Factorized NML Models

Petri Myllymäki Teemu Roos Tomi Silander Petri Kontkanen
Complex Systems Computation Group
Helsinki Institute for Information Technology
University of Helsinki and Helsinki University of Technology
e-mail: *firstname.lastname@hiit.fi*

Henry Tirri
Nokia Research Center
e-mail: *henry.tirri@nokia.com*

Abstract

We consider probabilistic graphical models where a directed acyclic graph represents a factorization of a joint probability distribution: the joint probability of the variables is represented as a product of conditional probabilities, one for each variable conditioned on its immediate parents in the graph. For this type of models, computing the normalized maximum likelihood (NML) is computationally very demanding. We suggest a computationally feasible alternative to NML, the factorized NML, where the normalization is done locally for each conditional distribution, and not globally.

1 Introduction

The Complex Systems Computation research group[1] (CoSCo) was established in the early 1990's at the Department of Computer Science of University of Helsinki. The group was first led by Professor Henry Tirri until 2002, and after that by Professor Petri Myllymäki. The first contact between CoSCo and Jorma Rissanen took place in 1996, in an evaluation of the HYPE research project, which was a part of a large research programme on Adaptive and Intelligent Systems, funded by Tekes, the Finnish Funding Agency for Technology and Innovation. For the evaluation, Jorma interviewed Henry in a one-to-one meeting, which did not go along quite the way we had expected. Namely, the very first thing Jorma did was to write the formula for Jeffreys prior on the board, and ask "What is this?". When Henry recognized the formula Jorma commented that he has read the papers and evidently Henry knows them so let's do some science. Then the rest of the session was spent on a pleasant conversation on recent developments of MDL. As a memento of this meeting, we still keep on the wall of our institute the drawings done during the session (see Figure 1).

All in all, it was apparent that Jorma had very carefully studied the material we had sent him beforehand, and he already had a clear opinion of our work. In his evaluation report, Jorma commends our work and points out that

The material in this paper was presented in part at the 2008 Information Theory and Applications Workshop (ITA-08), San Diego, CA, January–February 2008.

[1] `http://cosco.hiit.fi`

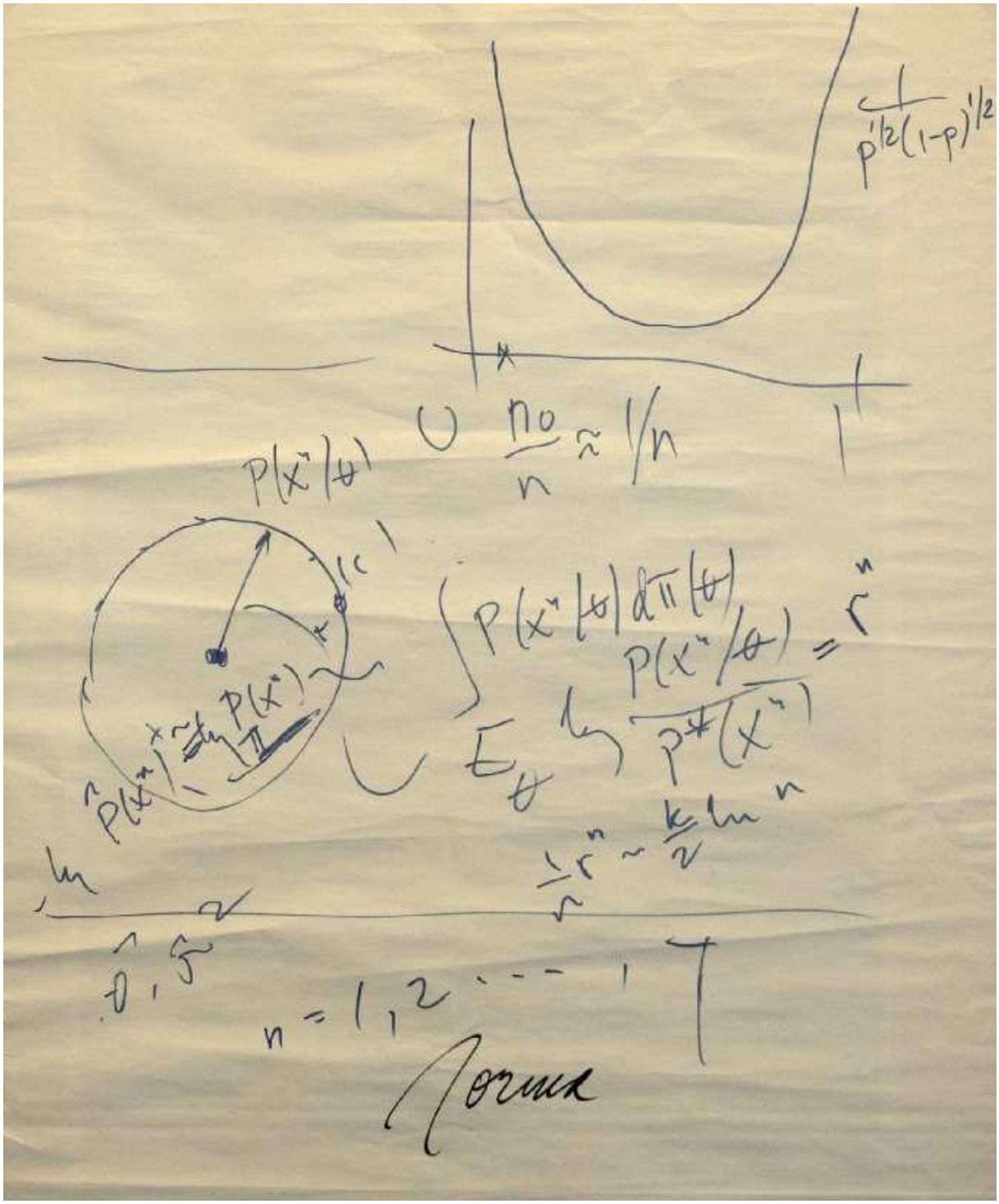

Figure 1: Jorma's notes from his first meeting with Henry Tirri in 1996.

> " It is particularly noteworthy that the difficult and important problem of determining the proper complexity of the models is done by new information-theoretic methods rather than resorting to usual ad hoc ones. "

He also had a quite clear opinion of the overall research programme, which focused on neural networks and genetic algorithms, which were popular at the time. Indeed, it is well known that Jorma is not scared to express his opinion quite directly, even if it is a negative one. In the evaluation of the research programme as a whole, Jorma chose to express his dissatisfaction somewhat indirectly, formulated cleverly in a seemingly positive statement:

> " On the whole, the research level of the teams using mainly the neural network techniques is in my opinion comparable to the general international level, which itself with a few exceptions, such as the work of A. Barron, is not particularly high. "

We were kind of an oddball in the programme, as we were just in the middle of a process of moving from neural networks and case-based reasoning to parametric probabilistic models. For the models we had started to explore—Bayesian networks, finite mixture models, Naive Bayes and other logistic regression type of classifiers—model regularization was clearly one of the central problems, and we were immediately intrigued by MDL. This interest had nothing to do with Jorma being Finnish, perhaps it was the information-theoretic approach that appealed to us as computer scientists. Actually, for a long time we discussed with Jorma in English only, and only later we have started to use more Finnish, at least for less technical discussions (involving often important topics like good food, beer, and soccer).

Our view of MDL was initially pretty ad hoc, and we, as many researchers still do, first employed the simple two-part/BIC types of codes, and later the "Bayes mixture" approach with various parameter priors [1–4]. Nevertheless, we soon felt increasing uneasiness with the arbitrariness of choosing the parameter priors, and we shared with Jorma the feeling that taking the subjective Bayesian approach is not as unproblematic as people often think, and that playing with the parameter priors is not an intuitively easy task after all, leading easily to anomalies in practical applications.

We kept seeing Jorma more and more often either in Helsinki or somewhere around the globe, and our appreciation towards him as a person and as a scientist was increasing. In addition to the pleasure of having a personal contact with Jorma, our work on MDL was greatly influenced by Peter Grünwald from CWI, Amsterdam, who met Petri Myllymäki in 1996 in a workshop organized by the NeuroCOLT working group of the European Union. Peter helped the CoSCo people to understand the new theoretical framework behind MDL, like the normalized maximum likelihood code, and we started working together in this field. Peter also came to Helsinki for a two month visit in 1997. Our joint work concentrated on issues like supervised learning, predictive distributions and choosing the parameter priors [5–11]. Quite interestingly, we were already then considering sequential (predictive) variants of MDL, which have recently gained popularity—more about them later. The co-operation between CWI and Helsinki has continued to this day, e.g. in the Pascal Network of Excellence[2], where Myllymäki and Grünwald are currently leading the Pascal Special Interest Group on Information-Theoretic Modeling. We are also jointly maintaining a popular web site[3] offering a (hopefully) useful portal to MDL-related work world-wide.

One of the active research areas in CoSCo nowadays is to study how to compute the NML criterion for Bayesian networks. This parametric model has become quite popular, and one

[2] `http://www.pascal-network.org`

[3] `http://www.mdl-research.org`

of the most popular freely available tools, the B-Course software[4], was developed and is being maintained by CoSCo. However, for practical applications, this model family introduces a couple of serious problems. First, the model structures are represented as acyclic directed graphs, which are superexponential in number. This makes the search for the best model structure a most difficult problem, which can currently be solved in reasonable time only for moderate size networks [12]. Nevertheless, perhaps even more crucial problem than how to find a good model, is the question of optimality: good in what sense?

Traditionally, in the Bayesian network community the models are evaluated by their posterior probability, which in the discrete Multinomial-Dirichlet setting can be computed in closed form, which leads to the popular BDe (Bayesian Dirichlet equivalent) score [13]. However, our recent work shows that the shape of the posterior is quite sensitive with respect to the choice of the hyperparameters of the Dirichlet prior [14]. NML would of course avoid this problem by offering a non-informative score that is not dependent on any parameter prior, but unfortunately, no efficient method for computing NML for Bayesian networks in general has yet been discovered. In the CoSCo group, we have gradually moved towards this goal by developing computationally efficient algorithms for independent multinomial variables, or equivalently, a Bayesian network with no arcs [15], for the Naive Bayes model [16, 17], and for tree-structured Bayesian networks [18, 19]. As an interesting application of the algorithm for computing the NML efficiently in the multinomial case, we can mention the minimax-optimal histogram density estimator suggested in [20].

As another active area of collaboration with Jorma, we have been focusing on MDL-based approaches to signal denoising. Starting from the original MDL denoising paper [21], we have been able to develop improved denoising methods [22], which are more robust with different levels of noise, achieve better frequency adaptivity, and employ the "soft thresholding" technique found very useful in denoising methods based on other approaches. For an illustration of denoising, see Figure 2. (The image in the example represents an Inter Milan soccer player in the 1950's. As many of us know, Jorma has always been a great fan of soccer, and a talented player himself: he even got an invitation in the early 1950's for a try-out in Milan, but the entrance examination for the Helsinki University of Technology was at the same time, and Jorma made, according to his own words, "a wrong decision" and chose science over football. Later he hurt his knee doing pole vault during his military service in the Finnish army, which finally ended any ideas about a potential career as a professional football player. This was a lucky strike for the IBM soccer team, who enjoyed having Jorma play for them for many years.)

One of the conclusions of the still ongoing work on denoising is the observation that the "model index", identifying the optimal subset of wavelet coefficients, forms a practically important part of the overall code length, and should not be ignored like was done in the original denoising paper. A similar phenomenon was observed already in the context of clustering [16]. However, Jorma was not after all very surprised by the result: he had of course always been aware of the missing part of his code, he just never thought it would make a difference in practice.

As the problem of computing NML for Bayesian networks is so difficult, we started to consider alternative solutions, other similar type of scoring functions that could be used instead of NML. It is probably appropriate to point out that also the non-informative Bayesian solution of using the Jeffreys prior is computationally NP-hard [10]. As already noted, we were already early on quite interested in predictive forms of MDL, while Jorma did not seem to share our interest. "Forget about prediction" was a frequently heard comment made by him when we tried to suggest exploring this area. One could have thought that Jorma did not want to touch the elaborate

[4] `http://b-course.hiit.fi`

Figure 2: The MDL denoising methods in action. *Top row* (from left) Original (size 128×128); noisy (noise std.dev. 20.0); original MDL denoising [21]. *Bottom row:* Left to right, gradual improvements of the MDL denoising method [22].

NML framework he had created, but as it would turn out, nothing was further from the truth. Already since 2004–2005, having studied a paper by Takimoto and Warmuth [23], we had started discussing the idea of sequential type NML variants in our group. Even though we found the topic potentially worthwhile, we couldn't see any obvious extensions beyond the basic idea. When we finally introduced the idea to Jorma in 2006, he was suddenly full of new ideas, leading to sequential NML (see Sec. 3 below) and many other novel innovations, and he was more than ready to abandon the "old" NML as obsolete—much more than we were! All in all, Jorma has often proved to be so fast and dynamic in his work that we, many being less than half of his age, have had hard time trying to keep up.

As the most recent result of our research on NML-like universal models for Bayesian networks, we introduce in this paper the *factorized NML* (fNML) model. The rest of the paper is organized as follows: In Sections 2 and 3 we discuss the normalized maximum likelihood (NML) and sequentially normalized maximum likelihood (sNML) models, respectively. In Section 4 we review the basics of Bayesian networks. The factorized NML model is introduced in Section 5, where it is also shown to be computationally feasible for all Bayesian networks. The new model is philosophically a relative of the sequential NML models discussed in Section 3. Finally, in Section 6, we present experimental results, demonstrating that fNML compares favorably in a model selection task, relative to the current state-of-the-art.

2 Normalized Maximum Likelihood Models

Before describing the sequential NML and factorized NML models, we fix some notation and review some basic properties of the well-known NML model. 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 $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$.

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

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

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 [24, 25]

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

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 dependent only on the sample size n:

$$\ln \frac{p(x^n \ ; \ \hat{\theta}(x^n))}{p_{\mathrm{NML}}(x^n)} = \ln C_n \ .$$

For some model classes, the normalizing factor is finite only if the range $\mathcal{X}^n$ of the data is restricted, see e.g. [21, 24, 26]. For discrete models, the normalizing constant, C_n, is given by a sum over all data matrices of size $m \times n$:

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

The practical problem arising in applications of the NML universal model is then to evaluate the normalizing constant. For continuous models the integral can be solved in closed form for only a few specific models. For discrete models, the time complexity of the naive solution, i.e., summing over all possible data matrices, grows exponentially in both n and m, and quickly becomes intractable. Even the second-most naive solution, summing over equivalence classes of matrices, sharing the same likelihood value, is usually intractable even though often polynomial in n.

The usual Fisher information approximation [24]

$$\ln C_n = \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int_{\Theta} \sqrt{\det I(\theta)} \, d\theta + o(1) \ ,$$

where k is the dimension of the parameter space, is also non-trivial to apply due to the integral involving the Fisher information $I(\theta)$. Using only the leading term (with or without 2π), i.e., the BIC criterion [27], gives a rough approximation which, as a rule, performs worse in model selection tasks than more refined approximations or, ideally, the exact solution, see e.g. [28, Chap. 9].

3 Sequentially Normalized ML Models[4]

A recent family of variants of NML, called the *sequentially* (or *conditional*) *normalized maximum likelihood* (sNML) [29, 30] has similar minimax properties like NML but is often significantly easier to use in practice.

For data matrix $x^n = (\mathbf{x}_{1,:}, \mathbf{x}_{2,:}, \ldots, \mathbf{x}_{n,:})'$, the sNML-1 model is defined as

$$p_{\mathrm{sNML1}}(x^n) := \prod_{i=1}^{n} \frac{p(\mathbf{x}_{i,:} \mid x^{i-1} \; ; \; \hat{\theta}(x^i))}{K_i(x^{i-1})} \quad , \tag{2}$$

$$K_i(x^{i-1}) := \int p(\mathbf{x}_{i,:} \mid x^{i-1} \; ; \; \hat{\theta}(x^i)) \, d\mathbf{x}_{i,:} \quad , \tag{3}$$

where normalization ensures that each factor in the product is a proper density function.

In some cases it is necessary to use a separate density, say $q(x^{n_0})$, for the first n_0 observations, with n_0 large enough, so that the maximimized likelihood is well-defined for longer sequences x^i with $i > n_0$. For instance, in linear regression n_0 has to be at least the number of regressor variables plus one.

Second variant (sNML-2). There is also another variant of sNML, which we call here sNML-2. It can be defined in analogy with (2) as follows:

$$p_{\mathrm{sNML2}}(x^n) := \prod_{i=1}^{n} \frac{p(x^i \; ; \; \hat{\theta}(x^i))}{K'_i(x^{i-1})} \quad , \tag{4}$$

$$K'_i(x^{i-1}) := \int p(x^i \; ; \; \hat{\theta}(x^i)) \, d\mathbf{x}_{i,:} \quad .$$

Using the sNML-2 model is equivalent to predicting the ith observation using the standard NML model defined for sequences of length i. Formally we have

$$p_{\mathrm{NML}}(\mathbf{x}_{i,:} \mid x^{i-1}) = p_{\mathrm{sNML2}}(\mathbf{x}_{i,:} \mid x^{i-1}) \quad .$$

Note that the standard NML model is not in general a stochastic process, which makes it possible that

$$p_{\mathrm{NML}}(\mathbf{x}_{i,:} \mid x^{i-1}) \neq \sum_{\mathbf{x}_{i+1,:}} p_{\mathrm{NML}}(\mathbf{x}_{i,:}, \mathbf{x}_{i+1,:} \mid x^{i-1}) \quad , \tag{5}$$

and hence, typically two NML models, defined for sequences of different lengths, give different predictions. In contrast, both sNML-1 and sNML-2 are by definition stochastic processes, so that for them we always have an equality in (5).

[4] This section is mostly based on as yet unpublished work by Rissanen, Myllymäki, and Roos.

Regrets Visualized. Figure 3 gives a visualization of the regrets of four universal models in the Bernoulli case: the Laplace predictor ("add-one"), the Krichevsky–Trofimov predictor ("add-half"), sNML-2, and NML. For NML, the initial sequence probabilities, $q(x^t)$, are obtained from a fixed NML model, defined for $n = 5$, by summing over the possible continuations of length $n - t$.

Note that for NML, while the intermediate regrets, for $t < n$, depend on the prefix x^t, the total regret for x^n is a constant. For sNML, the difference between the regret for x^t and x^{t+1} is constant with respect to x_t but varies with x^{t-1}; in the figure this means that each pair of edges originating from the same branching point are of equal length, but their length depends on the path from the origin. For the Bernoulli model, SNML-1 is equivalent to the Laplace predictor. Figure 4 shows the regrets with $n = 5$ as a function of the number of `1`s.

Related Work. The sNML-2 model has been analysed earlier in conjuction with discrete Markov models, including as a special case the Bernoulli model, by Shtarkov [25] (see his Eq. 45). Also, Takimoto and Warmuth [23] analyze a slightly more restricted minimax problem, the solution of which agrees with sNML-2 for Markov models. Grünwald [29] uses the term "conditional NML" (CNML) for a family of universal models, conditioned on an initial sequence without considering the joint model obtained as a product of such conditional densities. Our sNML-1 corresponds to his CNML-3, and our sNML-2 corresponds to his CNML-2. The conditional mixture codes studied by Liang and Barron [31] are also closely related to sNML, and have similar minimax properties.

4 Bayesian Networks

In Sec. 5, we describe a new NML variant, similar to the sNML models discussed in the previous section. This new variant gives a computationally feasible universal model, and a corresponding model selection criterion, for general Bayesian network models. This section presents the necessary background in Bayesian networks.

First, let us associate with the columns, $\mathbf{x}_{:,1}, \ldots, \mathbf{x}_{:,m}$, 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., $\mathrm{Pa}_j \subseteq \{1, \ldots, j-1\}$. 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 5 gives an example.

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, $p(X_j \mid \mathrm{Pa}_j), 1 \leq j \leq m$. 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, $\mathrm{pa}_{i,j} = \mathbf{s}$, is determined by the parameter

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

where the notation $\theta_{j|\mathrm{Pa}_j}(r, \mathbf{s})$ refers to the component of the parameter vector $\theta_{j|\mathrm{Pa}_j}$ indexed by the value r and the configuration $\mathbf{s}$ of the parents of X_j. For empty parent sets, we let $\mathrm{pa}_{i,j} \equiv 0$.

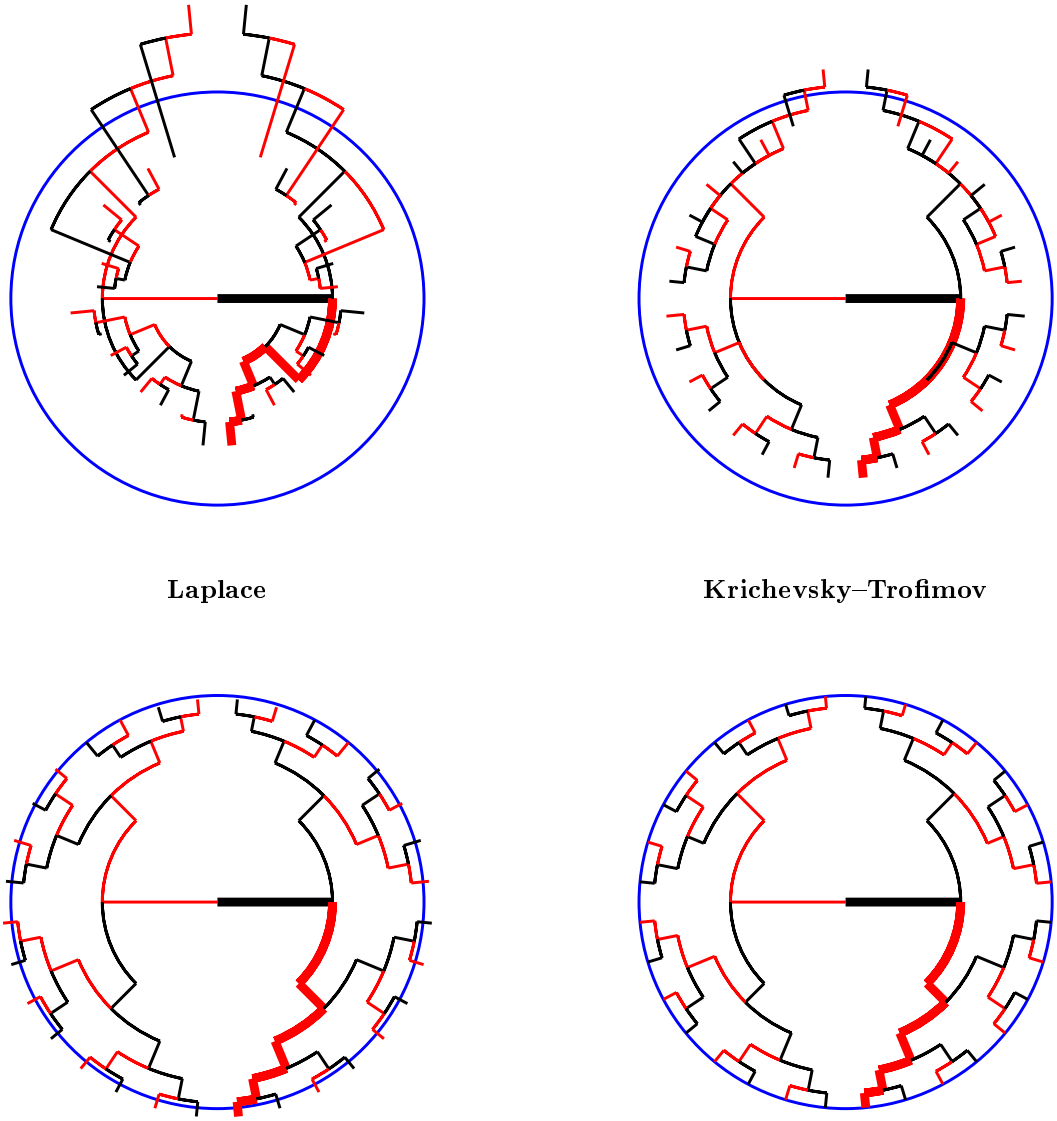

Figure 3: Regrets of four universal models in the Bernoulli case. Each path from the origin (center) to the boundary represents a binary sequence of length $n = 5$. Red edges correspond to 1s, black edges to 0s. The path for sequence 01111 is emphasized. The distances from the origin of the branching points are given by the regrets $\ln[p(x^t\ ;\ \hat{\theta}(x^t))/q(x^t)]$ for each prefix x^t. The blue circle shows the regret of NML. Note the similarity between sNML-2 and NML.

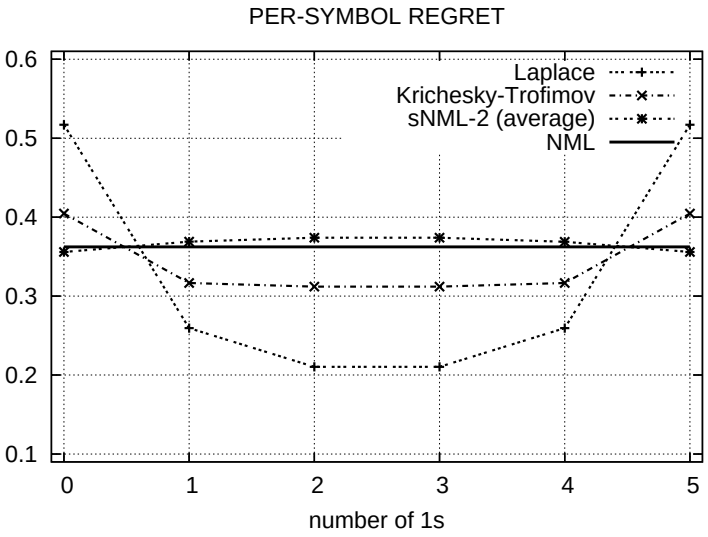

Figure 4: Per-symbol regrets of four universal models in the Bernoulli case as a function of the number of 1s in the sequence with $n = 5$ (for the same figure with $n = 30$, see [30]). For sNML-2 the regret depends not only on the number of 1s, but also on the actual sequence. (The dependency is *very* slight, see Fig. 3.) The graph shows the average regret.

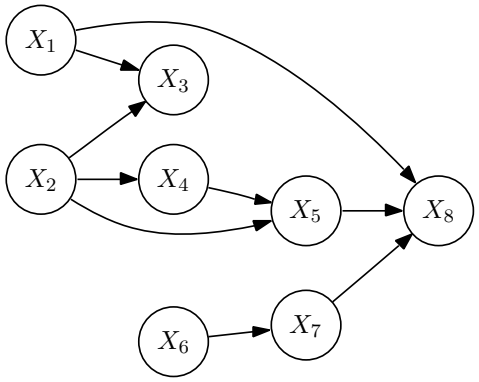

Figure 5: An example of a directed acyclic graph (DAG). The parents of node X_8 are $\{X_1, X_5, X_7\}$. The descendants of X_4 are $\{X_5, X_8\}$.

For instance, consider the graph of Fig. 5; on each row, $1 \leq i \leq n$, the parent configuration of column $j = 8$ is the vector $\mathrm{pa}_{i,8} = (x_{i,1}, x_{i,5}, x_{i,7})$; the parent configuration of column $j = 1$ is $\mathrm{pa}_{i,1} = 0$, etc.

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

$$p(x^n \; ; \; \theta) = \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \theta_{j|\mathrm{Pa}_j}) \;\; . \tag{6}$$

This type of probabilistic graphical models are called Bayesian networks [32]. Factorization (6) entails a set of conditional independencies, characterized by so called Markov properties, see [33]. 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.

It is now possible to define the NML model based on (6) and a fixed graph structure $\mathcal{G}$:

$$p_{\mathrm{NML}}(x^n \; ; \; \mathcal{G}) = \frac{\prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{C_n} \;\; , \tag{7}$$

where

$$C_n = \sum_{x^n} \prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n)) \;\; . \tag{8}$$

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

$$\hat{\theta}_{j|\mathrm{Pa}_j}(r, \mathbf{s}) = \frac{\big|\{i \; : \; x_{i,j} = r, \mathrm{pa}_{i,j} = \mathbf{s}\}\big|}{\big|\{i' \; : \; \mathrm{pa}_{i',j} = \mathbf{s}\}\big|} \;\; , \tag{9}$$

where $|S|$ denotes the cardinality of set S. However, as pointed out in Sec. 2, summing over all possible data matrices is not tractable except in toy problems where n and m are both very small. Efficient algorithms have been discovered only recently for restricted graph structures [17–19].

5 Factorized NML Models

As a computationally less demanding alternative to NML in the context of Bayesian networks, we define the *factorized NML* (fNML) in a similar spirit as sNML. We let the joint probability distribution be given by a product of *locally* normalized maximum likelihood distributions:

$$p_{\mathrm{fNML}}(x^n \; ; \; \mathcal{G}) := \prod_{j=1}^{m} \frac{p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{Z_j(\mathrm{Pa}_j)} \tag{10}$$

$$= \frac{\prod_{j=1}^{m} p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))}{Z(x^n)} \;\; , \tag{11}$$

where each of the local normalizing factors

$$Z_j(\mathrm{Pa}_j) = \sum_{X'_j} p(X'_j \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(X'_j, \mathrm{Pa}_j)) \tag{12}$$

is a sum over all possible instantiations of column $\mathbf{x}_{:,j}$, and the global normalizing factor

$$Z(x^n) = \prod_{j=1}^{m} \sum_{X'_j} p(X'_j \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(X'_j, \mathrm{Pa}_j)) \tag{13}$$

is a product of the local normalizing factors. The local normalizing factors $Z_j(\mathrm{Pa}_j)$ can be decomposed further into simple multinomial NML normalization constants, one for each parent configuration in Pa_j. Using the recently discovered linear-time algorithm [15] for the multinomial case, the total computation time becomes feasible even for large sample sizes and for many variables (columns).

In practice, we not only want to evaluate the likelihood of the data under a given model class, but we also wish to find the structure that maximizes the likelihood of the data. This is made hard by the fact that the number of possible DAG structure is superexponential. Unlike the standard NML criterion, the fNML criterion is 'modular' in the sense that it decomposes column-wise into independent terms. This enables the use dynamic programming techniques that find the global optimum in $o(n2^n)$ time, see [12], which is manageable for networks with up to about 30 nodes. For larger networks, local search heuristics are necessary.

Note that, as can be seen from (9), the maximum likelihood parameters of each local distribution, $\theta_{j|\mathrm{Pa}_j}$, depend only on column $\mathbf{x}_{:,j}$ and column(s) Pa_j. In particular, since we require $\mathrm{Pa}_j \subseteq \{1,\ldots,j-1\}$, we have

$$p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(x^n)) = p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(\mathbf{x}_{:,1},\ldots,\mathbf{x}_{:,j})) = p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(\mathbf{x}_{:,j},\mathrm{Pa}_j)) \;, \quad (14)$$

of which the second form, where only the first j columns appear, is the one that should be used in (10) by analogy with (2). Due to the above identity, the expressions are used interchangeably.

The sum-product view. It is interesting to compare the NML and fNML models. Consider Eqs. (7) and (11): the constant normalizer of NML, C_n, an exponential *sum of products*, is replaced in fNML by $Z(x^n)$, a *product of sums* that depends on the data. The fNML model can therefore be seen as 'cheating' by using a sum-product algorithm, where the distributive law (see [34])

$$\begin{cases} f(x_1,x_2) \equiv f(x_1) \\ g(x_1,x_2) \equiv g(x_2) \end{cases} \quad \Longrightarrow \quad \sum_{x_1,x_2} f(x_1,x_2)g(x_1,x_2) \;=\; \left(\sum_{x_1} f(x_1)\right)\left(\sum_{x_2} g(x_1)\right) \quad (15)$$

is applied to compute the sum in C_n even though the terms do not actually factor column-wise into independent parts. No cheating is necessary when the graph is empty, i.e., when $\mathrm{Pa}_j = \emptyset$ for all $1 \le j \le m$. This means that we have $Z(x^n) = C_n$, which by (7) and (11) implies that for empty graphs p_{NML} and p_{fNML} are equivalent.

The regrets of the two models are easily seen to be $\ln C_n$ and $\ln Z(x^n)$, for NML and fNML respectively. Notice also that the regret of fNML, $\ln Z(x^n)$, depends on the data only through the parents, $\mathrm{Pa}_j, 1 \le j \le m$, and hence, is independent of all the leaf nodes, i.e., nodes that have no descendants. Again, if the graph is empty, all nodes are leafs and $Z(x^n) = C_n$ for all x^n so that the NML and fNML models are equivalent.

Finally, we observe that for fNML the two variants of sNML, sNML-1 and sNML-2, coincide. Letting $x(j) := (\mathbf{x}_{:,1},\mathbf{x}_{:,2},\ldots,\mathbf{x}_{:,j})$ denote the first j columns, we obtain

$$\begin{aligned} p(x(j) \;;\; \hat{\theta}(x(j))) &= \prod_{l=1}^{j} p(\mathbf{x}_{:,l} \mid \mathrm{Pa}_l \;;\; \hat{\theta}(\mathbf{x}_{:,l},\mathrm{Pa}_l)) \\ &= p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \;;\; \hat{\theta}(x^n)) \prod_{l=1}^{j-1} p(\mathbf{x}_{:,l} \mid \mathrm{Pa}_l \;;\; \hat{\theta}(\mathbf{x}_{:,l},\mathrm{Pa}_l)) \;, \end{aligned}$$

where both equalities depend on (14). The last factor on the right-hand side is independent of column $\mathbf{x}_{:,j}$. When the above is normalized with respect to $\mathbf{x}_{:,j}$, this factor cancels and we are left with $p(\mathbf{x}_{:,j} \mid \mathrm{Pa}_j \; ; \; \hat{\theta}(x^n))$, which is exactly what is normalized in (10). Hence, it doesn't matter whether we define fNML as in (10) or as the product over $1 \leq j \leq m$ of the normalized versions of $p(x(j) \; ; \; \hat{\theta}(x(j)))$, and sNML-1 is equivalent to sNML-2 for Bayesian network model classes.

6 Experiments

To empirically test performance of the fNML-criterion in Bayesian network structure learning task, we generated several Bayesian networks, and then studied how different model selection criteria succeeded in learning the model structure from data. The most often used selection criterion for the task is the BDe (Bayesian Dirichlet equivalent) score [13], but due to its sensitivity to the choice of prior hyperparameter, we chose two different versions of it: $\mathrm{BDe}_{0.5}$ and $\mathrm{BDe}_{1.0}$. We also included the Bayesian Information Criterion, BIC. All these scores can be interpreted as implementing some version of the MDL criterion or an approximation thereof.

We present the results for an experiment in which we generated 1800 different Bayesian network models, which we tried to learn back using the data generated from these models. We generated the networks using 5, 10 and 15 variables, and also varied the density and the parameters of the networks. We then generated 1000, 10000 and 10000 data vectors from each network, and tried to learn the models back using these data samples and different scoring criteria. It turned out that learning the models back with these sample sizes was practically possible only for smallest networks containing 5 nodes. However, varying the number of arcs and parameters did not seem to have a strong effect on the outcome. This made it possible us to concentrate on comparing the performance of different scoring criteria for different sample sizes (Figure 6).

The results clearly show that fNML excels with small sample sizes. With large sample sizes, the difference is not that big, which is hardly surprising, since asymptotically, they all converge to the data generating model. This result is significant, since BDe score(s) can be regarded as the current state-of-the-art. Furthermore, the fNML score is computationally no more demanding than the BDe score.

Acknowledgment

This work was supported in part by the Finnish Funding Agency for Technology and Innovation under projects Kukot and PMMA, by the Academy of Finland under project Civi, and by the IST Programme of the European Community, under the Pascal Network of Excellence, IST-2002-506778. We thank the reviewers for useful comments.

References

[1] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri, "Comparing stochastic complexity minimization algorithms in estimating missing data," in *Proceedings of WUPES'97, the 4th Workshop on Uncertainty Processing*, Prague, Czech Republic, January 1997, pp. 81–90.

[2] ——, "On the accuracy of stochastic complexity approximations," in *Proceedings of the Causal Models and Statistical Learning Seminar*, London, UK, March 1997, pp. 103–117.

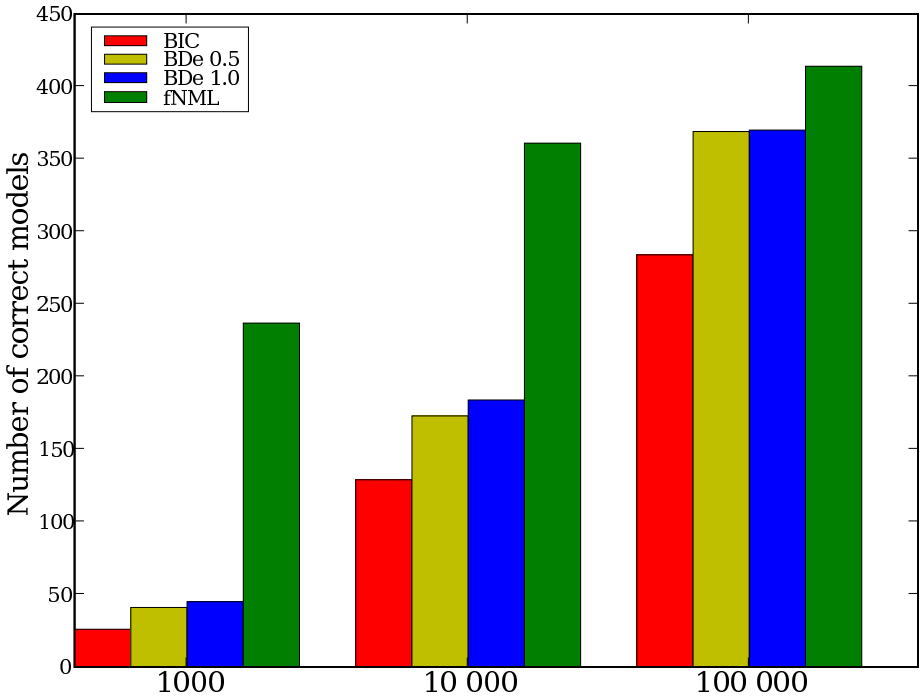

Figure 6: Number of correctly learned models in 1800 trials for sample sizes 1000, 10 000, and 100 000. For each sample size, the bars give the number of correctly learned models for (from left to right) the BIC, $\mathrm{BDe}_{0.5}$, $\mathrm{BDe}_{1.0}$, and fNML scores.

[3] P. Kontkanen, P. Myllymäki, and H. Tirri, "Comparing Bayesian model class selection criteria by discrete finite mixtures," in *Information, Statistics and Induction in Science*, D. Dowe, K. Korb, and J. Oliver, Eds. Proceedings of the ISIS'96 Conference, Melbourne, Australia: World Scientific, Singapore, August 1996, pp. 364–374.

[4] ——, "Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining," in *Proceedings of the Tenth International FLAIRS Conference*, D. Dankel, Ed., Daytona Beach, Florida, May 1997, pp. 204–211.

[5] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig, "Supervised posterior distributions," 2002, presented at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain.

[6] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri, "Minimum encoding approaches for predictive modeling," in *Proceedings of the 14th International Conference on Uncertainty in Artificial Intelligence (UAI'98)*, G. Cooper and S. Moral, Eds. Madison, WI: Morgan Kaufmann Publishers, San Francisco, CA, July 1998, pp. 183–192.

[7] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald, "Comparing predictive inference methods for discrete domains," in *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, Florida, January 1997, pp. 311–318.

[8] ——, "Bayesian and information-theoretic priors for Bayesian network parameters," in *Machine*

Learning: ECML-98, Proceedings of the 10th European Conference, ser. Lecture Notes in Artificial Intelligence, Vol. 1398, C. Nédellec and C. Rouveirol, Eds. Springer-Verlag, 1998, pp. 89–94.

[9] ——, "On the small sample behaviour of Bayesian and information-theoretic approaches to predictive inference," 1998, presented at the Sixth Valencia International Meeting on Bayesian Statistics, Alcossebre, Spain.

[10] ——, "On predictive distributions and Bayesian networks," *Statistics and Computing*, vol. 10, pp. 39–54, 2000.

[11] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, "On discriminative Bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, no. 3, pp. 267–296, 2005.

[12] T. Silander and P. Myllymäki, "A simple approach for finding the globally optimal Bayesian network structure," in *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, R. Dechter and T. Richardson, Eds. AUAI Press, 2006, pp. 445–452.

[13] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, no. 3, pp. 197–243, September 1995.

[14] T. Silander, P. Kontkanen, and P. Myllymäki, "On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter," in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, R. Parr and L. van der Gaag, Eds. AUAI Press, 2007, pp. 360–367.

[15] P. Kontkanen and P. Myllymäki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters*, vol. 103, no. 6, pp. 227–233, 2007.

[16] P. Kontkanen, P. Myllymäki, W. Buntine, J. Rissanen, and H. Tirri, "An MDL framework for data clustering," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I. Myung, and M. Pitt, Eds. The MIT Press, 2006.

[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] P. Kontkanen, H. Wettig, and P. Myllymäki, "NML computation algorithms for tree-structured multinomial Bayesian networks," *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.

[19] H. Wettig, 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.

[20] P. Kontkanen and P. Myllymäki, "MDL histogram density estimation," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, M. Meila and S. Shen, Eds., March 2007.

[21] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[22] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," 2006, submitted for publication. Preprint arXiv cs.IT/0609138.

[23] E. Takimoto and M. Warmuth, "The last-step minimax algorithm," in *Proc. 11th International Conference on Algorithmic Learning Theory*, 2000, pp. 279–290.

[24] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[25] Y. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, 1987.

[26] S. de Rooij and P. Grünwald, "An empirical study of minimum description length model selection with infinite parametric complexity," *Journal of Mathematical Psychology*, vol. 50, no. 2, pp. 180–192, 2006.

[27] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[28] J. Rissanen, *Information and Complexity in Statistical Modeling*. Springer, 2007.

[29] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.

[30] J. Rissanen and T. Roos, "Conditional NML models," in *Information Theory and Applications Workshop (ITA-07)*, San Diego, CA, January–February 2007.

[31] F. Liang and A. Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2708–2726, 2004.

[32] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[33] S. Lauritzen, *Graphical Models*. Oxford University Press, 1996.

[34] S. M. Aji and M. R. J., "The generalized distributive law," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 325–343, 2000.