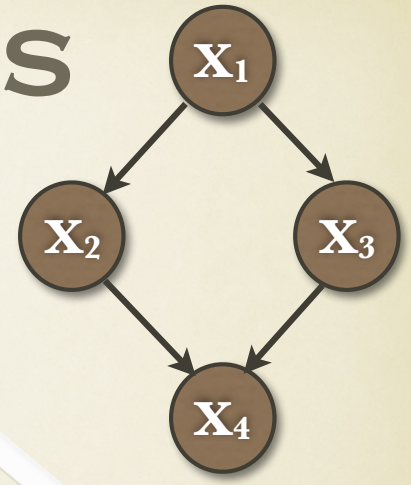
Towards Perfect Density Estimation

Petri Myllymäki
Professor, Department of Computer Science
Director, Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Aalto University
Finland

Bayesian Networks



- A Bayesian network structure ***G*** is a DAG that represents
  - A set of (pairwise) conditional independence statements
  - A ***set*** of joint probability distributions
  - A ***factorization*** of these joint probability distributions into a product of conditional

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

- The probability distributions can be indexed by a parameter vector θ instantiating the conditional probabilities determined by the factorization

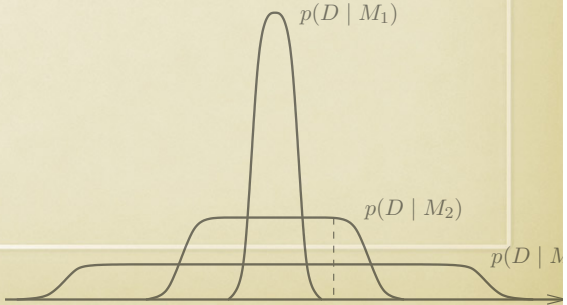
$$P(X_i = k \mid G_i = j, \Theta) = \theta_{ijk}$$

Model probability

- $P(G|D) = P(D|G)P(G)/P(D)$
- Assume (for the moment) that the ***structure prior*** $P(G)$ is uniform (and hence can be ignored)
  - $P(G|D) \propto P(D|G)$ ***marginal likelihood***
  - $P(D|G) = \int P(D|\theta,G)P(\theta|G)d\theta$
- Pick up a suitable ***parameter prior*** $P(\theta|G)$, and the rest is just calculus?
- ***BUT...which prior?***

Objective Learning

- Assume a "data-driven" scenario where no (or very little) useful prior information on the domain exists
  - New problem domains
  - Models with hidden variables
- How to construct a **non-informative, objective scoring function** for selecting or weighting different models?
- How to gain a better understanding on the effect of the parameter prior in structure learning?



Marginal Likelihood

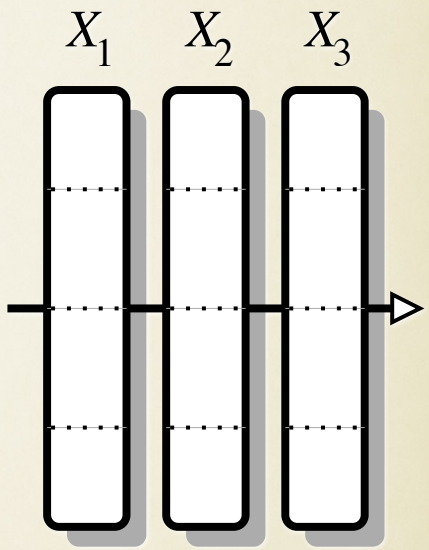
- Assuming
  - discrete i.i.d. data
  - multinomial distributions for the conditionals, with the Dirichlet (conjugate) priors
  - parameter independence and modularity
  - no missing data, no hidden variables

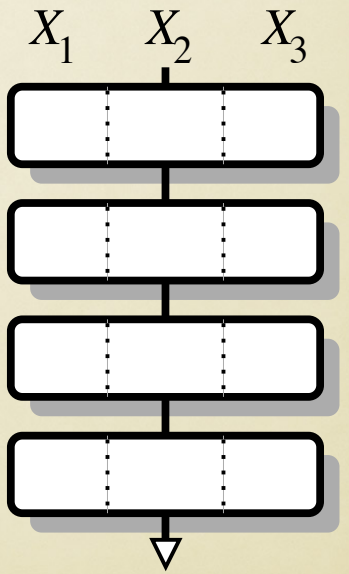
then the marginal likelihood of a BN can be computed by the ***Bayesian-Dirichlet (BD)*** score

$$P(D \mid G) = \int P(D \mid \theta, G) P(\theta \mid G) d\theta = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{r_i} \alpha_{ijk})}{\Gamma(\sum_{k=1}^{r_i} (\alpha_{ijk} + N_{ijk}))} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}.$$

Decomposability

- The BD score is decomposable: marginal likelihood factorizes like the likelihood, and can be computed as a product (a sum) of local scores
- With the model used here, the marginal likelihood distribution P(D | G) is actually equivalent to one of the distributions in the set defined by G, obtained by setting each parameter to its expected value
- N.B. Marginal likelihood can be also computed sequentially (row by row)



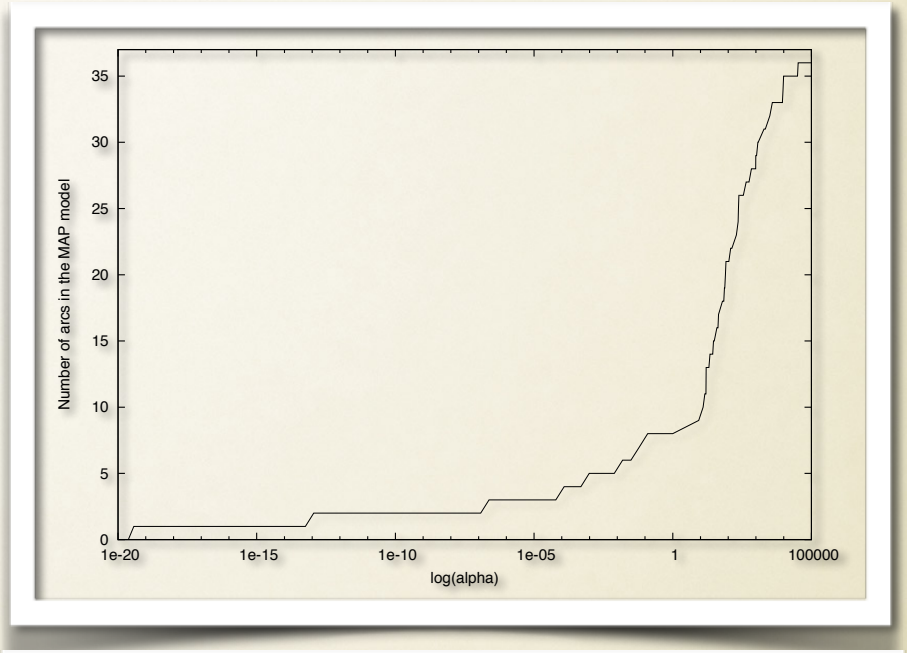


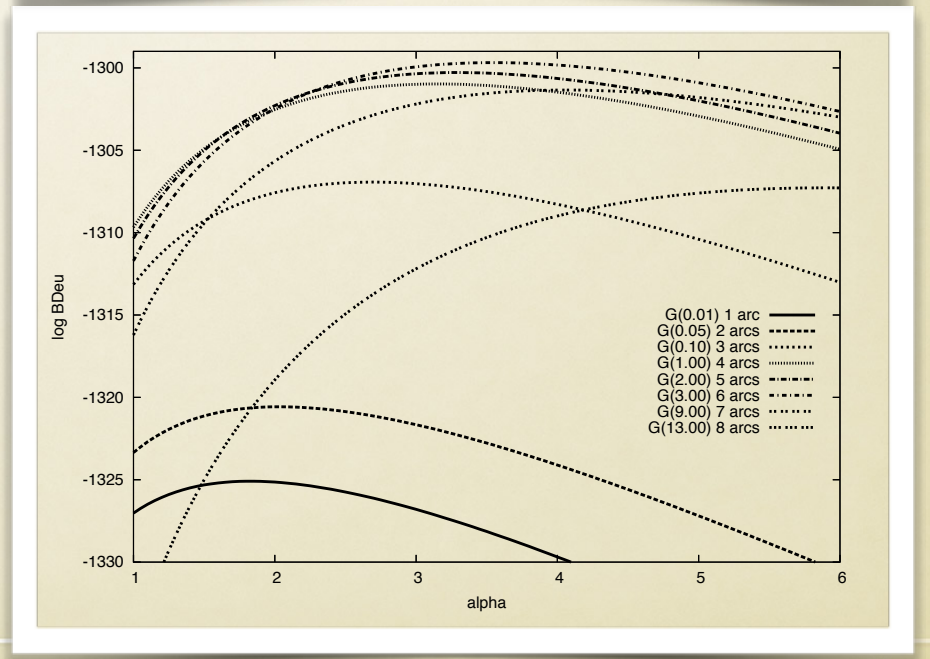
Likelihood equivalence

- Equivalence: If G_1 and G_2 represent the same set of conditional independence assumptions, then $\text{Score}(G_1) = \text{Score}(G_2)$
- $P(D \mid G_1) = P(D \mid G_2)$ if (and only if) the hyperparameters are of a certain form:
  - These ***BDeu priors*** are obtained by setting $\alpha_{ijk} = \boldsymbol{\alpha}/(q_i r_i)$, where $\boldsymbol{\alpha}$ is a single scaling parameter called the ***equivalent sample size (ESS)***

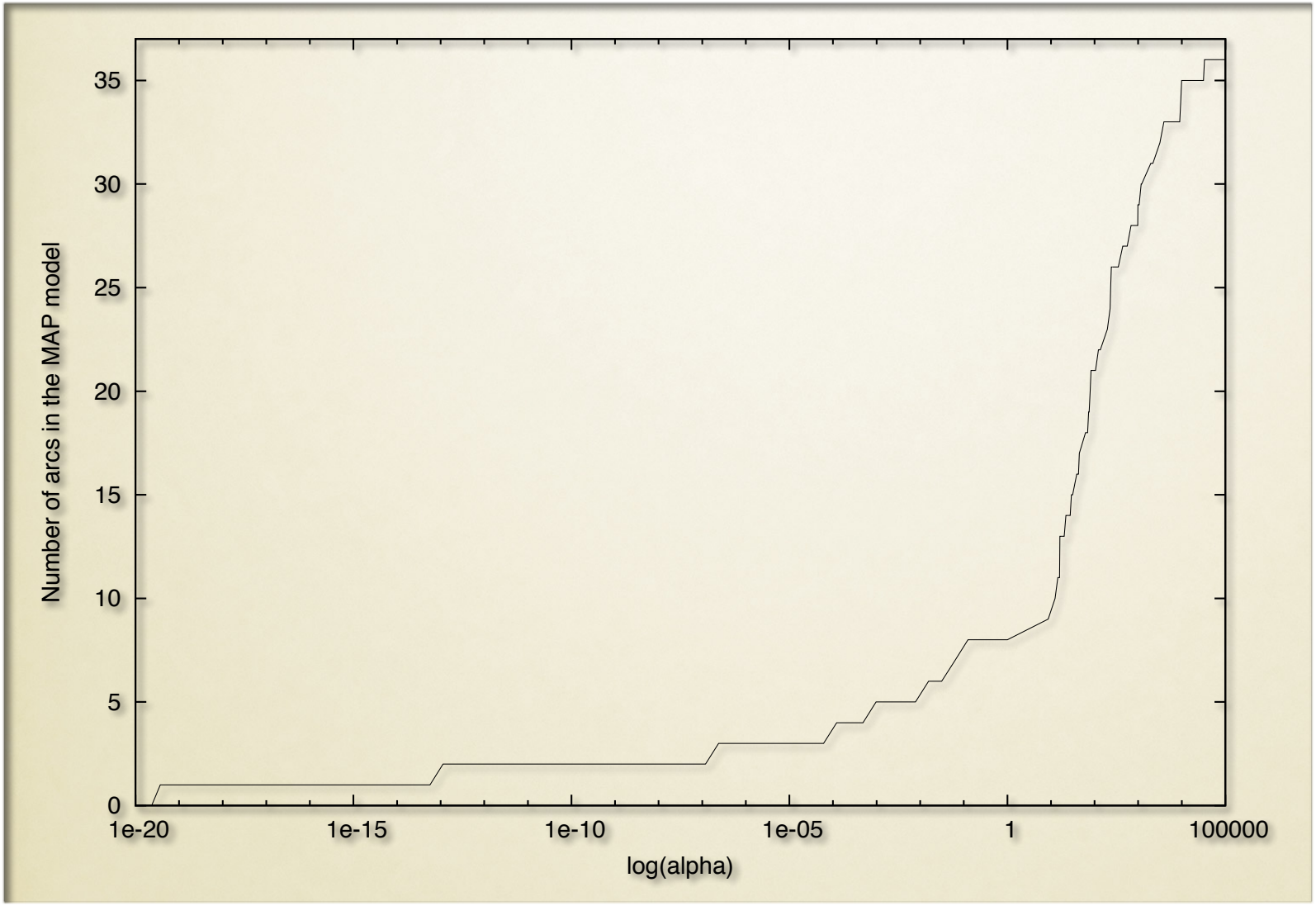
Sensitivity to Priors

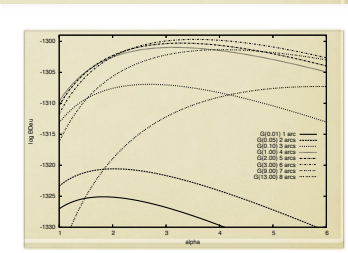
- The BDeu score is likelihood equivalent and decomposable, but requires setting the value of the ESS hyperparameter
- Unfortunately, the score is quite sensitive to this number (Silander & Myllymäki, UAI-2007)



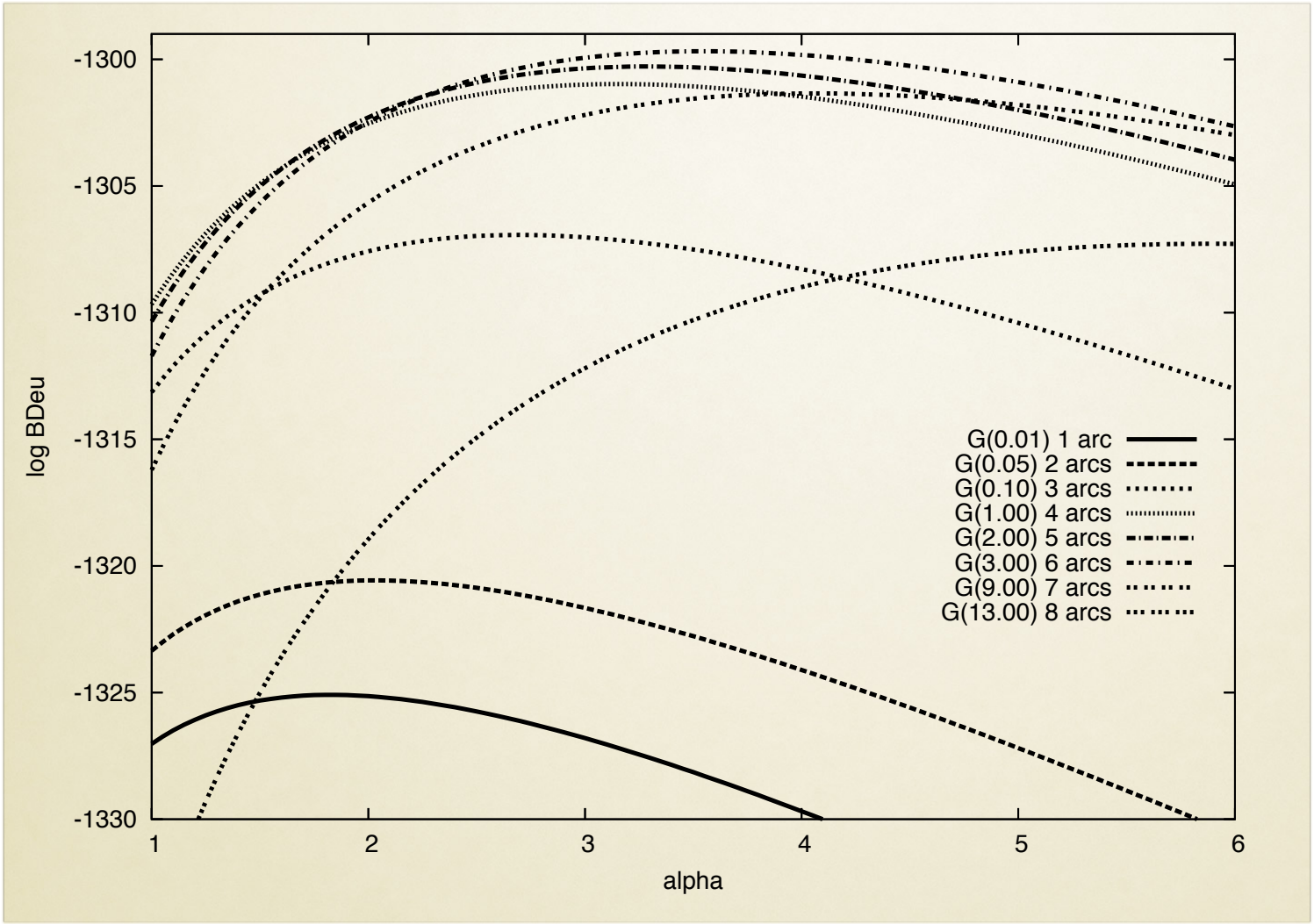


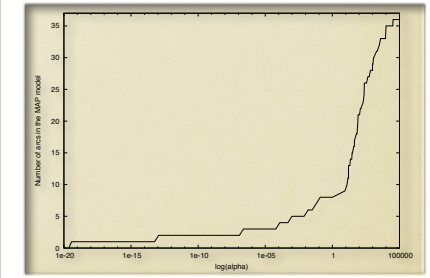
Sensitivity to Priors





Sensitivity to Priors



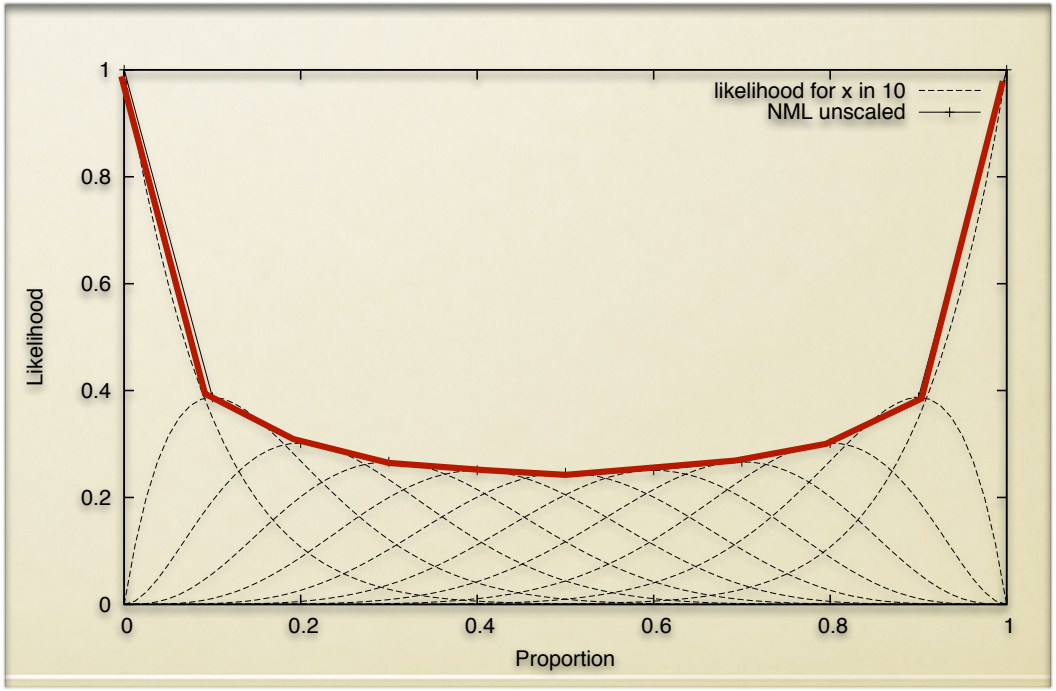


Normalized Maximum Likelihood (NML)

- Let D´ be the observed data, and denote the corresponding ***maximum likelihood estimator*** by $\theta^*(D')$
- $P(D \mid \theta^*(D'))$ is a distribution over D
- $P(D \mid \theta^*(D))$ is NOT and

  $\Sigma_D P(D \mid \theta^*(D)) > 1.$

The NML distribution can **not** be constructed as a BD distribution with some prior (but the Jeffreys prior gets close)



Optimality of NML

- The NML distribution is given by

$$P(D) = \frac{P(D;\hat{\theta}(D))}{\sum_{D'} P(D';\hat{\theta}(D'))}$$

- ***NML is the unique min-max (i.e. worst-case) optimal distribution*** satisfying

$$\min_{Q()} \max_{D} \log \frac{P(D;\hat{\theta}(D))}{Q(D)}$$

(Shtarkov, 1987)

Computing the NML

$$P(D) = \frac{P(D; \hat{\theta}(D))}{\sum_{D'} P(D'; \hat{\theta}(D'))}$$

- Brute-force calculation of the normalizing constant (also called parametric complexity) is obviously computationally infeasible
- NML is likelihood equivalent but ***not*** decomposable

Computing the NML

$$-\log P_G(D) = -\log P_G(D;\hat{\theta}(D)) + \log \sum_{D'} P_G(D';\hat{\theta}(D'))$$

- Fisher approximation

$$-\log P_G(D) = -\log P_G(D;\hat{\theta}(D)) + \frac{k}{2}\log\frac{n}{2\pi} + \log \int_\theta \sqrt{|\ I(\theta)\ |}d\theta + o(1)$$

- Bayesian Information Criterion (BIC)

$$-\log P_G(D) \approx -\log P_G(D;\hat{\theta}(D)) + \frac{k}{2}\log n$$

MDL ≠ BIC

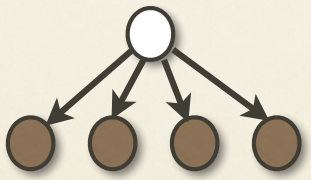
- NML-based approaches can be motivated by the information-theoretic ***Minimum Description Length (MDL)*** principle of Rissanen
- Historical background:
  - 1970-80: "Two-part MDL" = BIC
  - 1980s: Marginal likelihood codes (distributions)
  - 1996→: NML

Special Cases

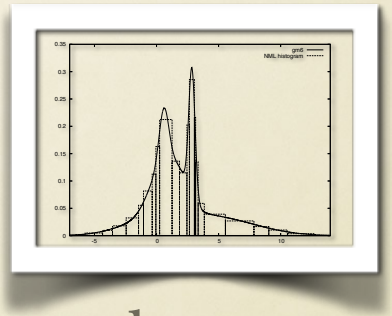
- Single multinomial:
  - The parametric complexity C(n,k) can be computed via linear-time recursion (Kontkanen & Myllymäki, 2007):

$$C(k+2,n) = C(k+1,n) + n/k\, C(k,n)$$

- This recursive formula can be used as a building block for
  - NML clustering



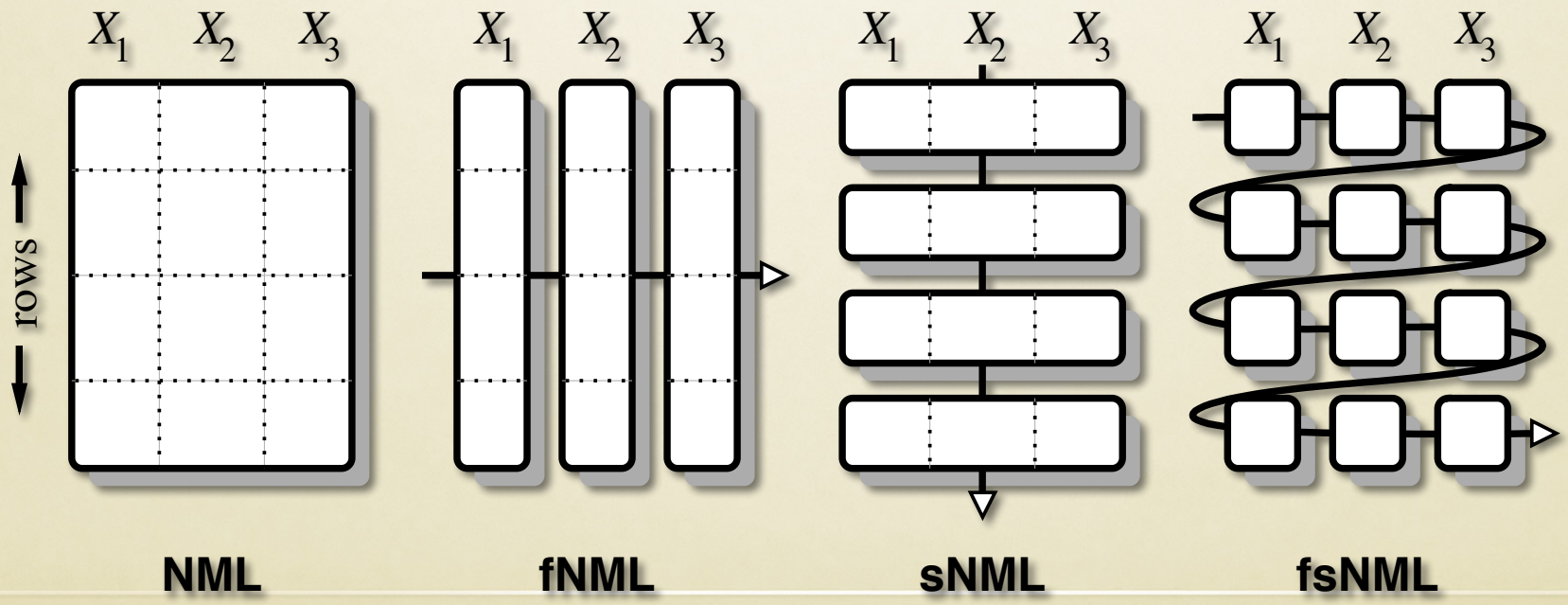
  - NML histogram density estimation



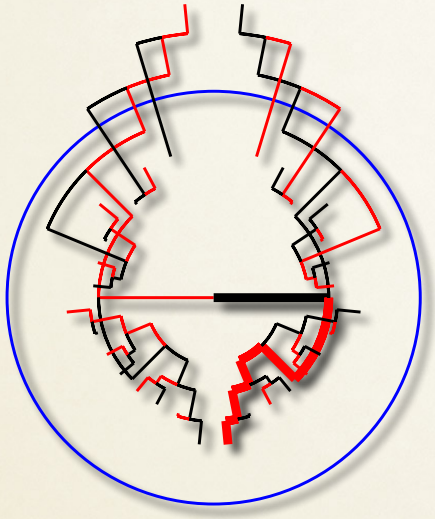
- Tree-structured Bayesian networks: NML can be computed in polynomial time with respect to data size (but the time is exponential wrt. number of values of non-leaf nodes)

Variants of NML

- *Factorized NML*
  - decomposable but not likelihood equivalent
- *Sequential NML*
  - likelihood equivalent but not decomposable

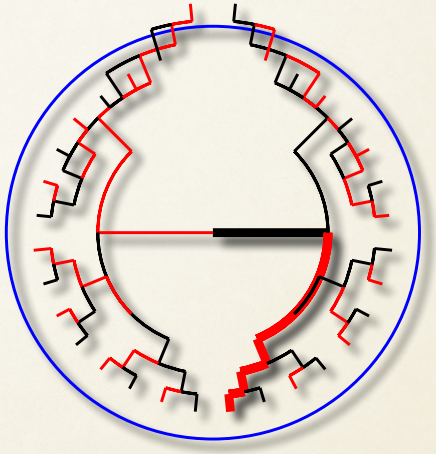


Non-informative priors?



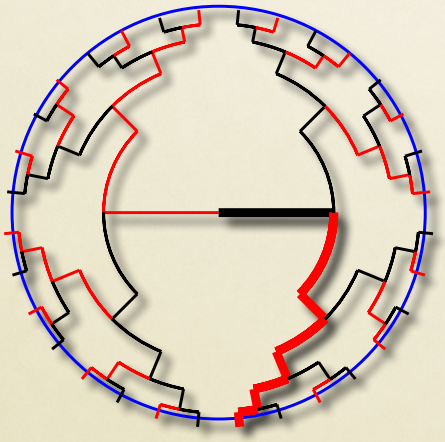
Laplace

BD + uniform prior

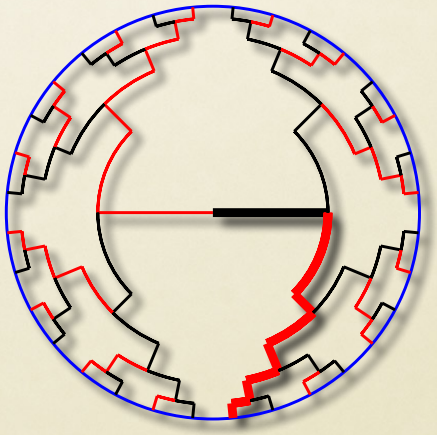


Krichevsky–Trofimov

BD + the Jeffreys prior

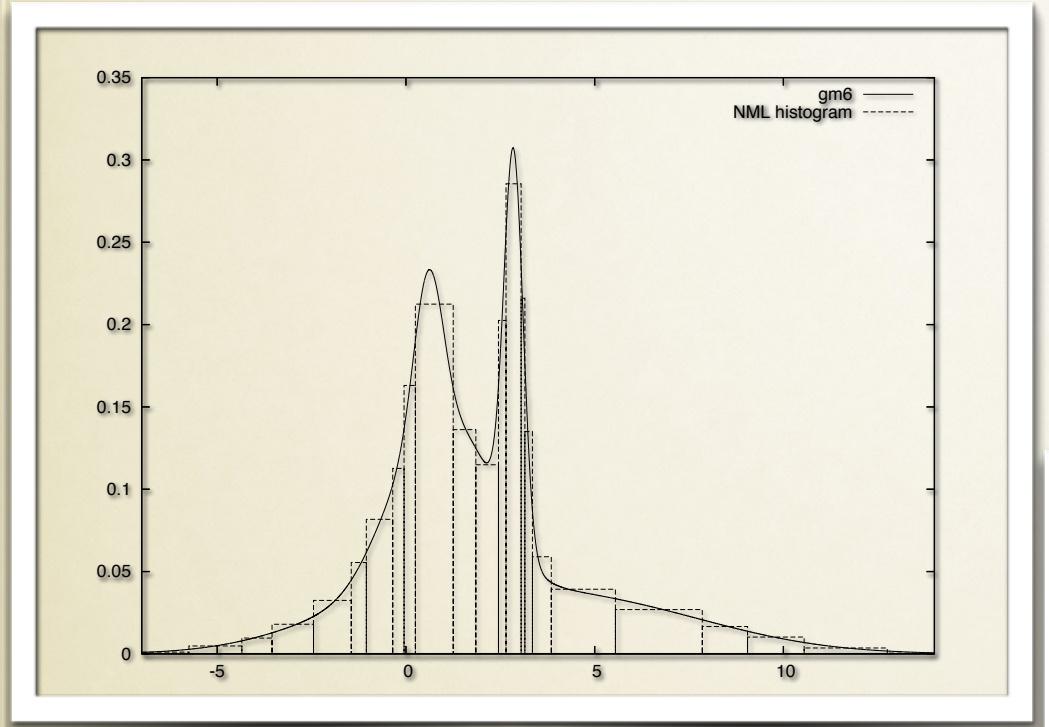


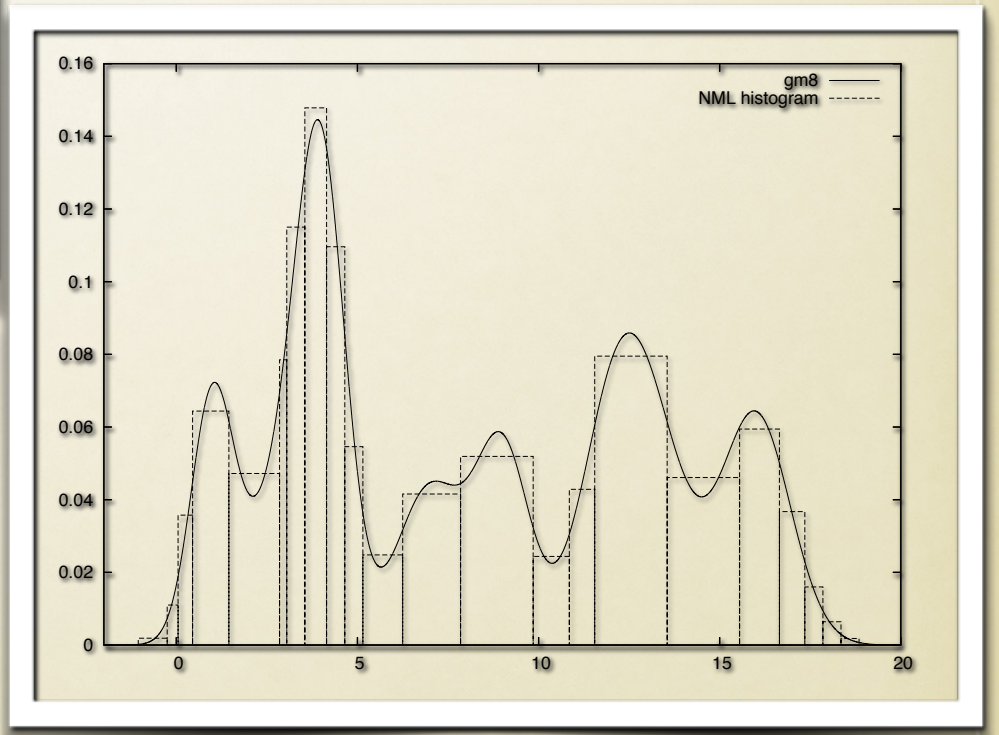
sNML-2



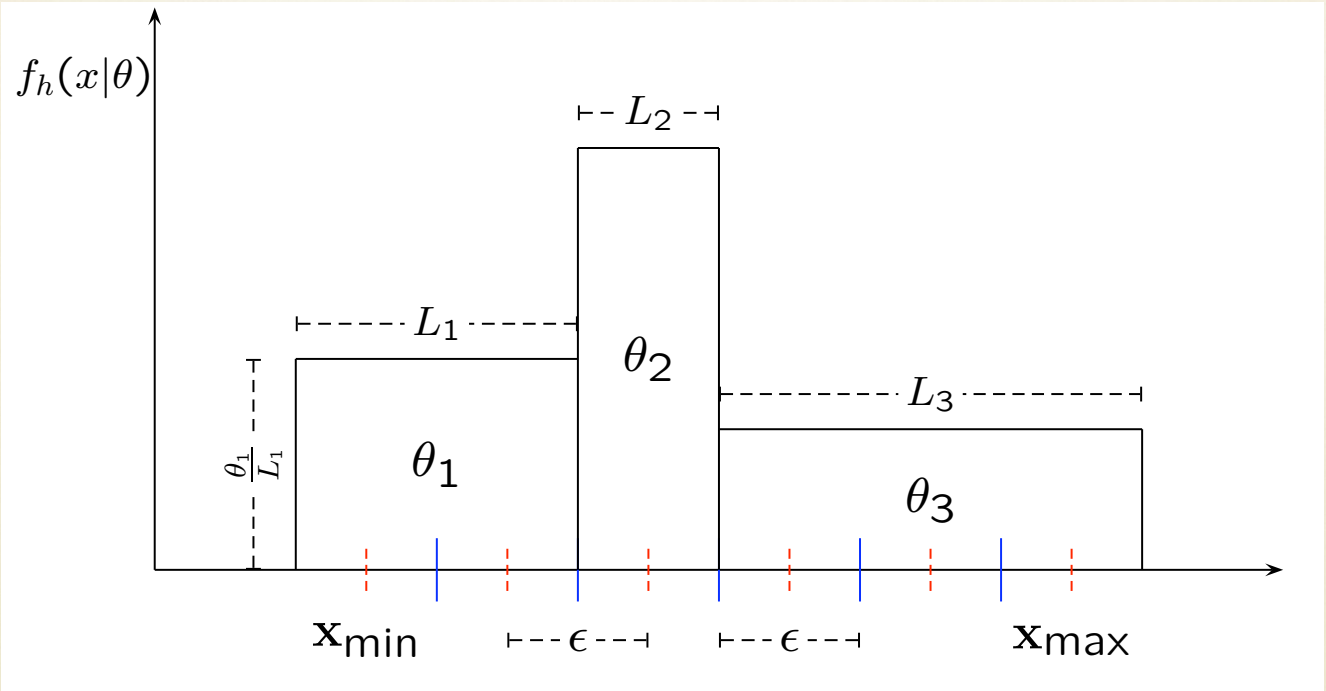
NML

NML Histograms



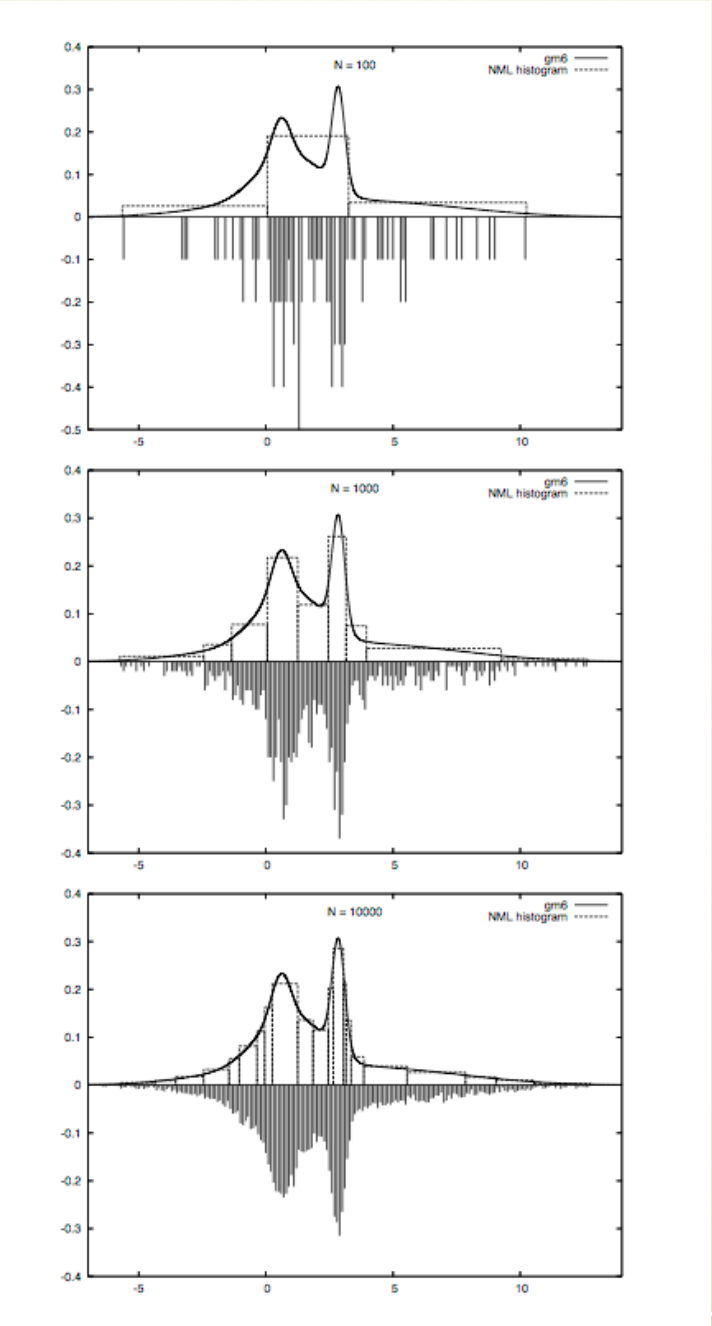


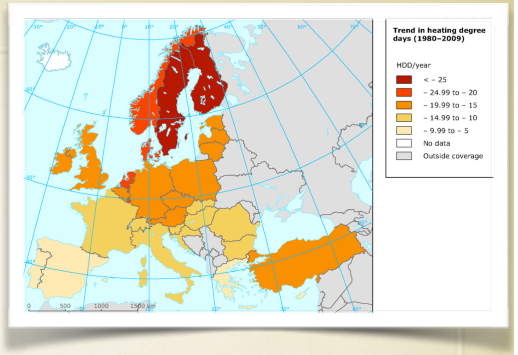
NML for Piecewise Constant Density Estimation

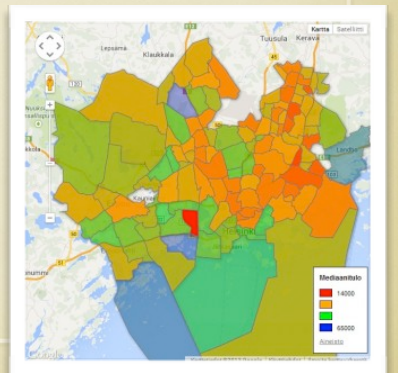


- Model structure (G): number of bins + locations of bin borders (bin widths)
- Model parameters (θ): bin probability masses
- Model structure selection criterion: multinomial NML (linear recursion)
- Model search: dynamic programming

It Works!


N = 100
gm6
NML histogram
N = 1000
gm6
NML histogram
N = 10000
gm6
NML histogram



Obvious Extensions

- Structure prior P(G): now uniform wrt. to potential bin border points? "Recursive NML"?
- Comparisons with Bayesian alternatives?
- Piecewise linear (or exponential) density functions?
- 2 (or n) -dimensional NML density functions?
  - The NML criterion is not the problem here, but the dynamic programming breaks down (?)
  - What type of shapes?
    - Rectangles? Voronoi Tessellations?