Locally Minimax Optimal Predictive Modeling with Bayesian Networks

Tomi Silander
Helsinki Institute for Information Technology
Finland

Teemu Roos
Helsinki Institute for Information Technology
Finland

Petri Myllymäki
Helsinki Institute for Information Technology
Finland

Abstract

We propose an information-theoretic approach for predictive modeling with Bayesian networks. Our approach is based on the minimax optimal Normalized Maximum Likelihood (NML) distribution, motivated by the MDL principle. In particular, we present a parameter learning method which, together with a previously introduced NML-based model selection criterion, provides a way to construct highly predictive Bayesian network models from data. The method is parameter-free and robust, unlike the currently popular Bayesian marginal likelihood approach which has been shown to be sensitive to the choice of prior hyperparameters. Empirical tests show that the proposed method compares favorably with the Bayesian approach in predictive tasks.

1 INTRODUCTION

Bayesian networks (Pearl, 1988) are one of the most popular model classes for discrete vector-valued i.i.d. data. The popular Bayesian BDeu criterion (Heckerman, Geiger, & Chickering, 1995) for learning Bayesian network structures has recently been reported to be very sensitive to the choice of prior hyper-parameters (Silander, Kontkanen, & Myllymäki, 2007). On the other hand, the general model selection criteria, AIC (Akaike, 1973) and BIC (Schwarz, 1978), are derived through asymptotics and their behavior is suboptimal for small sample sizes. Furthermore, it is not clear how to set the parameters for the structures selected with AIC or BIC, so that the resulting model would yield a good predictive distribution.

Silander et al. (2008) have recently proposed a new effective scoring criterion for learning Bayesian network structures, the factorized normalized maximum likelihood (fNML). This score has no tunable parameters thus avoiding the mentioned sensitivity problems of Bayesian scores. However, the question of learning the parameters for the selected model structure has not been previously addressed. The currently popular choice within the Bayesian paradigm is to use the expected parameter values since this choice yields the same predictive distribution as model averaging, i.e., integrating the parameters out using the posterior distribution of the parameters. Free from sensitivity problem of the Bayesian model selection criterion, it would be very disappointing to adhere to a Bayesian way to learn the parameters of Bayesian networks — this would bring us back to the question of the choice of prior hyperparameters.

In this paper we propose a novel method for learning the parameters for Bayesian networks, based on the sequential normalized maximum likelihood (sNML) criterion. The combination of the fNML model selection criterion and the new sNML-based parameter learning method yields a complete *non*-Bayesian method for learning Bayesian networks. Computationally, the new method is as efficient as its Bayesian counterparts. It has already been shown to perform well in structure learning task, and in this paper we will show that it performs well in the predictive sense as well.

In the following, we will first introduce the notation needed for learning Bayesian networks. We will then briefly review the currently popular Bayesian scoring criterion and its sensitivity problem. To set the stage for our method for learning the network parameters, we will explain the fNML criterion for learning Bayesian network structures, after which we introduce our sNML-based methods for setting the parameters

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

for predictive purposes. The performance of the proposed method is be demonstrated by experiments.

2 BAYESIAN NETWORKS

A Bayesian network defines a joint probability distribution for an m-dimensional multivariate data vector $X = (X_1, \ldots, X_m)$, where each X_i may have r_i different values which, without loss of generality, can be denoted as $\{1, \ldots, r_i\}$.

2.1 MODEL CLASS

A Bayesian network consists of a directed acyclic graph (DAG) G and a set of conditional probability distributions. We specify the DAG with a vector $G = (G_1, \ldots, G_m)$ of parent sets so that $G_i \subset \{X_1, \ldots, X_m\}$ denotes the parents of variable X_i, i.e., the variables from which there is an arc to X_i. Each parent set G_i has q_i ($q_i = \prod_{X_p \in G_i} r_p$) possible values that are the possible value combinations of the variables belonging to G_i. We assume an enumeration of these values and denote the fact that G_i holds the j^{th} value combination simply by $G_i = j$.

The conditional probability distributions $P(X_i \mid G_i)$ are determined by a set of parameters, Θ, via the equation

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

We denote the set of parameters associated with variable X_i by Θ_i. Given a Bayesian network (G, Θ), the joint distribution can be factorized as

$$P(x \mid G, \Theta) = \prod_{i=1}^{m} P(x_i \mid G_i, \Theta_i). \tag{1}$$

This factorization induces a so called parental Markov condition: all the parametrizations of the structure G produce a joint distribution in which the variable X_i is independent of all its non-descendants given the values of its parents G_i. (Descendants of X_i are those variables that can be reached from node X_i in network G by following directed arcs.)

2.2 DATA

To learn Bayesian network structures, we assume a data set D with N i.i.d instantiations $(D^{(1)}, \ldots, D^{(N)})$ of the vector X, i.e., an $N \times m$ data matrix without missing values. We select columns of the data matrix D by subscripting it with a corresponding variable index or variable set; D_i, for instance, denotes the data corresponding to variable X_i.

Since the rows of D are assumed to be i.i.d, the probability of a data matrix can be calculated by just taking the product of the row probabilities. Combining equal terms yields

$$P(D \mid G, \Theta) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}, \tag{2}$$

where N_{ijk} denotes number of rows in which $X_i = k$ and its parents contain the j^{th} value combination.

For a given structure G, the *maximum likelihood* parameters $\widehat{\Theta}(D)$ are simply the relative frequencies found in data: $\hat{\theta}_{ijk} = N_{ijk} / \sum_{k'} N_{ijk'}$. Setting parameters $\hat{\theta}_{ijk}$ to their maximum likelihood values for data D gives the predictive distribution $P(x \mid G, \widehat{\Theta}(D))$. In the following, we denote the value $P(D \mid G, \widehat{\Theta}(D))$ by $\widehat{P}(D \mid G)$[1]. We also use a shorthand notation $\widehat{P}(A \mid B) = P(A \mid B, \widehat{\Theta}(A, B))$.

3 MODEL SELECTION

The number of possible Bayesian network models is super exponential with respect to the number of variables, and the model selection task has been shown to be NP-hard for practically all model selection criteria such as AIC, BIC and marginal likelihood (Chickering, 1996). However, all popular Bayesian network selection criteria $S(G, D)$ feature a convenient *decomposability* property,

$$S(G, D) = \sum_{i=1}^{m} S(D_i, D_{G_i}), \tag{3}$$

which makes implementing a heuristic search for models easier (Heckerman et al., 1995).

Many popular scoring functions avoid overfitting by balancing the fit to the data against the complexity of the model. A common form of this idea can be expressed as

$$S(G, D) = \log \widehat{P}(D \mid G) - \Delta(D, G), \tag{4}$$

where $\Delta(D, G)$ is a complexity penalty. For example, $\Delta^{\mathrm{BIC}} = \sum_i \frac{q_i(r_i - 1)}{2} \ln N$, and $\Delta^{\mathrm{AIC}} = \sum_i q_i(r_i - 1)$, where $q_i(r_i - 1)$ is the number of free parameters needed to specify the local distribution of variable X_i.

3.1 BAYESIAN DIRICHLET SCORES

The current state-of-the art is to use the marginal likelihood scoring criterion

$$S_{\bar{\alpha}}(D_i, D_{G_i}) = \log \int_{\theta_i} P(D_i \mid D_{G_i}, \theta_i) W(\theta_i \mid \alpha_i). \tag{5}$$

[1] We often drop the dependency on G when the dependency is clear from the context.

The most convenient form of this, the Bayesian Dirichlet (BD) score, uses conjugate priors in which the parameter vectors Θ_{ij} are assumed independent of each other and follow Dirichlet distributions so that

$$W(\theta_i \mid \alpha_i) = \prod_{j=1}^{q_i} P(\theta_{ij} \mid \alpha_{ij*}), \tag{6}$$

where $\theta_{ij} \sim Dir(\alpha_{ij1}, \ldots, \alpha_{ijr_i})$. With a choice of $\alpha_{ijk} = \alpha/(q_i r_i)$ we get a popular family of BDeu scores that give equal scores for different Bayesian network structures encoding same independence assumptions.

The BDeu score depends only on a single parameter α, but the outcome of model selection is very sensitive to it: it has been previously shown (Steck & Jaakkola, 2002) that the extreme values of α strongly affect the model selected by the BDeu score, and moreover, recent empirical studies have demonstrated great sensitivity to this parameter even within a completely normal range of values (Silander et al., 2007).

3.2 INFORMATION THEORY SCORES

Our preferred model selection criterion would be to use the normalized maximum likelihood (NML) distribution (Shtarkov, 1987; Rissanen, 1996):

$$P_{\mathrm{NML}}(D \mid \mathcal{M}) = \frac{\widehat{P}(D \mid \mathcal{M})}{\sum_{D'} \widehat{P}(D' \mid \mathcal{M})}, \tag{7}$$

where the normalization is over all data sets D' of a fixed size N. The log of the normalizing factor is called the *parametric complexity*. NML is the unique *minimax regret optimal* model, i.e., the unique minimizer in

$$\min_{Q} \max_{D'} \log \frac{\widehat{P}(D' \mid \mathcal{M})}{Q(D')},$$

where Q is allowed to be any distribution. The log-likelihood ratio is called the *regret* of distribution Q for data D' (wrt. model class $\mathcal{M}$). It can be interpreted as the excess logarithmic loss over the minimum loss achievable with model class $\mathcal{M}$.

However, there is no known method to compute the parametric complexity or the NML distribution for Bayesian networks efficiently (in less than exponential time). Therefore, in accordance with S_{BD} above, Silander et al. (2008) have proposed the following local score

$$\begin{aligned} S_{\mathrm{fNML}}(D_i, D_{G_i}) &= \log P_{\mathrm{NML}}(D_i \mid D_{G_i}) \\ &= \log \left(\frac{\widehat{P}(D_i \mid D_{G_i})}{\sum_{D_i'} \widehat{P}(D_i' \mid D_{G_i})} \right), \end{aligned} \tag{8}$$

where the normalizing sum goes over all the possible D_i-column vectors of length N, i.e., $D_i' \in \{1, \ldots, r_i\}^N$; the structure G is implicit in the notation. The *factorized NML* criterion is obtained as a sum of such local scores:

$$S_{\mathrm{fNML}}(G, D) = \sum_{i=1}^{m} S_{\mathrm{fNML}}(D_i, D_{G_i}). \tag{9}$$

Even though the normalizing sum in (8) is has an exponential number of terms, it can be evaluated efficiently using the recently discovered linear time algorithm for calculating the parametric complexity for a single r-ary multinomial variable (Kontkanen & Myllymäki, 2007).

It is immediate from the construction that fNML is decomposable. Thus it can be used efficiently in heuristic local search. Empirical tests show that selecting the network structure with fNML compares favourably to the state-of-the art model selection using BDeu scores even when the prior hyperparameter is optimized (with "hindsight") to maximize the performance (Silander, Roos, Kontkanen, & Myllymäki, 2008).

4 PREDICTION

The scoring methods described in the previous section can be used for selecting the best Bayesian network structure. However, much of the appeal of the Bayesian networks rests on the fact that *with the parameter values instantiated*, they define a joint probability distribution that can be used for probabilistic inference. For that reason, the structure selection is usually followed by a parameter learning phase. Next we will first review the standard Bayesian solution, and then in Section 4.2 introduce our new information-theoretic parameter learning scheme.

4.1 BAYESIAN PARAMETERS

In general, the Bayesian answer for learning the parameters amounts to inferring their posterior probability distribution. Consequently, the answer to determining the predictive probability

$$P(d \mid D, G) = \int P(d \mid \theta, G) P(\theta \mid D, G) d\theta$$

avoids selecting any particular parameter values. The actual calculation of the integral can be hard, but with the assumptions behind the BDeu score, the task becomes trivial since the predictive probability coincides with the joint probability of the data vector calculated using the expected parameter values

$$\tilde{\theta}_{ijk}^{BD} = \frac{N_{ijk} + \alpha_{ijk}}{\sum_{k'=1}^{r_i} [N_{ijk'} + \alpha_{ijk'}]}. \tag{10}$$

This choice of parameters can be further backed up by a prequential model selection principle (Dawid, 1984). Since the BDeu score is just a marginal likelihood $P(D \mid G, \alpha)$, it can be expressed as a product of predictive distributions

$$\begin{aligned} P(D \mid G, \alpha) &= \prod_{n=1}^{N} P(D^{(n)} \mid D^{(<n)}, \alpha) \\ &= \prod_{n=1}^{N} P(D^{(n)} \mid \tilde{\theta}(D^{(<n)}, \alpha)), \quad (11) \end{aligned}$$

where $D^{(<n)} = (D^{(1)}, \ldots, D^{(n-1)})$ denotes the first $n-1$ rows of D. Since we have selected the structure that has the strongest predictive record when using the expected parameter values, it is very natural to continue using the expected parameter values after the selection.

4.2 SEQUENTIAL NML PARAMETERS

Having proposed a non-Bayesian method for structure learning, it would be intellectually dissatisfactory to fall back to the Bayesian solution in the parameter learning task — in particular, as the Bayesian solution again depends on the hyperparameters. Hence, in accordance with the information-theoretic approach we introduce a solution to the parameter learning task based on minimax rules.

The so called *sequential NML* model (Rissanen & Roos, 2007; Roos & Rissanen, 2008) is similar in spirit to the factorized NML model in the sense that the idea is to obtain a joint likelihood as a product of locally minimax (regret) optimal models. In sNML, the normalization is done separately for each observation (vector) in a sequence.

$$P_{\mathrm{sNML}}(D) = \prod_{n=1}^{N} \frac{\widehat{P}(D^{(n)}, D^{(<n)})}{\sum_{d'} \widehat{P}(d', D^{(<n)})}. \quad (12)$$

One advantage of a row-by-row normalization is that it immediately leads to a natural prediction method: having seen a data-matrix of size $(N-1) \times m$, we can use the locally minimax optimal model for the N'th observation vector, obtained from (12), as a predictive distribution. The joint distribution (12) depends on the order of the data. However, we will later use it only for finding a parametrization that gives a good predictive distribution, and the $P_{\mathrm{sNML}}(d \mid D)$ does not depend on the order of data D.

That sNML defines a good predictive method can be demonstrated by showing that predicting with it never yields much worse a result than predicting the data while taking advantage of knowledge of the post-hoc optimal parameter value(s).

For a simple Bernoulli model, a result by Takimoto and Warmuth (2000) implies a neat bound on the regret of sNML.

Proposition 1 (Takimoto and Warmuth (2000)) *For the Bernoulli model, the worst-case regret $\bar{R}_{\mathrm{sNML}}(N, 2)$ over all binary sequence D of length N is upper-bounded by*

$$\begin{aligned} \bar{R}_{\mathrm{sNML}}(N, 2) &:= \max_{D} \left[\log P_{\mathrm{sNML}}(D) - \log P(D|\widehat{\Theta}(D)) \right] \\ &\leq \frac{1}{2} \log(N+1) + \frac{1}{2}. \end{aligned}$$

This is better than, for instance, what can be obtained by either the Laplace predictor, i.e., mixture with uniform prior, or the Krichevsky-Trofimov prediction, i.e., mixture with Dirichlet$(1/2, \ldots, 1/2)$ prior, see (Takimoto & Warmuth, 2000).

For a categorical datum with K different values, the following bound can be obtained.

Proposition 2 *For categorical (discrete) data, the worst-case regret of the sNML model is upper-bounded by*

$$\bar{R}_{\mathrm{sNML}}(N, K) \leq \frac{1}{K} \sum_{k=1}^{K-1} N \log \frac{N+k}{N} + k \log \frac{N+k}{k}.$$

We give an elementary proof of this statement in Appendix B. A relaxed version of the bound is as follows:

$$\bar{R}_{\mathrm{sNML}}(N, K) \leq (K-1) \left[\frac{K-1}{K} \log(\frac{N}{K-1} + 1) + \frac{1}{2} \right];$$

for $K = 2$, this agrees with the binary case above.

In theory, using sNML for determining a predictive distribution $P(d \mid D, G)$ would be straightforward. Furthermore, since the fNML was introduced as a computationally feasible version of the NML, we would still want to use a prediction scheme based on NML, thus the sNML would be a natural choice. In practice, however, using sNML for Bayesian networks faces two major problems. Firstly, it is not computationally feasible to calculate the normalizing term, since the number of possible values of a single data vector may be prohibitively large. Secondly, we set ourselves to learn the parameters for the selected Bayesian network, and it turns out that the predictive distribution $P_{\mathrm{sNML}}(d \mid D, G)$ cannot necessarily be obtained with any parametrization of the structure G (see Appendix A for an example). In the Bayesian case, the predictive probability can be obtained with the expected parameter values, but for NML we have no such luck.

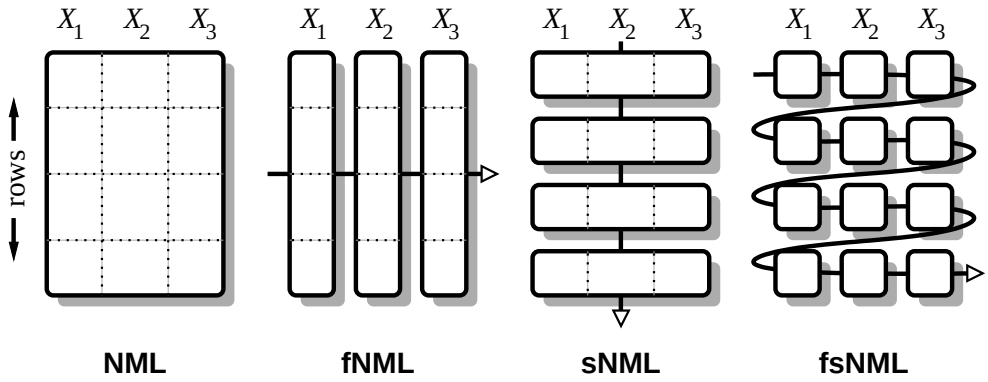

Figure 1: A schematic illustration of alternative ways to obtain minimax optimal models by normalizing the maximized likelihood $\widehat{P}(D \mid G)$. *Left to right:* In NML, the normalization is done over the whole data matrix in one go. In factorized NML (fNML), each column is normalized separately. In sequential NML (sNML), each row is normalized separately. In factorized-sequential NML (fsNML), the normalization is done entry-by-entry, in either the row or column order (the result is the same either way).

On the other hand, the Bayesian expected parameters can be interpreted as predictive probabilities for a one-dimensional categorical datum:

$$\theta_{ijk}^{BD} = P(d_i = k \mid D_i^{G_i=j}, G, \alpha_{ij}).$$

In analogy to this, we propose to use the corresponding sNML predictive probability distribution to set the parameters, i.e,

$$\theta_{ijk}^{\text{fsNML}} = P_{\text{sNML}}(d_i = k \mid D_i^{G_i=j}, G).$$

We call this approach *factorized sequential NML*. For categorical data this yields a spiced-up version of the Laplace's rule of succession

$$\theta_{ijk} = \frac{e(N_{ijk})(N_{ijk}+1)}{\sum_{k'=1}^{r_i} e(N_{ijk'})(N_{ijk'}+1)}, \qquad (13)$$

where $e(n) = (\frac{n+1}{n})^n$; $(e(0) = 1)$.

This selection of parameters also defines a joint probability distribution

$$P_{\text{fsNML}}(D \mid G) = \prod_{i=1}^{m} P_{\text{sNML}}(D_i | D_{G_i}) \qquad (14)$$
$$= \prod_{i=1}^{m} \prod_{n=1}^{N} \frac{\widehat{P}(D_i^{(n)}, D_i^{(<n)} \mid D_{G_i}^{(\leq n)})}{\sum_{k=1}^{r_i} \widehat{P}(k, D_i^{(<n)} \mid D_{G_i}^{(\leq n)})}.$$

Comparing P_{fsNML} (14) with the equations (9) and (8) for the logarithmic version of the $P_{\text{fNML}}(D \mid G)$ reveals their similar spirit. In contrast with NML, where normalization is done over the whole data matrix in a single, huge summation, or sNML, where normalization is done over data vectors of length m, the normalization in fsNML is very simple since it only involves a single entry at a time (see Figure 1).

Proposition 3 *Given a Bayesian network structure G, the regret of the fsNML distribution is upper-bounded by*

$$\bar{R}_{\text{fsNML}}(N, G) \leq \sum_{i=1}^{m} q_i \bar{R}_{\text{sNML}}\left(N/q_i, r_i\right),$$

where q_i and r_i denote the number of parent configurations and the arity of variable X_i, respectively, and $\bar{R}_{\text{sNML}}(N/q_i, r_i)$ is the univariate bound given in Prop. 2.

The proof is rather straightforward by using the decomposition of the likelihood function according to the network structure, and the concavity of the regret functions with respect to the counts N_{ij}. We omit the details.

5 EXPERIMENTS

To empirically test our method, we selected 20 UCI data sets[2] with fewer than 20 variables, so that we can use exact structure learning algorithms (Silander & Myllymäki, 2006) that eliminate the uncertainty due to the heuristic search for the best structure. We then compared our method, the fNML-based structure learning + fsNML parametrization, with the state-of-the-art Bayesian method, the BDeu score with expected parameters. The equivalent sample size hyperparameter α for the Bayesian learning was set to 1.0, a common and convenient "non-informative" choice.

The comparison was done by creating 100 random train and test splits (50%–50%) of each data set,

[2]Continuous values in these data sets were discretized into three equal width bins.

and then using each training data set for learning two Bayesian networks, one with each method. The Bayesian networks were then used to determine the predictive probability $P(d \mid G, \Theta)$ for each vector in the test data.

Table 1: Summary of the prediction experiment.

Data	N	m	#vals	$\frac{P_{\mathrm{fNML}}(d\mid D)}{P_{\mathrm{BDeu}}(d\mid D)}$
iris	150	5	3.0	0.968
thyroid	215	6	3.0	0.996
shuttle	58000	10	3.0	0.998
page blocks	5473	11	3.2	1.001
yeast	1484	9	3.7	1.027
abalone	4177	9	3.0	1.029
liver	345	7	2.9	1.050
diabetes	768	9	2.9	1.070
adult	32561	15	7.9	1.088
ecoli	336	8	3.4	1.094
balance	625	5	4.6	1.100
glass	214	11	3.3	1.139
tic tac toe	958	10	2.9	1.246
heart hungarian	294	14	2.6	1.316
breast cancer	286	10	4.3	1.519
bc wisconsin	699	11	2.9	1.550
wine	178	14	3.0	1.665
heart statlog	270	14	2.9	1.888
heart cleveland	303	14	3.1	2.587
post operative	90	9	2.9	2.621

The results of the predictive experiment are presented in Table 1. For each data set, the table lists the number of data vectors N, the number of variables m, the average number of values per variable (#vals), and the ratio of average predictive probabilities obtained with our method and the Bayesian method. In 17 data sets (out of 20) the NML-based method predicted better, and never did it predict significantly worse. From the graphical illustration in Figure 2 we can see that the difference in favor of the NML-based approach is especially large with the more difficult data sets where the predictive probabilities are small.

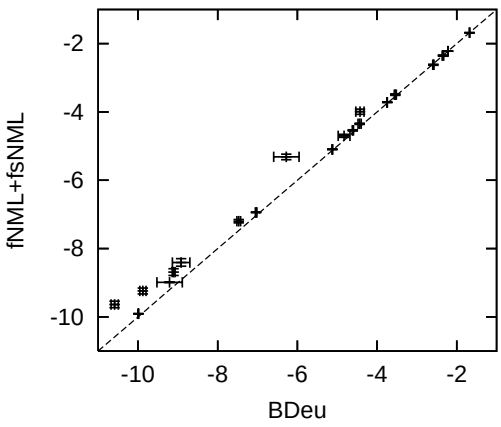

Figure 2: Visualization of the results summarized in Table 1. Each point gives the predictive accuracies obtained with the two methods, in terms of average log-likelihood per data vector (greater values are better). Error-bars show $\pm 1.96\times$ standard deviation over 100 random train-test splits. Point above the diagonal line represent cases where the fNML+fsNML method performs better than the Bayesian approach.

6 CONCLUSION

We have presented a sequential NML-based method for learning Bayesian network parameters. Combined with the previously presented NML-based structure learning method, this work provides a parameter-free non-Bayesian way to automatically construct Bayesian networks from the data. Empirical tests show the feasibility of the proposed method.

Plenty of questions remain. Both the structure learning and the parameter learning may be seen as practical approximations to the theoretically more desirable methods of NML and sNML. While the empirical tests demonstrate the computational efficiency and good performance of the proposed method, theoretical results about the accuracy of these approximations would definitely be welcome.

The parameter learning procedure proposed in this paper can be used regardless of the structure selection method. While we have here only compared the MDL-based and Bayesian approaches, it would be possible to also study mixed methods like the combination of BDeu or BIC model selection and fsNML parametrization. While one can be sceptical about the performance of such mixes, these kind of additional experiments could clarify the role the different methods of structure and parameter learning play in predictive performance.

Acknowledgements

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

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrox & F. Caski (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Chickering, D. (1996). Learning Bayesian networks is NP-Complete. In D. Fisher & H. Lenz (Eds.), *Learning from data: Artificial intelligence and statistics v* (pp. 121–130). Springer-Verlag.

Dawid, A. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Kontkanen, P., & Myllymäki, P. (2007). A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, *103*(6), 227–233.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann Publishers, San Mateo, CA.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Rissanen, J., & Roos, T. (2007). Conditional NML models. In *Proceedings of the information theory and applications workshop (ITA-07).* San Diego, CA.

Roos, T., & Rissanen, J. (2008). On sequentially normalized maximum likelihood models. In *Workshop on information theoretic methods in science and engineering (WITMSE-08).* Tampere, Finland.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shtarkov, Y. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, *23*, 3–17.

Silander, T., Kontkanen, P., & Myllymäki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In R. Parr & L. van der Gaag (Eds.), *Proceedings of the 23rd conference on uncertainty in artificial intelligence* (pp. 360–367). AUAI Press.

Silander, T., & Myllymäki, P. (2006). A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter & T. Richardson (Eds.), *Proceedings of the 22nd conference on uncertainty in artificial intelligence* (pp. 445–452). AUAI Press.

Silander, T., Roos, T., Kontkanen, P., & Myllymäki, P. (2008). Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In *Proceedings of the 4th European workshop on probabilistic graphical models (PGM-08)* (pp. 257–264). Hirtshals, Denmark.

Steck, H., & Jaakkola, T. S. (2002). On the Dirichlet prior and Bayesian regularization. In *Advances in neural information processing systems 15* (pp. 697–704). Vancouver, Canada: MIT Press.

Takimoto, E., & Warmuth, M. (2000). The last-step minimax algorithm. In *Proceedings of the 11th international conference on algorithmic learning theory* (pp. 279–290).

Appendix A

The following example shows that the joint probability distribution $P_{\mathrm{sNML}}(d \mid D, G)$ cannot necessarily be presented with any parametrization of the network G.

Let G be a simple v-structure $G = (\{\}, \{X_1, X_3\}, \{\})$, and let the data D consist of just a single 3-dimensional binary-vector $[(0, 0, 0)]$.

A direct calculation of

$$P_{\mathrm{sNML}}(d \mid D, G) = \frac{P(d, D \mid G, \hat{\theta}(D, d))}{\sum_{d'} P(d', D \mid G, \hat{\theta}(D, d'))}$$

yields a probability distribution

$P(d \mid D)$	$\frac{8}{19}$	$\frac{2}{19}$	$\frac{2}{19}$	$\frac{2}{19}$	$\frac{2}{19}$	$\frac{1}{38}$	$\frac{2}{19}$	$\frac{1}{38}$
d	000	001	010	011	100	101	110	111

.

In this joint probability distribution X_1 and X_3 are not marginally independent, i.e. $P(X_1, X_3) \neq P(X_1)P(X_3)$. However, all the parametrizations of the structure v-structure G yield distributions where X_1 and X_3 are independent.

After marginalizing out X_2, we get $P(X_1, X_3 \mid D, G)$

$P(x_1, x_3 \mid D)$	$\frac{10}{19}$	$\frac{4}{19}$	$\frac{4}{19}$	$\frac{1}{19}$
x_1x_3	00	01	10	11

.

However, the product of marginals $P(X_1) = (\frac{14}{19}, \frac{5}{19})$ and $P(X_3) = (\frac{14}{19}, \frac{5}{19})$ yields a different distribution

$P(x_1 \mid D)P(x_3 \mid D)$	$\frac{196}{361}$	$\frac{70}{361}$	$\frac{70}{361}$	$\frac{25}{361}$
x_1x_3	00	01	10	11

.

Appendix B: Proof of Proposition 2

We derive a regret bound for the categorical data of size N with K categories. We start by reviewing the probability distribution of interest

$$P_{\mathrm{sNML}}(D) = \prod_{n=1}^{N} \frac{\widehat{P}(D^{(n)}, D^{(<n)})}{\sum_{d'} \widehat{P}(d', D^{(<n)})},$$

where we have denoted with $D^{(<n)}$ the first $n-1$ data items of the sequence D, and with $\widehat{P}(X)$ the maximum likelihood of the data X, $\widehat{P}(X) = P(X|\widehat{\theta}(X))$. We denote with $k_{<n}$ the number of times the value k appears in $D^{(<n)}$.

To anticipate the comparison of the P_{sNML} with the $\widehat{P}$, we write the $\widehat{P}$ in the form

$$\widehat{P}(D) = \prod_{n=1}^{N} \frac{\widehat{P}(D^{(n)}, D^{(<n)})}{\widehat{P}(D^{(<n)})}.$$

Now we compare the ratio

$$\begin{aligned}
Q(D) &= \frac{\widehat{P}(D)}{P_{\mathrm{sNML}}(D)} \\
&= \prod_{n=1}^{N} \frac{\widehat{P}(D^{(n)}, D^{(<n)}) \sum_{d'} \widehat{P}(d', D^{(<n)})}{\widehat{P}(D^{(<n)}) \widehat{P}(D^{(n)}, D^{(<n)})} \\
&= \prod_{n=1}^{N} \frac{\sum_{d'} \widehat{P}(d', D^{(<n)})}{\widehat{P}(D^{(<n)})} \\
&= \prod_{n=1}^{N} \frac{\sum_{d'} \prod_{k=1}^{K} (\frac{k_{<n}+[d'=k]}{n})^{(k_{<n}+[d'=k])}}{\prod_{k=1}^{K} (\frac{k_{<n}}{n-1})^{k_{<n}}} \\
&= \prod_{n=1}^{N} \frac{\frac{1}{n^n} \sum_{d'} \prod_{k=1}^{K} (k_{<n} + [d'=k])^{(k_{<n}+[d'=k])}}{\frac{1}{(n-1)^{n-1}} \prod_{k=1}^{K} k_{<n}^{k_{<n}}} \\
&= \prod_{n=1}^{N} \frac{(n-1)^{n-1}}{n^n} \sum_{k=1}^{K} \frac{(k_{<n}+1)^{k_{<n}+1}}{{k_{<n}}^{k_{<n}}} \\
&= \prod_{n=1}^{N} \frac{(n-1)^{n-1}}{n^n} \sum_{k=1}^{K} (k_{<n}+1) e(k_{<n}),
\end{aligned}$$

where we have used the function $e(x) = (\frac{x+1}{x})^x$ that approaches the real number e from below ($e(0) = 1$) when x grows. The sum within the product obtains it largest when all the $k_{<n}$ are equal. Therefore we can bound the ratio by

$$\begin{aligned}
Q(D) &\leq \prod_{n=1}^{N} \frac{(n-1)^{n-1}}{n^n} \sum_{k=1}^{K} (\frac{n-1}{K} + 1) e(\frac{n-1}{K}) \\
&= \prod_{n=1}^{N} \frac{(n-1)^{n-1}}{n^n} (n+K-1) e(\frac{n-1}{K}) \\
&= \prod_{n=1}^{N} \frac{(n-1)^{n-1}}{n^n} (n+K-1) (\frac{n+K-1}{n-1})^{\frac{n-1}{K}} \\
&= \prod_{n=1}^{N} \frac{(n-1)^{(\frac{K-1}{K})(n-1)} (n+K-1)^{\frac{n+K-1}{K}}}{n^n} \\
&= \prod_{n=1}^{N} \frac{(n-1)^{(\frac{K-1}{K})(n-1)}}{n^{\frac{K-1}{K}n}} \frac{(n+K-1)^{\frac{n+K-1}{K}}}{n^{\frac{n}{K}}} \\
&= \frac{1}{N^{N\frac{K-1}{K}}} \frac{\prod_{k=1}^{K-1} (N+k)^{\frac{N+k}{K}}}{\prod_{k=1}^{K-1} k^{\frac{k}{K}}} \\
&= \prod_{k=1}^{K-1} (\frac{N+k}{N})^{\frac{N}{K}} (\frac{N+k}{k})^{\frac{k}{K}}.
\end{aligned}$$

By taking the logarithm we get a bound for the regret

$$\begin{aligned}
R(N,K) &= \max_{D} \ln(Q(D)) \\
&\leq \frac{1}{K} \sum_{k=1}^{K-1} \left[\ln(\frac{N+k}{N})^N + \ln(\frac{N+k}{k})^k \right].
\end{aligned}$$

This concludes the proof.

By noticing that $(\frac{N+k}{N})^N \leq e^k$ and that $(\frac{N+k}{k})^k \leq (\frac{N+K-1}{K-1})^{K-1}$ we get the relaxed version

$$R(N,K) \leq (K-1) \left[\frac{K-1}{K} \ln\left(\frac{N}{K-1} + 1 \right) + \frac{1}{2} \right].$$