
Efficient Computation of NML for Bayesian Networks

Petri Myllymäki

Department of Computer Science & Helsinki Institute for Information Technology
P.O. Box 68, FI-00014 University of Helsinki, Finland

Abstract

Bayesian networks are parametric models for multidimensional domains exhibiting complex dependencies between the dimensions (domain variables). A central problem in learning such models is how to regularize the number of parameters; in other words, how to determine which dependencies are significant and which are not. The *normalized maximum likelihood (NML)* distribution or code offers an information-theoretic solution to this problem. Unfortunately, computing it for arbitrary Bayesian network models appears to be computationally infeasible, but we show how it can be computed efficiently for certain restricted type of Bayesian network models.

1 Normalized Maximum Likelihood

Let

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

be a data matrix where each row, $\mathbf{x}_{i,:} = (x_{i,1}, x_{i,2}, \dots, x_{i,m})$, $1 \leq i \leq n$, is an m -dimensional observation vector, and columns of x^n are denoted by $\mathbf{x}_{:,1}, \dots, \mathbf{x}_{:,m}$.

A parametric probabilistic model $\mathcal{M} := \{p(x^n; \theta) : \theta \in \Theta\}$, where Θ is a parameter space, assigns a probability mass or density value to the data. A *universal model* for \mathcal{M} is a single distribution that, roughly speaking, assign almost as high a probability to any data as the the maximum likelihood parameters $\hat{\theta}(x^n)$.

Formally, a universal model $\hat{p}(x^n)$ satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \frac{p(x^n; \hat{\theta}(x^n))}{\hat{p}(x^n)} = 0 , \quad (1)$$

i.e., the log-likelihood ratio, often called the ‘regret’, is allowed to grow sublinearly in the sample size n . The celebrated *normalized maximum likelihood (NML)* universal model [19, 22]

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

is the unique minimax optimal universal model in the sense that the worst-case regret is minimal. In fact, it directly follows from the definition that the regret is a constant depending only on the sample size n :

$$\ln \frac{p(x^n; \hat{\theta}(x^n))}{p_{\text{NML}}(x^n)} = \ln C_{\mathcal{M}}(n) .$$

For some model classes, the normalizing factor is finite only if the range \mathcal{X}^n of the data is restricted, see e.g. [19, 20, 2]. For discrete models, the normalizing constant, $C_{\mathcal{M}}(n)$, is given by a sum over all data matrices of size $m \times n$:

$$C_{\mathcal{M}}(n) = \sum_{x^n \in \mathcal{X}^n} p(x^n; \hat{\theta}(x^n)) .$$

2 Bayesian Networks

Let us associate with the columns, $\mathbf{x}_{:,1}, \dots, \mathbf{x}_{:,m}$, a directed acyclic graph (DAG), \mathcal{G} , so that each column is represented by a node. Each node, $X_j, 1 \leq j \leq m$, has a (possibly empty) set of *parents*, Pa_j , defined as the set of nodes with an outgoing edge to node X_j . Without loss of generality, we require that all the edges are directed towards increasing node index, i.e., $\text{Pa}_j \subseteq \{1, \dots, j-1\}$. If this is not the case, the columns in the data, and the corresponding nodes in the graph, can be simply relabeled, which does not change the resulting model. Figure 1 gives an example.

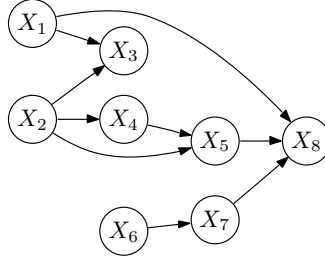


Figure 1: An example of a directed acyclic graph (DAG). The parents of node X_8 are $\{X_1, X_5, X_7\}$. The descendants of X_4 are $\{X_5, X_8\}$.

The idea is to model dependencies among the nodes (i.e., columns) by defining the joint probability distribution over the nodes in terms of *local distributions*: each local distribution specifies the conditional distribution of each node given its parents, $p(X_j | \text{Pa}_j), 1 \leq j \leq m$. It is important to notice that these are *not* dependencies among the subsequent rows of the data matrix x^n , but dependencies ‘inside’ each row, $\mathbf{x}_{i,:}, 1 \leq i \leq n$. Indeed, in all of the following, we assume that the rows are independent realizations of a fixed (memoryless) source.

The local distributions can be modeled in various ways, but here we focus on the discrete case. The probability of a child node taking value $x_{i,j} = r$ given the parent nodes’ configuration, $\text{pa}_{i,j} = \mathbf{s}$, is determined by the parameter

$$\theta_{j|\text{Pa}_j}(r, \mathbf{s}) = p(x_{i,j} = r | \text{pa}_{i,j} = \mathbf{s}; \theta_{j|\text{Pa}_j}) \quad , \quad 1 \leq i \leq n, 1 \leq j \leq m \quad ,$$

where the notation $\theta_{j|\text{Pa}_j}(r, \mathbf{s})$ refers to the component of the parameter vector $\theta_{j|\text{Pa}_j}$ indexed by the value r and the configuration \mathbf{s} of the parents of X_j . For empty parent sets, we let $\text{pa}_{i,j} \equiv 0$. For instance, consider the graph of Fig. 1; on each row, $1 \leq i \leq n$, the parent configuration of column $j = 8$ is the vector $\text{pa}_{i,8} = (x_{i,1}, x_{i,5}, x_{i,7})$; the parent configuration of column $j = 1$ is $\text{pa}_{i,1} = 0$, etc.

The joint distribution is obtained as a product of local distributions:

$$p(x^n; \theta) = \prod_{j=1}^m p(\mathbf{x}_{:,j} | \text{Pa}_j; \theta_{j|\text{Pa}_j}) . \quad (2)$$

This type of probabilistic graphical models are called Bayesian networks [18]. Factorization (2) entails a set of conditional independencies, characterized by so called Markov properties, see [16]. For instance, the *local Markov property* asserts that each node is independent of its non-descendants given its parents, generalizing the familiar Markov property of Markov chains.

3 NML for Bayesian Networks

The NML distribution based on (2) and a fixed Bayesian network graph structure \mathcal{G} is given by

$$p_{\text{NML}}(x^n; \mathcal{G}) = \frac{\prod_{j=1}^m p(\mathbf{x}_{:,j} | \text{Pa}_j; \hat{\theta}(x^n))}{C_{\mathcal{G}}(n)}, \quad (3)$$

where

$$C_{\mathcal{G}}(n) = \sum_{x^n} \prod_{j=1}^m p(\mathbf{x}_{:,j} | \text{Pa}_j; \hat{\theta}(x^n)). \quad (4)$$

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

$$\hat{\theta}_{j|\text{Pa}_j}(r, \mathbf{s}) = \frac{|\{i : x_{i,j} = r, \text{pa}_{i,j} = \mathbf{s}\}|}{|\{i' : \text{pa}_{i',j} = \mathbf{s}\}|}, \quad (5)$$

where $|S|$ denotes the cardinality of set S . However, direct summing over all possible data matrices is not tractable except in toy problems where n and m are both very small.

For a single (independent) multinomial variable with K values, the normalizing constant can be computed in quadratic time using the recursion [7, 11]:

$$C_K(n) = \sum_{r_1+r_2=0}^n \frac{n!}{r_1!r_2!} \left(\frac{r_1}{n}\right)^{r_1} \left(\frac{r_2}{n}\right)^{r_2} \cdot C_{K^*}(r_1) \cdot C_{K-K^*}(r_2), \quad (6)$$

which holds for all $K^* = 1, \dots, K-1$. A straightforward algorithm based on this formula can be used to compute $C_K(n)$ in time $\mathcal{O}(n^2 \log K)$. In [5, 9] the quadratic-time algorithm was further improved to $\mathcal{O}(n \log n \log K)$ by writing (6) as a convolution-type sum and then using the Fast Fourier Transform algorithm. However, the relevance of this result is unclear due to severe numerical instability problems it easily produces in practice. Moreover, although these results have succeeded in removing the exponentiality of the computation of the multinomial NML, they are still superlinear with respect to n . In [10] a linear-time algorithm based on the mathematical technique of generating functions was derived for the problem. In this paper it was shown how the properties of the so-called *Cayley's tree function* [4, 1] can be used to prove the following remarkably simple recurrence formula:

$$C_{K+2}(n) = C_{K+1}(n) + \frac{n}{K} \cdot C_K(n). \quad (7)$$

It is now straightforward to write an $\mathcal{O}(n+K)$ time algorithm for computing the multinomial NML based on this result. The algorithm is also very easy to implement and does not suffer from any numerical instability problems.

The one-dimensional single multinomial case is of course not adequate for many real-world situations, where data is typically multi-dimensional and involves complex dependencies between the domain variables, but it is a useful building block that can be exploited with more complex Bayesian networks. An example of a domain where the multinomial NML can be directly applied is histogram density estimation, as demonstrated in [10].

In [11], a quadratic-time algorithm for computing the NML for a specific Bayesian network structure, usually called the Naive Bayes, was derived. In this case the Bayesian network forms a single-layer tree where one of the variables is the root, and the other variables form the leaves. This model family has been very successful in practice in mixture modeling [14], clustering of data [11], case-based reasoning [12], classification [3, 13] and data visualization [8]. The time complexity of the algorithm is $\mathcal{O}(n^2 + L)$, where L denotes the number of values of the root variable. This result was further improved in [17] to $\mathcal{O}(n^2)$. For more complex Bayesian network structures, we have been able to derive an algorithm

which runs in polynomial time with respect to the the number of values of the leave nodes, but is exponential with respect to the number of values of the non-leave nodes [15, 23]. For Bayesian networks of arbitrary complexity, it appears that the problem of computing the NML is not feasible [6]. However, recently developed new variants [21] of the standard NML offer an alternative, computationally efficient information-theoretic approach for regularizing Bayesian network models.

Acknowledgments. This work was supported in part by the Academy of Finland under the project Civi and by the Finnish Funding Agency for Technology and Innovation under the projects Kukot and PMMA. In addition, this work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence.

References

- [1] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.
- [2] S. de Rooij and P. Grünwald. An empirical study of minimum description length model selection with infinite parametric complexity. *Journal of Mathematical Psychology*, 50(2):180–192, 2006.
- [3] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Minimum encoding approaches for predictive modeling. In G. Cooper and S. Moral, editors, *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 183–192, Madison, WI, July 1998. Morgan Kaufmann Publishers, San Francisco, CA.
- [4] D.E. Knuth and B. Pittel. A recurrence related to trees. *Proceedings of the American Mathematical Society*, 105(2):335–349, 1989.
- [5] M. Koivisto. *Sum-Product Algorithms for the Analysis of Genetic Risks*. PhD thesis, Report A-2004-1, Department of Computer Science, University of Helsinki, 2004.
- [6] M. Koivisto. Parent assignment is hard for the MDL, AIC, and NML costs. In *Proceedings of The 19th Annual Conference on Learning Theory (COLT 2006)*, Lecture Notes in Computer Science 4005, pages 289–303. Springer, 2006.
- [7] P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. Efficient computation of stochastic complexity. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, pages 233–238. Society for Artificial Intelligence and Statistics, 2003.
- [8] P. Kontkanen, J. Lahtinen, P. Myllymäki, T. Silander, and H. Tirri. Supervised model-based visualization of high-dimensional data. *Intelligent Data Analysis*, 4:213–227, 2000.
- [9] P. Kontkanen and P. Myllymäki. A fast normalized maximum likelihood algorithm for multinomial data. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.
- [10] P. Kontkanen and P. Myllymäki. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233, 2007.
- [11] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri. An MDL framework for data clustering. In P. Grünwald, I.J. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2006.
- [12] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On Bayesian case matching. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning, Proceedings of the 4th European Workshop (EWCBR-98)*, volume 1488 of *Lecture Notes in Artificial Intelligence*, pages 13–24. Springer-Verlag, 1998.
- [13] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. *Statistics and Computing*, 10:39–54, 2000.
- [14] P. Kontkanen, P. Myllymäki, and H. Tirri. Constructing Bayesian finite mixture models by the EM algorithm. Technical Report NC-TR-97-003, ESPRIT Working Group on Neural and Computational Learning (Neuro-COLT), 1996.
- [15] P. Kontkanen, H. Wettig, and P. Myllymäki. NML computation algorithms for tree-structured multinomial Bayesian networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.
- [16] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

- [17] T. Mononen and P. Myllymäki. Fast NML computation for naive Bayes models. In *Proc. 10th International Conference on Discovery Science*, Sendai, Japan, October 2007.
- [18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.
- [19] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.
- [20] J. Rissanen. MDL denoising. *IEEE Transactions on Information Theory*, 46(7):2537–2543, 2000.
- [21] T. Roos, T. Silander, P. Kontkanen, and Myllymäki P. Bayesian network structure learning using factorized NML universal models. In *Information Theory and Applications Workshop*, San Diego, CA, January 2008.
- [22] Yu.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17, 1987.
- [23] H. Wette, P. Kontkanen, and P. Myllymäki. Calculating the normalized maximum likelihood distribution for Bayesian forests. In *Proc. IADIS International Conference on Intelligent Systems and Agents*, Lisbon, Portugal, July 2007.