A comparison of non-informative priors for Bayesian networks

Petri Kontkanen, Petri Myllymäki, Tomi Silander, Henry Tirri
University of Helsinki, Finland[1]

Peter Grünwald
CWI, Amsterdam[2]

NeuroCOLT2 Technical Report Series

NC2-TR-1998-002

May, 1998[3]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

NeuroCOLT2 Coordinating Partner

Royal Holloway
University of London

Department of Computer Science
Egham, Surrey TW20 0EX, England

For more information contact John Shawe-Taylor at the above address
or email neurocolt@neurocolt.com

[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
[3]Received 06-MAR-1998

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 various 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. The results with several public domain datasets show that the choice of the prior distribution can have a significant effect on the results obtained, especially if the amount of the data available is small. Inspired by our empirical observations, we also introduce a new heuristics for determining the prior distribution. The empirical results show that the heuristics gives consistently very good results with respect to the results obtained by alternative methods.

1 Introduction

Let us consider the following prediction problem for discrete domains: given some sample data D (the *training data*), and an incomplete *query vector* $\vec{u}$, where the values of some of the problem domain variables are not given, we want to predict the values of the missing part $\vec{v}$ of the query vector. The predictions are made probabilistically by using a *predictive distribution* $\mathcal{P}$. We wish to use the training D to learn about the regularities in our domain so that we can arrive at a "good" predictive distribution $\mathcal{P}$, i.e., to a $\mathcal{P}$ with which we 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. We consider two ways to arrive at a $\mathcal{P}$: the *MAP* and the *evidence* approach, both of which are well-known in the Bayesian literature [2]. These approaches 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 and empirically compare several different "optimal" prior distributions P using a model class of practical importance, the class of Bayesian networks [6]. 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 [12], and a prior based on Wallace & Boulton's *Minimum Message Length (MML)* principle [18]. 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, 7]. The contribution of this section is to derive explicit formulas for the MDL and MML priors for the case of Bayesian networks. This involves computing the (expected) Fisher information matrix for Bayesian networks[1].

In section 4 we report on empirical results obtained by applying the predictive distributions described in section 3. For the experiments we have used publicly available real world classification data from the Irvine database. For computational reasons, in the experiments we restrict ourselves to the *Naive Bayes classifier*, a very simple subclass of Bayesian networks. This allows us to provide a systematic comparison of standard Bayesian, MDL and MML-based methods. Section 5 gives the conclusions.

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* $\vec{d} = (x_1, \ldots, x_m)$ is a vector in which all the variables X_i have been assigned a value: by $\vec{X} = \vec{d}$ we mean that $X_1 = x_1, \ldots, X_m = x_m$, where $x_i \in \mathbf{X}_i$. A *random sample* $D = (\vec{d}_1, \ldots, \vec{d}_N)$ is a set of N i.i.d. (independent and identically distributed) data instantiations, where each $\vec{d}_j$ is assumed to be sampled from $\mathcal{P}$, the joint distribution of the variables in X.

Given the values of some of the variables in $\vec{d}$, and some *training data* D, we want to arrive at good predictions for the values of the rest of the variables. More precisely, given D and the values of a subset U (the *clamped* variables) of the variables in X, we wish to determine probabilities for each of the possible values of the variables in the set $\mathrm{V} = \mathrm{X} \setminus \mathrm{U}$ (the *free* variables). Without loss of generality we may assume that the variables are indexed in such a way that we can write $\mathrm{U} = \{X_1, \ldots, X_k\}$, $\mathrm{V} = \{X_{k+1}, \ldots, X_m\}$. We will use $\vec{U} = \vec{u}$ as an abbreviation for a partial data instantiation $X_1 = u_1, \ldots, X_k = u_k$ where $u_i \in \mathbf{X}_i$. Similarly, $\vec{V} = \vec{v}$ stands for $X_{k+1} = v_{k+1}, \ldots, X_m = v_m$ where $v_i \in \mathbf{X}_i$. Hence each data instantiation $\vec{d}$ can be written as $\vec{d} = (\vec{u}, \vec{v})$ for some $\vec{u}$ and $\vec{v}$. We can now state our aim more precisely as follows: we wish to compute, for all possible $\vec{v}$, the probabilities $\mathcal{P}(\vec{V} = \vec{v} \mid \vec{U} = \vec{u}, D = (\vec{d}_1, \ldots, \vec{d}_N))$ which we abbreviate to $\mathcal{P}(\vec{v}|\vec{u}, D)$ and which we can rewrite as follows:

$$\mathcal{P}(\vec{v}|\vec{u}, D) = \frac{\mathcal{P}(\vec{u}, \vec{v}, D)}{\mathcal{P}(\vec{u}, D)} = \frac{\mathcal{P}((\vec{u}, \vec{v})|D)}{\sum_{\vec{v}} \mathcal{P}((\vec{u}, \vec{v})|D)}, \tag{1}$$

where the summing goes over all possible instantiations of the variables in the set V. We see that, in order to compute (1), it suffices to compute the *predictive distribution* $\mathcal{P}(\vec{d}|D)$ for a number of different vectors $\vec{d}$. In most practical applications, this number will be small enough to make the computations feasible.

[1] It should be observed that although the *observed* information matrix for Bayesian networks has been presented in [16], computation of the required *expected* Fisher information is one of the novel contributions of this paper.

We investigate several candidates for $\mathcal{P}$, relative to a parametric family $\mathcal{M}$ of probabilistic models: each model $\Theta \in \mathcal{M}$ defines a probability $P(\vec{d}|\Theta)$ for each data instantiation $\vec{d}$, and, under the i.i.d assumption, for each dataset D. All candidates for $\mathcal{P}$ we consider depend on a prior distribution $P(\Theta)$ defined for all $\Theta \in \mathcal{M}$. The various candidates differ in two ways: first, they differ in the way *how* a given prior distribution is used for constructing a predictive distribution $\mathcal{P}$. Second, they differ in the way the prior distribution is defined in the first place. We now discuss these two issues in turn.

2.1 MAP and Evidence-Based Predictive Distributions

Given a prior distribution $P(\Theta)$, we can arrive at a predictive distribution in two different ways, which we call the *MAP (Maximum A Posteriori probability)* approach and the *evidence* approach, respectively. In order to explain these two approaches, we note first that together with the training data D, our prior determines a posterior distribution $P(\Theta|D)$. By Bayes' rule [2] we have:

$$P(\Theta|D) \propto P(D|\Theta)P(\Theta) \tag{2}$$

Both approaches employ this posterior distribution, but the MAP approach does this using only a single model Θ, while the evidence approach takes into account all the models at once. More precicely, in the MAP approach, the predictive distribution $\mathcal{P}$ is as the distribution corresponding to the single model $\hat{\Theta}(D)$ maximizing the posterior:

$$\hat{\Theta}_\Phi(D) = \arg\max_\Theta P(\Theta|D,\Phi),$$

where Φ denotes the (hyper)parameters used for defining the prior distribution $P(\Theta)$. In the sequel, this type of an instantiation of $\mathcal{P}$ will be denoted by $\mathcal{P}_{\mathrm{MAP}}$. The corresponding predictive distribution is

$$\mathcal{P}_{\mathrm{MAP}}(\vec{d} \mid D, \Phi) = P(\vec{d} \mid D, \hat{\Theta}_\Phi(D)) \overset{\text{i.i.d.}}{=} P(\vec{d} \mid \hat{\Theta}_\Phi(D)). \tag{3}$$

A more sophisticated approach is to average (integrate) over all the models $\Theta \in \mathcal{M}$. In the Bayesian literature the corresponding integral is called the *evidence* or *marginal likelihood.* In the following, evidence-based instantiations of the predictive distribution will be denoted by $\mathcal{P}_{\mathrm{EV}}$. For a data set D, $\mathcal{P}_{\mathrm{EV}}$ is defined as $\mathcal{P}_{\mathrm{EV}}(D) = \int P(D|\Theta)P(\Theta)d\Theta$. Hence the conditional version $\mathcal{P}_{\mathrm{EV}}(\vec{d} \mid D)$ is given by

$$\mathcal{P}_{\mathrm{EV}}(\vec{d}|D,\Phi) = \int P(\vec{d}|D,\Theta,\Phi)P(\Theta|D,\Phi)d\Theta \overset{\text{i.i.d.}}{=} \int P(\vec{d}|\Theta)P(\Theta|D,\Phi)d\Theta. \tag{4}$$

2.2 Prior Distributions for $\mathcal{P}_{\mathrm{EV}}$ and $\mathcal{P}_{\mathrm{MAP}}$

2.2.1 The Uniform Prior

The conceptually simplest non-informative prior is the *uniform prior*, in which case the prior distribution $P(\Theta)$ is a constant: $\forall \Theta \in \mathcal{M} : P(\Theta) = C$. This prior distribution is commonly used because of its mathematical simplicity. One can see from (2) that in this case the MAP predictive distribution becomes $\mathcal{P}_{\text{MAP}}(\vec{d} \mid D, \Phi) = P(\vec{d} \mid \tilde{\Theta}(D))$ where $\tilde{\Theta}(D)$ is the *Maximum Likelihood (ML) model* of classical statistics, i.e., the model $\tilde{\Theta}$ maximizing the data likelihood $P(D|\Theta)$. Hence using the uniform prior in combination with $\mathcal{P}_{\text{MAP}}$ is justified from the point of view of *classical* frequentist (non-Bayesian) statistics.

2.2.2 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. We note first that, given a sufficiently regular model class $\mathcal{M}$, we can construct priors Φ for $\mathcal{M}$ so that for all $\vec{d}$ and all D of any size the following property holds:

$$\mathcal{P}_{\text{EV}}(\vec{d} \mid D, \Phi) = P(\vec{d} \mid \tilde{\Theta}(D \cup D')) \tag{5}$$

Here $\tilde{\Theta}$ is the maximum likelihood model (see above) for the training data D and 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 predicting using $\mathcal{P}_{\text{EV}}$ 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 [6]. Priors Φ which have property (5) will be referred to as *Equivalent Sample Size (ESS) Priors* (for an explanation of this terminology and a more extensive explanation of the ESS concept, see [6]). To see how such priors relate to "degrees of belief", notice that (5) suggests the following interpretation: "prior $\Phi_{D'}$ expresses a prior degree of belief in parameter value Θ that corresponds to the belief one would have in Θ if one first had no prior knowledge at all about Θ and had then seen the data D'."

2.2.3 An MDL Prior

The *Minimum Description Length (MDL) Principle* [12, 13, 14] can be seen as a mathematical formalization of Occam's Razor. It states that the more we are able to compress a set of data, the more we have learnt about it, and the better we will be able to predict future data. *Stochastic Complexity* is the central notion in MDL theory [13, 14]. It is defined as follows [13]:

> *The stochastic complexity S of a data set D relative to a class of models $\mathcal{M}$ is 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 [13]) that the stochastic complexity S can be written as $S = -\log \mathcal{P}_{\text{SC}}$ where $\mathcal{P}_{\text{SC}}$ is some probability distribution. One can show ([12], section 3) that under certain conditions, predictive distributions based on $\mathcal{P}_{\text{SC}}$ lead to optimal predictive performance.

The definition above gives no explicit formula for the stochastic complexity. There are several reasons [13] why $-\log \mathcal{P}_{\mathrm{EV}}(D)$ is a good candidate for the stochastic complexity. Therefore, this formula was originally proposed by Rissanen as the formalization of the stochastic complexity: $S(D) = -\log \mathcal{P}_{\mathrm{EV}}(D)$. Recently, however, Rissanen [14] has shown that there exists a code that is itself not dependent on any prior distributions of parameters, and which in general yields even shorter codelengths than the code with lengths $-\log \mathcal{P}_{\mathrm{EV}}(D)$. Here "shorter" means that for some data sets the codelength will be considerably shorter, while for *all* other data sets it will be negligibly longer. This led Rissanen to redefine stochastic complexity using this new code. The newly defined stochastic complexity can, in the light of the discussion above, once again be written as $S'(D) = -\log \mathcal{P}_{\mathrm{SC}}(D)$ for some probability distribution $\mathcal{P}_{\mathrm{SC}}$. Rissanen [14] gives an explicit formula for $\mathcal{P}_{\mathrm{SC}}$ which does not depend on any prior distribution. Basing a predictive distribution directly on $\mathcal{P}_{\mathrm{SC}}$ however leads to computational problems. In the same paper [14], however, by extending on results of Clarke & Barron [3], Rissanen showed that under suitable technical conditions we have

$$-\log \mathcal{P}_{\mathrm{SC}}(D) = -\log \mathcal{P}_{\mathrm{EV}}(D) + o(1) \tag{6}$$

where $\mathcal{P}_{\mathrm{EV}}$ is the evidence distribution $\mathcal{P}_{\mathrm{EV}} = \int P(D|\Theta)\pi(\Theta)d\Theta$ instantiated to a particular prior known in the literature as *Jeffrey's prior* [2, 3], denoted in the sequel by $\pi(\Theta)$. We remark that for the model class of Bayesian networks, which will be used in the experimental part of this paper, the regularity conditions required for (6) all apply (we discuss this in more detail in [9]). Furthermore the $o(1)$-term in (6) shrinks to zero very rapidly for the class of multinomial models, as verified by Rissanen [14]. The fact that a discrete Bayesian network consists of collections of multinomial models suggests a similar behavior of the $o(1)$ term also for the Bayesian network model class.

Jeffrey's prior [3] is defined by

$$\pi(\Theta) = |I(\Theta)|^{1/2} / \int |I(\eta)|^{1/2} d\eta, \tag{7}$$

where $|I(\Theta)|$ denotes the determinant of the *(expected) Fisher information matrix* $I(\Theta)$ (see [2]). Denoting $\Theta = (\theta_1, \ldots, \theta_r)$, entry (i, j) of matrix $I(\Theta)$ is defined as

$$[I(\Theta)]_{i,j} = -E_{\Theta}\left[\frac{\partial^2 \log P(D|\Theta)}{\partial\theta_i \partial\theta_j}\right]. \tag{8}$$

We see that from this MDL point of view, the optimal predictive distribution is equal to $\mathcal{P}_{\mathrm{EV}}$ with the prior instantiated to the Jeffrey's prior $\pi(\Theta)$.

2.2.4 An MML Prior

Minimum Message Length (MML) Inference [17, 18] is based on a similar philosophy to the *MDL* principle; however, 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 comparison of both the MDL and MML approaches). For our purposes, it is sufficient to note the following two differences: first, in MML modeling we always end up with a *single* model $\Theta_{\mathrm{MML}}(D)$ which is the "MML-optimal" model for the data set D. This model can then be used to arrive at a predictive distribution $P(\vec{d}|\Theta_{\mathrm{MML}}(D))$. The approach is in contrast with the MDL/Stochastic Complexity approach where all the models in the class are taken into account when forming the predictive distribution. A second difference between MML and MDL is that, though both employ priors, the priors used in MDL serve as a technical tool only, namely to compute an approximation to $\mathcal{P}_{\mathrm{SC}}$ (which itself does not depend on any prior). Indeed, as we have seen, there is only one particular prior which can be used for this purpose. On the other hand, MML adopts a Bayesian philosophy regarding priors—prior distributions in MML are taken to represent prior knowledge or beliefs about the models in the model class used. Hence in order to apply MML, the user must provide a prior $P(\Theta)$ to reflect his/her prior beliefs. Omitting all mathematical details (which can be found in [18]), the MML estimator $\Theta_{\mathrm{MML}}(D)$ for training data D and prior P is defined as follows:

$$\Theta_{\mathrm{MML}}(D) = \arg\max_{\Theta\in\mathcal{M}} \frac{P(D|\Theta)P(\Theta)}{|I(\Theta)|^{1/2}} \tag{9}$$

Here $|I(\Theta)|$ is the determinant of the Fisher information matrix $I(\Theta)$ as defined in equation (8). Combining this with equation (7) we see that $\Theta_{\mathrm{MML}}(\mathrm{D})$ for prior $P(\Theta)$ is equal to the MAP-model $\hat{\Theta}(\mathrm{D})$ for prior $P'(\Theta) \propto P(\Theta)/\pi(\Theta)$. The MML predictive distribution $\mathcal{P}_{\mathrm{MML}}$ with respect to prior P is then equal to $\mathcal{P}_{\mathrm{MAP}}$ with respect to prior P'. In the empirical part of the paper we report experiments with the prior P instantiated to ESS priors, since they have a clear interpretation as representing prior beliefs (corresponding to their interpretation in MML theory). 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 Bayesian Network Predictive Distributions and Their Priors

A Bayesian (belief) network [11, 15] 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 $X_1, \ldots, X_m$. Each network topology defines a set of independence assumptions which allow the joint probability distribution for a data vector $\vec{d}$ to be written as a product of simple conditional probabilities,

$$P(\vec{d}) = P(X_1 = x_1, \ldots, X_m = x_m) = \prod_{i=1}^{m} P(X_i = x_i|\mathrm{pa}_i = q_i), \tag{10}$$

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 $P(\vec{d} \mid \Theta)$ is uniquely determined by fixing the values of the parameters $\Theta = (\theta^1, \ldots, \theta^m)$, where $\theta^i = (\theta^i_{11}, \ldots, \theta^i_{1n_i}, \ldots, \theta^i_{c_i1}, \ldots, \theta^i_{c_in_i})$, 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 we assume that $\theta^i_{q_ix_i} > 0$ and $\theta^i_{q_in_i} = 1 - \sum_{x_i=1}^{n_i-1} \theta^i_{q_ix_i}$. Furthermore, all the conditional distributions of the variables given their parents are assumed to be multinomial, i.e., $X_{i|q_i} \sim \mathrm{Multi}(1; \theta^i_{q_i1}, \ldots, \theta^i_{q_in_i})$. Since the family of Dirichlet distributions is *conjugate* (see e.g. [2]) to the family of multinomials, i.e., the functional form of parameter distribution remains invariant in the prior-to-posterior transformation, 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 $(\theta^i_{q_i1}, \ldots, \theta^i_{q_in_i}) \sim \mathrm{Di}(\mu^i_{q_i1}, \ldots, \mu^i_{q_in_i})$, where $(\mu^i_{q_i1}, \ldots, \mu^i_{q_in_i})$ are the hyperparameters of the corresponding distributions. From the definition of Dirichlet distributions [2] we see that this implies the following functional form for the prior distributions $P(\theta^i_{q_i1}, \ldots, \theta^i_{q_in_i})$:

$$P(\theta^i_{q_i1}, \ldots, \theta^i_{q_in_i}) \propto \prod_{j=1}^{n_i} (\theta^i_{q_ij})^{(\mu^i_{q_ij}-1)} \tag{11}$$

It turns out that both the uniform prior and the ESS priors are Dirichlet distributions [7]. We will see that for the subclass of Bayesian Networks used in our experiments, Jeffrey's prior is of the Dirichlet form too. Moreover, we have seen in the previous section that the priors $P'(\Theta)$ we need for the MML predictive distributions are arrived at by dividing an ESS prior by Jeffrey's prior. Since both of these priors are Dirichlet, it follows immediately from (11) that the resulting prior is of the Dirichlet form too. Consequently, all the priors of interest in here are Dirichlet, and we may therefore safely assume that our priors have the functional form given by (11). This allows us to derive explicit expressions for $\mathcal{P}_{\mathrm{MAP}}$ and $\mathcal{P}_{\mathrm{EV}}$. Since this has been done before [4, 7] we will omit it here. The computation of ESS priors for Bayesian Networks can also be found in the literature [6], hence it is also omitted. Instead, in the sequel we focus on our contribution: the computation of Jeffrey's prior $\pi(\Theta)$.

3.1 Computing Jeffrey's prior for Bayesian Networks

In the following, we denote by $I()$ 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 $\vec{d_j}$; q_{ji} stands for the configuration of the parent variables of X_i in $\vec{d_j}$.

We first compute the Fisher information matrix (8). 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 $\log P(\vec{X}|\Theta)$ 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:

$$\mathrm{E}_\Theta \left[\frac{-\partial^2 \log P(\vec{d_j} \mid \Theta)}{\partial(\theta^i_{q_il})^2} \right] = \frac{P(\mathrm{pa}_i = q_i \mid \Theta)}{\theta^i_{q_in_i}} + I(l_1, l_2) \cdot \frac{P(\mathrm{pa}_i = q_i \mid \Theta)}{\theta^i_{q_il_1}}$$

We now have an expression for each element of $I(\Theta)$. This allows us to calculate $\sqrt{|I(\Theta)|}$ using a combination of several standard methods for calculating determinants, yielding

$$\pi(\Theta) \propto \sqrt{|I(\Theta)|} = \prod_{i=1}^{m} \prod_{q_i=1}^{c_i} (N \cdot P^i_{q_i})^{\frac{n_i-1}{2}} \prod_{l=1}^{n_i} (\theta^i_{q_i l})^{-\frac{1}{2}} \propto \prod_{i=1}^{m} \prod_{q_i=1}^{c_i} (P^i_{q_i})^{\frac{n_i-1}{2}} \prod_{l=1}^{n_i} (\theta^i_{q_i l})^{-\frac{1}{2}}. \tag{12}$$

The derivation of this result is somewhat involved and omitted here, but the details can be found in [9].

Computing the Jeffrey's prior as formulated above requires computing for each variable the marginal distribution of its parents. Unfortunately, for multi-connected Bayesian networks, this problem is known to be NP-hard [5]. In the experiments reported in the next section, we used a simple tree-structured Bayesian network, in which case Jeffrey's prior is of the proper conjugate form, and can be computed efficiently.

4 Empirical Results

In our experiments we have concentrated on the standard classification problem, where the task is to predict the value of a single *classification variable* (denoted by X_m), given the values of all the other variables. For computational reasons, we restricted ourselves in this set of experiments to the simple Naive Bayes classifier case, where the variables in $X_1, \ldots, X_{m-1}$ are assumed to be independent given the value of variable X_m. Consequently, we can regard the Naive Bayes model as a simple tree-structured Bayesian network, where the classification variable X_m forms the root of the tree, and the other variables are represented by the leaves. Despite of its apparent simplicity, this model has been proved to perform surprisingly well compared to more complex models [10, 8].

In Naive Bayes case, the Jeffrey's prior can be written as a product of Dirichlet distributions, yielding

$$\Theta \sim \mathrm{Di}(\frac{1}{2}(\sum_{i=1}^{m-1}(n_i-1)+1), \ldots, \frac{1}{2}(\sum_{i=1}^{m-1}(n_i-1)+1)) \prod_{i=1}^{m-1} \prod_{k=1}^{K} \mathrm{Di}(\frac{1}{2}, \ldots, \frac{1}{2}),$$

where n_i denotes the number of values of variable X_i, and K is the number of classes (the number of values of the class variable X_m). For the ESS priors in this case, it is easy to see that if the hyperparameters for the prior distribution of the class variable have value μ, then the hyperparameters for a variable with n_i values should be set to μ/n_i.

In the experiments we used eight public domain classification data sets of varying size (the data sets can be obtained from the UCI data repository[2]). Table 1 describes the size (N), the number of attributes (m), and the number of classes (K) for each of these data sets.

[2] "http://www.ics.uci.edu/~mlearn/".

Table 1 The datasets used in the experiments.

Dataset	Data vectors	Attributes	Classes
Heart Disease	270	14	2
Iris	150	5	3
Lymphography	148	19	4
Australian	690	15	2
Breast Cancer	286	10	2
Diabetes	768	9	2
Glass	214	10	6
Hepatitis	150	20	2
Primary Tumor	339	18	21

In order to compare the predictive accuracy of different predictive distributions, we used two different utility functions: the *log-score* and the *0/1-score*. The log-score of a predictive distribution $\mathcal{P}(X_m|\vec{u}, D)$ is defined as $-\log \mathcal{P}(X_m = k|\vec{u}, D)$ where k denotes the actual ("true") classification, i.e., the correct value of X_m. When using logarithm of base two, the log-score has the following coding interpretation: If one encodes the data using the code corresponding to $\mathcal{P}(X_m|\vec{u}, D)$, then $\log \mathcal{P}(X_m = k|\vec{u}, D)$ is equal to the number of bits one needs to describe the actual outcome k. On the other hand, if we think the correct predictive distribution to give in this case probability one to the actual outcome, and probability zero to other outcomes, we see that the log-score is equivalent to the cross-entropy or Kullback-Leibler distance between the above defined degenerate distribution and the predictive distribution produced.

For the 0/1-score, we simply first determine the k for which the probability $\mathcal{P}(X_m = k|\vec{u}, D)$ is maximized, and the 0/1-score is then defined to be 1, if the actual outcome indeed was k, otherwise 0.

Table 2 gives the averaged leave-one-out crossvalidated scores over all the datasets used. In order to be able to compare the results between different datasets, the crossvalidation scores are normalized so that the highest result gets always a normalized score of 100.0, and a non-normalized score s is normalized to the value $s/b*100.0$, where b denotes the non-normalized best score. In this table, by AB we mean that the corresponding result is obtained by the corresponding predictive distribution ($\mathcal{P}_{\mathrm{EV}}$ or $\mathcal{P}_{\mathrm{MAP}}$) with A as the prior distribution for the class variable, and B as the prior distribution for the other variables. The possible values for A are U (uniform prior), J (Jeffrey's prior), and H (a prior determined by a heuristic rule to be explained later). In addition to the values U and J, B can also have value E, which means that the hyperparameters for the variables $X_1, \ldots, X_{m-1}$ are taken to be the ESS values as defined above, computed from the hyperparameter values for the class variable X_m determined by A. By 'n/a' we mean that the resulting prior distribution is not of the proper Dirichlet form, and hence the corresponding predictive distributions can not be determined without major computational difficulties.

Table 2 The crossvalidation scores obtained, normalized and averaged over the datasets.

Used prior:	UU	UJ	UE	JU	JU/J	JJ	JE	JE/J	HE	HE/J
Log-score										
EV	89.5	87.9	86.5	88.3	90.5	86.8	94.1	93.9	95.9	95.1
MAP	77.6	n/a	n/a	77.8	87.9	n/a	92.8	92.9	96.3	95.5
0/1-score										
EV	98.5	98.9	98.9	98.0	98.4	97.9	97.2	97.0	96.7	96.8
MAP	98.1	n/a	n/a	97.9	98.8	n/a	97.3	97.6	97.5	96.8

From Table 2 we can see that the 0/1-score results do not differ very much between different priors and predictive distributions. This means that for purely discriminative tasks, where the goal is only to choose one of the possible values of the class variable, the choice of the predictive distribution or the prior distributions does not have a significant role. Nevertheless, if one wishes to estimate the expected risks following the classifications made, the predictive distribution should give good results in the log-score sense. In this case, the results may vary a lot depending on the predictive distribution used, and also on the prior distribution. Therefore in the following, we concentrate on the log-score results only.

From the results obtained, the first observation is that the evidence-based predictive distribution gives usually better results than the MAP approach. If the prior distribution for the class variable is assumed to be the same as for the other variables, then we can see that on the average, the uniform prior (UU) gives slightly better results than the Jeffrey's prior (JJ). This can be explained by the fact that the Jeffrey's prior is optimal in the worst-case sense, thus the average-case behavior may not be optimal. On the other hand, if the ESS priors are to be used, then we can see that the results obtained by using the Jeffrey's prior for the class variable (JE) gives much better results than the uniform case (UE). This raises an interesting question about the optimal value for the class variable hyperparameters in the ESS case. Figure 1 shows some illustrative examples of the behavior of the crossvalidated log-score as the function of this parameter. From these pictures we can see that the value 1 (corresponding to the uniform distribution) and the value corresponding to the Jeffrey's prior (marked by the vertical line on the left) are in many cases far from optimal, and that the optimal value is in many cases much larger than either of these. This observation is the motivation for the following heuristics: if we wish in the ESS prior case our prior information to have the same "weight" as our sample data D, i.e., the prior and the data to have a similar effect on the predictive distribution, then the hyperparameters for the class variable should be set to N/K, where N is the size of the dataset D, and K is the number of the classes (the number of values of the class variable X_m). The corresponding point in

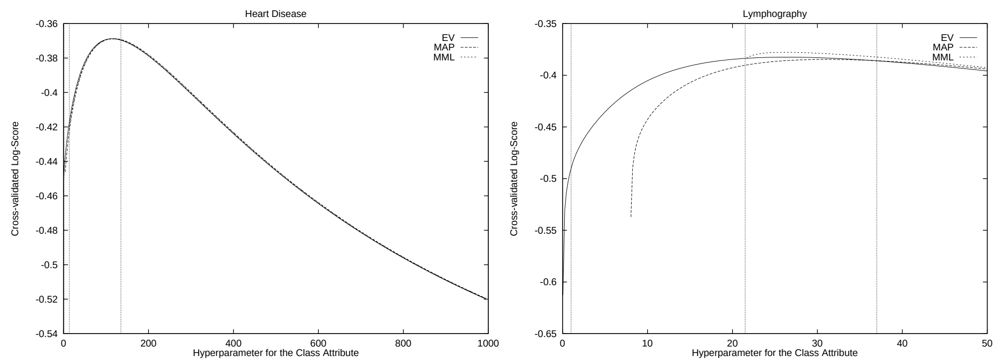

Figure 1 Examples of crossvalidation results with the ESS priors as a function of the hyperparameter value for the class variable.

the X-axis in the figures is marked with a vertical line (right from the "J" line). As can be seen from these pictures, and also from Table 2, the "HE" heuristics performs consistently very well both with the evidence prediction, and with the MAP prediction. It should be noted that using the MML-based approach (normalizing the ESS prior with the Jeffrey's prior) seems to have no effect on the results obtained (all the three lines in the picture coincide nearly perfectly).

As already mentioned above, the evidence predictive distribution produces usually better results than the MAP predictive distribution. In [8, 9] it was observed that the difference becomes larger if the size of the training data is decreased. To see whether the ESS priors would make a difference with the small data size behavior, we ran a series of tests where we studied how the prediction quality of our various approaches depends on the size of the training set D. In these experiments, the data was first randomly partitioned into a training set and test set of equal size. We then iteratively computed the log-score for the test set by using in the first phase only the first vector of the training set as the training data D, then the two first vectors, then three, and so on, until at the last step we would eventually be using all the vectors in the training set. This procedure was then repeated 100 times by using each time a different, randomly chosen partitioning of the data for producing the training set and the test set. The statistics of the behavior of different predictive distributions as a function of increasing amount of training data should now give us some idea as to the typical behavior of our prediction methods.

As a general observation, the results with all the datasets used show very similar behavior. Figure 2 depicts illustrative examples of average log-score behavior plotted as a function of the training set size. From these pictures we see that the evidence-based approaches are quite robust in the sense that they predict very well even with small training sets, regardless of the prior used. This shows that the data sets used here are quite redundant, and when properly used, only a very small sample of these data sets is needed for constructing good models. On the other hand, the MAP approach with uniform priors (UU) is much more sensitive to the amount of data available. Nevertheless, with the ESS priors, the MAP becomes much more robust, producing curves similar to the evidence-based methods. This phenomenon can be explained by the fact

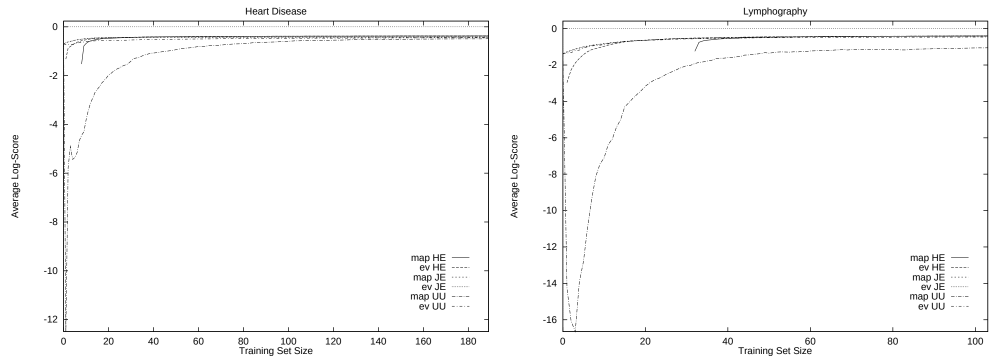

Figure 2 Illustrative examples of the behavior of the prediction error as a function of the training set size.

that the evidence-based approaches are more conservative methods as they base their predictions on expected values of the parameters, i.e., on integrating over all the parameter values, while the MAP-UU approach makes more "eager" predictions based on the maximum likelihood estimator. Changing the priors from UU to JE or HE increases the absolute value of the hyperparameters, which makes the system less eager, and hence more robust with respect to the amount of data available. These 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. In [9] the same experiments were performed using the 0/1-loss function, with similar results.

5 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 various equivalent sample size priors. To study the relevance of the various approaches in practice, we instantiated the methods for the family of Bayesian networks. In the experimental part of the paper we restricted ourselves for computational reasons to a simple subclass of Bayesian networks, the Naive Bayes classifier. While in the case of large training samples all methods give very good results, we observe that 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. Inspired by our empirical observations, we also introduced a new heuristics for determining the prior distribution. The empirical results show that the heuristics gives consistently very good results with respect to the results obtained by alternative methods.

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. The primary tumor, breast cancer, and the lymphography domains were obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklič for providing the data.

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] G.F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, March 1990.

[6] 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.

[7] 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.

[8] 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*, pages 311–318, Ft. Lauderdale, Florida, January 1997.

[9] 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.

[10] P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 223–228, San Jose, CA, July 1992. MIT Press.

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

[12] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society*, 49(3):223–239 and 252–265, 1987.

[13] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey, 1989.

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

[15] R.D. Shachter. Probabilistic inference and influence diagrams. *Operations Research*, 36(4):589–604, July-August 1988.

[16] B. Thiesson. Score and information for recursive exponential models with incomplete data. Technical Report R-95-2020, Aalborg University, Institute for Electronic Systems, Department of Mathematics and Computer Science, 1995.

[17] C.S. Wallace and D.M Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

[18] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49(3):240–265, 1987.