Bayesian and Information-Theoretic Priors for Bayesian Network Parameters

Petri Kontkanen[1], Petri Myllymäki[1], Tomi Silander[1], Henry Tirri[1], and
Peter Grünwald[2]

[1] Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland
[2] CWI, Department of Algorithms and Architectures
P.O. Box 94079, NL-1090 GB Amsterdam, The Netherlands

Abstract. We consider Bayesian and information-theoretic approaches for determining non-informative prior distributions in a parametric model family. The information-theoretic approaches are based on the recently modified definition of *stochastic complexity* by Rissanen, and on the Minimum Message Length (MML) approach by Wallace. The Bayesian alternatives include the uniform prior, and the equivalent sample size priors. In order to be able to empirically compare the different approaches in practice, the methods are instantiated for a model family of practical importance, the family of Bayesian networks.

1 Introduction

Given some sample data, our goal is to learn about the regularities in the problem domain so that we can arrive at a "good" predictive distribution $\mathcal{P}$ that can be used to predict well. In the following we restrict the search for such a $\mathcal{P}$ to a class $\mathcal{M}$ of probabilistic models, which all share the same parametric form. All the approaches considered here depend on a prior distribution $P(\Theta)$ over all the models (parameter instantiations) Θ in the class $\mathcal{M}$. In this paper we study different alternatives for choosing $P(\Theta)$ in an *non-informative setting*, where no "data independent" prior knowledge about the problem domain is available.

The statistical literature contains many proposals for "optimal" non-informative prior distributions. While all of these satisfy some optimality criterion, in practice they give rise to different predictions. The main purpose of this paper is to compute several different "optimal" prior distributions P for a model class of practical importance, the class of Bayesian networks (see, e.g., [5]). In particular, we will compare priors which are in accordance with the Bayesian interpretation of probability, to priors motivated by information-theoretic considerations: a prior based on Rissanen's *Minimum Description Length (MDL)* principle [10], and a prior based on Wallace & Boulton's *Minimum Message Length (MML)* principle [15]. Though MDL and MML are similar in spirit, we will see that they do not lead to the same prior distribution.

In Section 2 we introduce the general setting of the problem by discussing and motivating the priors we will use. In Section 3 the priors and predictive distributions are instantiated for the special case where the models are defined by a Bayesian network structure with a particular arbitrary, but fixed, topology. For some of the priors, this instantiation has been presented in [4, 6]. The contribution of this section is to derive explicit formulas for the MDL and MML priors for the case of Bayesian networks, which involves computing the (expected) Fisher information matrix for Bayesian networks. For comparing the predictive distributions presented in this paper, we have run an extensive series of tests on real world data, but due to space constrains the results of the tests are presented elsewhere [8].

2 Predictive Distributions and Their Priors

We model the problem domain by a set X of m discrete random variables, $\mathrm{X} = \{X_1, \ldots, X_m\}$, where a random variable X_i can take on any of the values in the set $\mathbf{X}_i = \{x_{i1}, \ldots, x_{in_i}\}$. A *data instantiation* $\boldsymbol{d} = (x_1, \ldots, x_m)$ is a vector in which all the variables X_i have been assigned a value: by $\boldsymbol{X} = \boldsymbol{d}$ we mean that $X_1 = x_1, \ldots, X_m = x_m$, where $x_i \in \mathbf{X}_i$. A *random sample* $D = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N)$ is a set of N i.i.d. (independent and identically distributed) data instantiations, where each $\boldsymbol{d}_j$ is assumed to be sampled from the joint distribution of the variables in X.

Given a random sample D, we are interested in the question of how to define the *predictive distribution* $\mathcal{P}(\boldsymbol{d}|D)$ for a given vector $\boldsymbol{d}$. We investigate several candidates for $\mathcal{P}(\boldsymbol{d}|D)$, relative to a parametric family $\mathcal{M}$ of probabilistic models: each model $\Theta \in \mathcal{M}$ defines a probability $P(\boldsymbol{d}|\Theta)$ for each data instantiation $\boldsymbol{d}$, and, under the i.i.d. assumption, a probability $P(D|\Theta)$ (the *likelihood*) for each dataset D. Given the likelihood, and a prior distribution $P(\Theta)$ for all $\Theta \in \mathcal{M}$, we can arrive at a posterior distribution for the models:

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta). \tag{1}$$

The *MAP (maximum a posteriori probability)* predictive distribution is given by

$$\mathcal{P}_{\mathrm{MAP}}(\boldsymbol{d} \mid D, \Phi) = P(\boldsymbol{d} \mid D, \hat{\Theta}_{\Phi}(D)) \overset{\text{i.i.d.}}{=} P(\boldsymbol{d} \mid \hat{\Theta}_{\Phi}(D)), \tag{2}$$

where Φ denotes the (hyper)parameters used for defining the prior distribution $P(\Theta)$, and $\hat{\Theta}(D)$ is the MAP model maximizing the posterior (1).

A more sophisticated approach is to average (integrate) over all the models $\Theta \in \mathcal{M}$, which produces the *evidence* or *marginal likelihood* predictive distribution

$$\mathcal{P}_{\mathrm{EV}}(\boldsymbol{d}|D, \Phi) = \int P(\boldsymbol{d}|D, \Theta, \Phi)P(\Theta|D, \Phi)d\Theta \overset{\text{i.i.d.}}{=} \int P(\boldsymbol{d}|\Theta)P(\Theta|D, \Phi)d\Theta. \tag{3}$$

Both the MAP predictive distribution and the evidence predictive distribution are defined by using the posterior $P(\Theta|D)$, which depends on the prior $P(\Theta)$. We now consider different alternatives for determining the prior distribution.

The Uniform Prior The conceptually simplest non-informative prior is the *uniform prior*, in which case the prior distribution $P(\Theta)$ is a constant. One can see from (1) that in this case the MAP predictive distribution becomes the *Maximum Likelihood (ML) model* of classical statistics, i.e., the model $\tilde{\Theta}$ maximizing the data likelihood $P(D|\Theta)$.

Equivalent Sample Size Priors In the Bayesian philosophy the prior probability of a model Θ can be regarded as a prior (initial) *degree of belief* in the model Θ. Given a sufficiently regular model class $\mathcal{M}$, we can construct *Equivalent Sample Size (ESS)* priors Φ for $\mathcal{M}$ so that the following property holds for all $\boldsymbol{d}$ and all D of any size:

$$P(\boldsymbol{d} \mid D, \Phi) = P(\boldsymbol{d} \mid \tilde{\Theta}(D \cup D')), \tag{4}$$

where $\tilde{\Theta}$ is the maximum likelihood model (see above) for the training data D plus some additional "virtual data" D'. This virtual data D' depends only on the prior Φ, i.e., for each dataset D' of any size there is exactly one prior $\Phi_{D'}$ that corresponds to it. Hence using P in combination with Φ is always equivalent to predicting using the model that renders the training data D plus the extra data D' in the most likely manner. We can now interpret D' as *a priori data* that governs how strongly we let our predictions be influenced by the actual sample D (see [5]).

An MDL Prior Intuitively speaking, the *Minimum Description Length (MDL) Principle* [10–12] states that the more we are able to compress a set of data, the more we have learned about it, and the better we will be able to predict future data. *Stochastic complexity* of a data set D relative to a class of models $\mathcal{M}$ is defined as the code length of D when it is encoded using the shortest code obtainable with the help of the class $\mathcal{M}$. Here by the "shortest code" one means the code that gives as short as possible a code length to *all* possible data sets D. It follows from the Kraft Inequality (see for example [11]) that the stochastic complexity S can be written as $S = -\log \mathcal{P}_{\mathrm{SC}}$ where $\mathcal{P}_{\mathrm{SC}}$ is some probability distribution.

There are several reasons why $S = -\log \mathcal{P}_{\mathrm{EV}}(D) = \int P(D|\Theta)\pi(\Theta)d\Theta$ is a good candidate for defining the stochastic complexity explicitly [11]. Recently, however, Rissanen [12] has shown that there exists a code that is itself not dependent on any prior distribution of parameters, and which yields even shorter codelengths than the code with lengths $-\log \mathcal{P}_{\mathrm{EV}}(D)$, except for the special case where a particular prior $\pi(\Theta) \propto |I(\Theta)|^{1/2}$, the so-called *Jeffrey's prior* [2, 3], is used for $\mathcal{P}_{\mathrm{EV}}(D)$. Here $|I(\Theta)|$ denotes the determinant of the *Fisher information matrix* $I(\Theta)$ as defined in [2]. In this case it can be shown [12] that under suitable technical conditions, $\mathcal{P}_{\mathrm{EV}}$ and $\mathcal{P}_{\mathrm{SC}}$ asymptotically coincide:

$$-\log \mathcal{P}_{\mathrm{SC}}(D) = -\log \mathcal{P}_{\mathrm{EV}}(D) + o(1), \tag{5}$$

which means that from the MDL point of view, the optimal predictive distribution is obtained by using $\mathcal{P}_{\mathrm{EV}}$ with Jeffrey's prior.

An MML Prior *Minimum Message Length (MML) Inference* [14, 15] is based on a similar philosophy to the *MDL* principle, but there are also some subtle differences which cause the actual formulas used in MDL and MML estimation to differ considerably (see [1] for a detailed discussion on this subject). For our purposes, it is sufficient to note the following two differences: first, in MML modeling the predictive distribution $P(\boldsymbol{d}|\Theta_{\mathrm{MML}}(D))$ is defined by using a single "MML-optimal" model, whereas $\mathcal{P}_{\mathrm{SC}}$ as defined above uses an integral over all the models in the given class. Second, although both employ priors, the priors used in MDL serve only as a technical tool for computing an approximation to $\mathcal{P}_{\mathrm{SC}}$ (which itself does not depend on any prior), while MML adopts a Bayesian philosophy regarding priors, and assumes the user to provide a *subjective prior* $P(\Theta)$ to reflect his/her prior beliefs. Omitting all mathematical details (which can be found in [15]), the MML-optimal model $\Theta_{\mathrm{MML}}(D)$ is defined by

$$\Theta_{\mathrm{MML}}(D) = \arg\max_{\Theta\in\mathcal{M}} \frac{P(D|\Theta)P(\Theta)}{|I(\Theta)|^{1/2}}, \tag{6}$$

where $|I(\Theta)|$ is the determinant of the Fisher information matrix. We now see that $\Theta_{\mathrm{MML}}(\mathrm{D})$ for prior $P(\Theta)$ is equal to the MAP-model $\hat{\Theta}(\mathrm{D})$ for prior $P'(\Theta) \propto P(\Theta)/\pi(\Theta)$. Interestingly, while the formula for the MDL predictive distribution involves *multiplying* $P(D|\Theta)$ by Jeffrey's prior, the formula for the MML predictive distribution involves *dividing* $P(D|\Theta)$ by Jeffrey's prior.

3 Priors for Bayesian Networks

A Bayesian (belief) network [9, 13] is a representation of a probability distribution over a set of discrete variables, consisting of an acyclic directed graph, where the nodes correspond to domain variables $X_1, \dots, X_m$. Each network topology defines a set of independence assumptions which allow the joint probability distribution for a data vector $\boldsymbol{d}$ to be written as a product of simple conditional probabilities,

$$P(\boldsymbol{d}) = P(X_1 = x_1, \dots, X_m = x_m) = \prod_{i=1}^{m} P(X_i = x_i|\mathrm{pa}_i = q_i), \tag{7}$$

where q_i denotes a configuration of (the values of) the parents of variable X_i. Consequently, in the Bayesian network model family, a distribution $P(\boldsymbol{d} \mid \Theta)$ is uniquely determined by fixing the values of the parameters $\Theta = (\theta^1, \dots, \theta^m)$, where $\theta^i = (\theta^i_{11}, \dots, \theta^i_{1n_i}, \dots, \theta^i_{c_i1}, \dots, \theta^i_{c_in_i})$, n_i is the number of values of X_i, c_i is the number of configurations of pa_i, and $\theta^i_{q_ix_i} := P(X_i = x_i \mid \mathrm{pa}_i = q_i)$.

In the following all the conditional distributions of the variables, given their parents, are assumed to be multinomial, i.e., $X_{i|q_i} \sim \mathrm{Multi}(1; \theta^i_{q_i1}, \dots, \theta^i_{q_in_i})$. Since the family of Dirichlet distributions is *conjugate* (see e.g. [2]) to the family of multinomials, it would be convenient if we could assume that the prior distributions of the parameters are from this family. More precisely, this would mean that $(\theta^i_{q_i1}, \dots, \theta^i_{q_in_i}) \sim \mathrm{Di}(\mu^i_{q_i1}, \dots, \mu^i_{q_in_i})$, where $(\mu^i_{q_i1}, \dots, \mu^i_{q_in_i})$ are the hyperparameters of the corresponding distributions. From the definition of Dirichlet distributions [2], it is relatively easy to see that both the uniform prior and the ESS priors are Dirichlet distributions (see, e.g., [6]). For the subclass of Bayesian Networks used in our experiments reported in [8], Jeffrey's prior is of the Dirichlet form too. Moreover, we have seen in the previous section that the priors we need for the MML predictive distributions are arrived at by dividing the user's subjective prior by Jeffrey's prior. If the subjective prior is Dirichlet, it is easy to see that the resulting prior is of the Dirichlet form too. Consequently, all the priors used here are Dirichlet, which allows us to derive explicit expressions for $\mathcal{P}_{\mathrm{MAP}}$ and $\mathcal{P}_{\mathrm{EV}}$, as shown in [4,6]. The computation of the ESS priors for Bayesian Networks can be found in [5]. In the following we show how to compute the Jeffrey's prior $\pi(\Theta)$ for Bayesian networks, which is required for determining the MDL and MML priors discussed above.

Let $I(\cdot)$ denote the indicator function, i.e., $I(a, b) = 1$ if $a = b$ and 0 otherwise. We write d_{ji} for the i-th entry of data instantiation $\boldsymbol{d}_j$; q_{ji} stands for the configuration of the parent variables of X_i in $\boldsymbol{d}_j$. For computing the Fisher information matrix $I(\Theta)$, let us consider the element $[I(\Theta)]_{r,s}$ where (r, s) is the entry corresponding to $\theta^{i_1}_{q_{i_1}l_1}$ and $\theta^{i_2}_{q_{i_2}l_2}$. By deriving an explicit expression for $\log P(\boldsymbol{X}|\Theta)$ one can show that if either the variable indices i_1, i_2 or the parent configurations q_{i_1}, q_{i_2} are different, then $[I(\Theta)]_{r,s} = 0$. If $q_{i_1} = q_{i_2}$ and $i_1 = i_2$, one obtains after some calculations:

$$[I(\Theta)]_{r,s} = \mathrm{E}_\Theta \left[\frac{-\partial^2 \log P(\boldsymbol{d}_j|\Theta)}{\partial(\theta^i_{q_il})^2} \right] = \frac{P(\mathrm{pa}_i = q_i|\Theta)}{\theta^i_{q_in_i}} + I(l_1, l_2) \frac{P(\mathrm{pa}_i = q_i|\Theta)}{\theta^i_{q_il_1}}.$$

We now have an expression for each element of $I(\Theta)$, which gives us

$$\begin{aligned} \pi(\Theta) \propto \sqrt{|I(\Theta)|} &= \prod_{i=1}^{m} \prod_{q_i=1}^{c_i} (N \cdot P^i_{q_i})^{\frac{n_i-1}{2}} \prod_{l=1}^{n_i} (\theta^i_{q_il})^{-\frac{1}{2}} \\ &\propto \prod_{i=1}^{m} \prod_{q_i=1}^{c_i} (P^i_{q_i})^{\frac{n_i-1}{2}} \prod_{l=1}^{n_i} (\theta^i_{q_il})^{-\frac{1}{2}}. \end{aligned}$$

Details of the derivation of this result can be found in [7].

4 Conclusion

In this paper we have discussed various Bayesian and information-theoretic approaches for determining non-informative prior distributions in a parametric model family: Minimum Description Length (MDL) prior, Minimum Message Length (MML) prior, uniform prior, and equivalent sample size priors. To be able to study the relevance of the various approaches in practice, we instantiated the methods for the family of Bayesian networks. Our empirical results reported in [8] show that while in the case of large training samples all methods give very good results, some of them perform close to optimal already when only a very small amount of training data is available. The results suggest that if the size of the training data is small, it would be a good idea to use either the evidence-based approach (with any prior), or the MAP approach with the ESS priors.

Acknowledgments This research has been supported by the ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), the Technology Development Center (TEKES), and the Academy of Finland.

References

1. R.A. Baxter and J.O. Oliver. MDL and MML: Similarities and differences. Technical Report 207, Department of Computer Science, Monash University, 1994.
2. J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.
3. B.S. Clarke and A.R. Barron. Jeffrey's prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.
4. G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
5. D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, One Microsoft Way, Redmond, WA 98052, 1996.
6. D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, September 1995.
7. P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. Technical Report NC-TR-97-032, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1997.
8. P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. A comparison of non-informative priors for Bayesian networks. Technical Report NC-TR-98-002, ESPRIT Working Group on Neural and Computational Learning (NeuroCOLT), 1998.
9. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, San Mateo, CA, 1988.
10. J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239 and 252–265, 1987.
11. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, New Jersey, 1989.
12. J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.
13. R.D. Shachter. Probabilistic inference and influence diagrams. *Operations Research*, 36(4):589–604, July-August 1988.
14. C.S. Wallace and D.M Boulton. An information measure for classifiation. *Computer Journal*, 11:185–194, 1968.
15. C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49(3):240–265, 1987.