Fast computation of NML for Bayesian networks

Petri Myllymäki[1]

Complex Systems Computation Group (CoSCo)
Department of Computer Science and Helsinki Institute for Information Technology
University of Helsinki, Finland

July 9, 2008

[1] Joint work with Petri Kontkanen, Tommi Mononen, Jorma Rissanen, Teemu Roos, Tomi Silander, Hannes Wettig

Outline

1. The data
2. Bayesian networks
3. NML for Bayesian networks
4. Application: histogram density estimation
5. Conclusions

The data

- Let

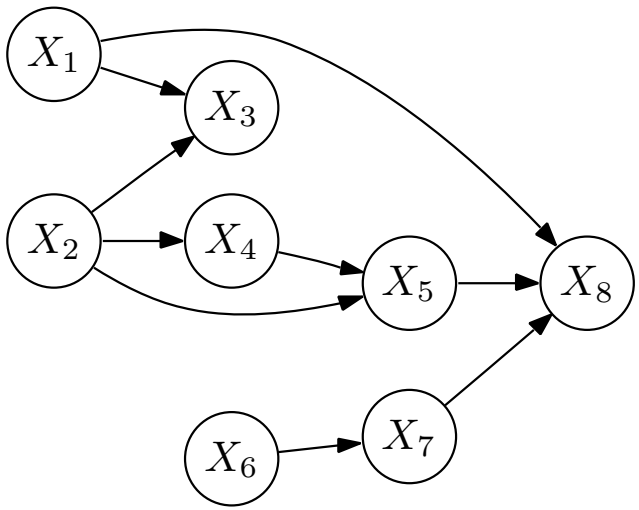
$$x^n := \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1,:} \\ \mathbf{x}_{2,:} \\ \vdots \\ \mathbf{x}_{n,:} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{:,1} \mathbf{x}_{:,2} \cdots \mathbf{x}_{:,m} \end{pmatrix}$$

  be a data matrix where each row, $\mathbf{x}_{i,:} = (x_{i,1}, x_{i,2}, \ldots, x_{i,m}), 1 \leq i \leq n$, is an m-dimensional observation vector, and columns of x^n are denoted by $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$.
- The multidimensional rows $\mathbf{x}_{i,:}$ are assumed i.i.d.
- There can be dependencies between the dimensions (columns $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$).

Bayesian networks

- In general, a Bayesian network is a DAG representing a set of independence assumptions
- In particular, given a Bayesian network, the joint distribution factorizes as a product of local distributions, each conditioned on the parents of a node:

$$p(x^n \,;\, \theta) = \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \,;\, \theta_{j|\mathrm{Pa}_j})$$



- E.g., given the network above, we get

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_2)P(X_5|X_2, X_4)P(X_6)P(X_7|X_6)P(X_8|X_7).$$

NML for Bayesian networks (Definition)

$$p_{\mathrm{NML}}(x^n) := \frac{p(x^n ; \hat{\theta}(x^n))}{C_{\mathcal{M}}(n)} \quad , \qquad C_{\mathcal{M}}(n) = \int_{\mathcal{X}^n} p(x^n ; \hat{\theta}(x^n)) \, dx^n .$$

Assuming discrete data x^n, given a Bayesian network structure $\mathcal{G}$, we get

$$p(x^n ; \hat{\theta}(x^n)) = \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j ; \hat{\theta}(x^n)), \text{and}$$

$$C_{\mathcal{G}}(n) = \sum_{x^n} \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j ; \hat{\theta}(x^n)) \quad .$$

The required maximum likelihood parameters are easily evaluated since it is well known that the ML parameters are equal to the relative frequencies:

$$\hat{\theta}_{j|\mathrm{Pa}_j}(r, \mathbf{s}) = \frac{\left|\{i \ : \ x_{i,j} = r, \mathrm{pa}_{i,j} = \mathbf{s}\}\right|}{\left|\{i' \ : \ \mathrm{pa}_{i',j} = \mathbf{s}\}\right|} \quad ,$$

where $|S|$ denotes the cardinality of set S.

NML for Bayesian networks (Computation)

- In general, the problem is NP-hard (Koivisto, 2006)
  - Solution: redefine the goal (factorized NML)
- A single multinomial: linear
  - $\mathcal{C}_{K+2}(n) = \mathcal{C}_{K+1}(n) + \frac{n}{K} \cdot \mathcal{C}_K(n)$.
  - Applies to products of multinomials
  - Applies to histogram density estimation (more later)
- The Naive Bayes structure: quadratic
- Tree-structured structures: exponential with respect to the number of values of the inner (=not the root or a leaf) nodes

Open problems

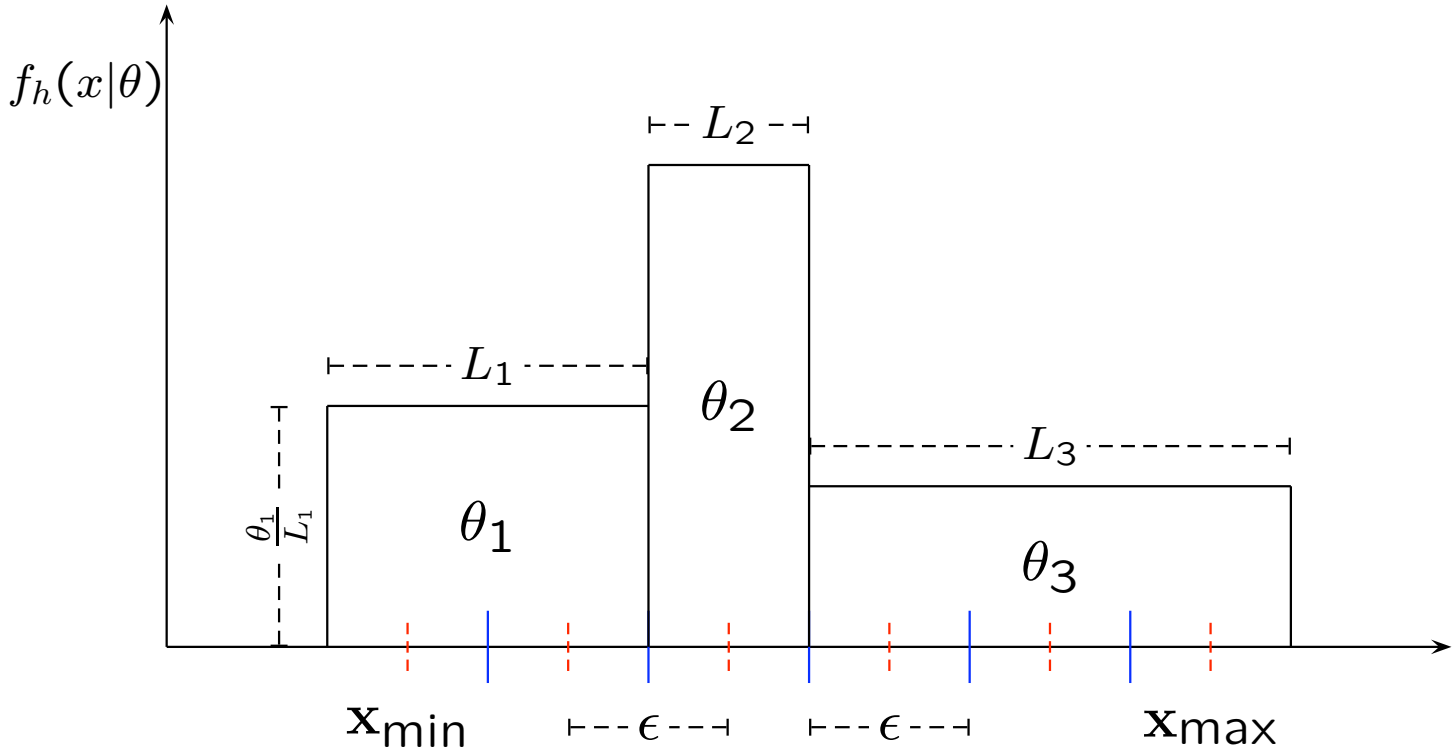
- Can the NB or tree-structured case be made more efficient?
- Are there other subclasses of Bayes nets of interest?
- Approximations of NML, e.g. for the multinomial:

$$\log C_K(n) = \frac{K-1}{2}\log\frac{n}{2} + \log\frac{\sqrt{\pi}}{\Gamma(K/2)} + \frac{\sqrt{2}K\cdot\Gamma(K/2)}{3\Gamma\left(\frac{K}{2}-\frac{1}{2}\right)}\cdot\frac{1}{\sqrt{n}} + \left(\frac{3+K(K-2)(2K+1)}{36} - \frac{\Gamma^2(K/2)\cdot K^2}{9\Gamma^2\left(\frac{K}{2}-\frac{1}{2}\right)}\right)\cdot\frac{1}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

- Model search
  - Search space is super-exponential
- How the encode the model structure?

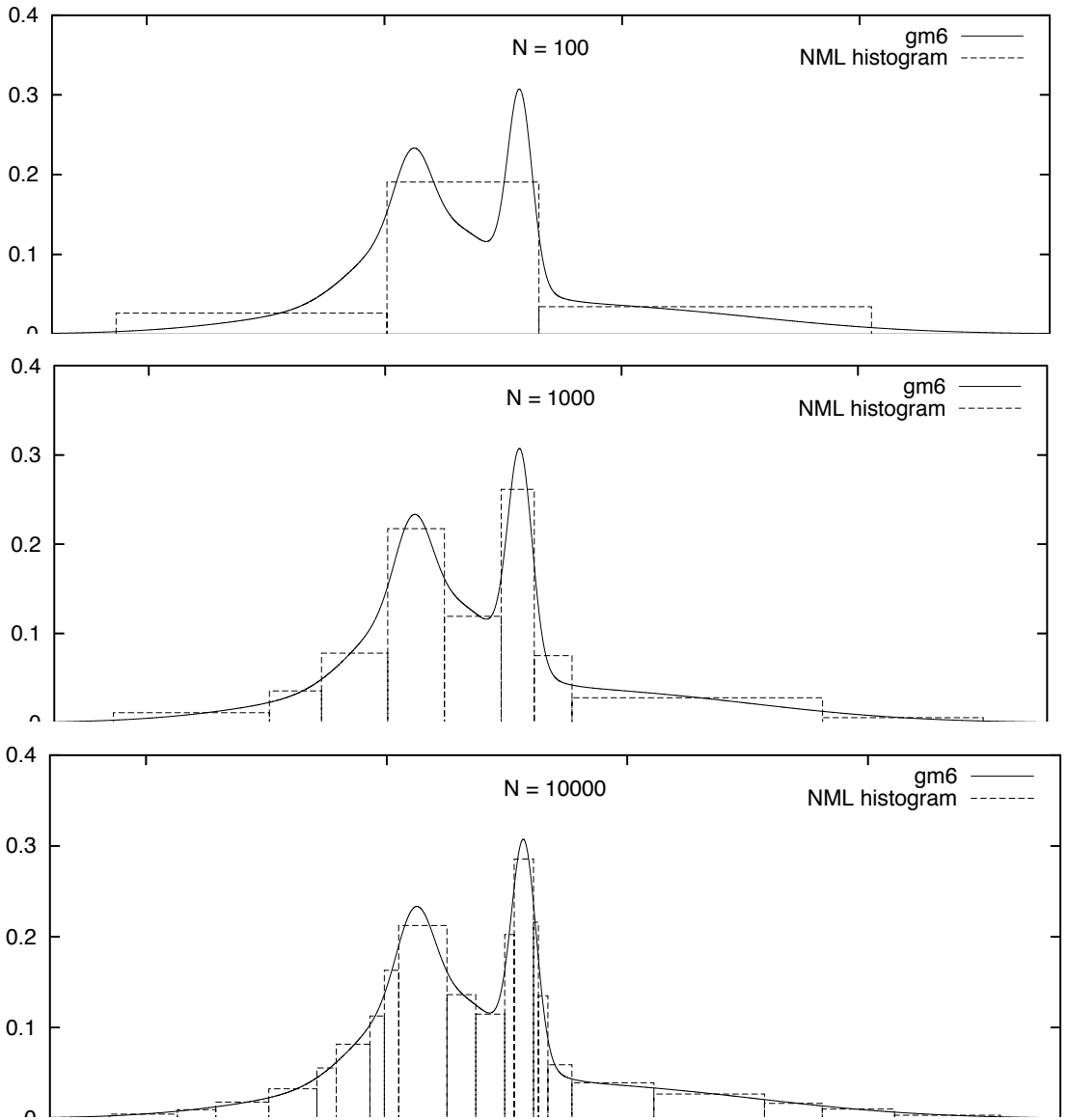
$$P(x^n, \mathcal{G}) = P(x^n \mid \mathcal{G})P(\mathcal{G}).$$

Application: histogram density estimation



- Model structure: bin borders (bin widths)
- Model parameters: bin probability masses
- Model structure selection criterion: multinomial NML
- Model search: dynamic programming

MDL histogram density estimation



Conclusions

- The "standard" Bayesian mixture model selection criterion can be problematic (more about this in Silander's talk)
- NML avoids these problems but is computationally challenging
- The factorized NML criterion offers a computationally efficient alternative
- There are several interesting open problems related to Bayesian network structure learning
- A nice application of the results obtained so far: histogram density estimation (code available by request)