac{p(\mathrm{X}_j \mid \mathrm{X}_{MB_1(j)})}{p(\mathrm{X}_j \mid \mathrm{X}_{MB_2(j)})}.$$

Reformulation of the direct approach

- By denoting $MB(G) = \{MB(1), \ldots, MB(d)\}$, we reformulate the original problem:

$$\arg\max_{G \in \mathcal{G}} \hat{p}(\mathrm{X} \mid G)$$

$$\Leftrightarrow$$

$$\arg\max_{MB(G) \in \times_{j \in V} \mathcal{P}(V \setminus j)} \prod_{j=1}^{d} p(\mathrm{X}_j \mid \mathrm{X}_{MB(j)})$$

$$\text{subject to} \quad i \in MB(j) \Rightarrow j \in MB(i) \text{ for all } i, j \in V$$

Relaxation of the direct approach

- Relaxed version of the reformulated problem:

$$\underset{MB(G)\in \times_{j\in V}\mathcal{P}(V\setminus j)}{\arg\max} \prod_{j=1}^{d} p(X_j \mid X_{MB(j)})$$

- We now have d independent subproblems:

$$\underset{MB(j)\subseteq V\setminus j}{\arg\max}\, p(X_j \mid X_{MB(j)}) \quad \text{for } j = 1, \ldots, d.$$

- High-dimensional problems - Parallel solving!

Forming a MN structure from inconsistent Markov blankets

- Solutions to the relaxed problem are in general inconsistent in the sense that $i \in MB(j)$ but $j \notin MB(i)$.
- Post-process the solution to satisfy the structure of a MN.
- Simple approaches:

$$E_{AND} = \{\{i,j\} \in \{V \times V\} : i \in MB(j) \textbf{ AND } j \in MB(i)\}$$
$$E_{OR} = \{\{i,j\} \in \{V \times V\} : i \in MB(j) \textbf{ OR } j \in MB(i)\}$$

- A more elaborate approach - Treat the Markov blanket discovery phase as a pre-scan and solve

$$\arg\max_{G \in \mathcal{G}_{OR}} \hat{p}(\mathbf{X} \mid G)$$

where $\mathcal{G}_{OR} = \{G \in \mathcal{G} : E \subseteq E_{OR}\}$.

References

Julian Besag.
Statistical analysis of non-lattice data.
Journal of the Royal Statistical Society. Series D (The Statistician), 24(3):pp. 179–195, 1975.

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

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie.
Dependency networks for inference, collaborative filtering, and data visualization.
J. Mach. Learn. Res., 1:49–75, September 2001.

D. Koller and N. Friedman.
Probabilistic Graphical Models: Principles and Techniques.
MIT Press, 2009.