Journal of Logic, Language and Information (2006) 15: 5–20
DOI: 10.1007/s10849-005-9004-8
© Springer 2006

Bayesian Model Learning Based on Predictive Entropy

JUKKA CORANDER and PEKKA MARTTINEN
Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FIN-00014, Finland
E-mail: jukka.corander@helsinki.fi

(Received 20 April 2005; in final form 24 July 2005)

Abstract. Bayesian paradigm has been widely acknowledged as a coherent approach to learning putative probability model structures from a finite class of candidate models. Bayesian learning is based on measuring the predictive ability of a model in terms of the corresponding marginal data distribution, which equals the expectation of the likelihood with respect to a prior distribution for model parameters. The main controversy related to this learning method stems from the necessity of specifying proper prior distributions for all unknown parameters of a model, which ensures a complete determination of the marginal data distribution. Even for commonly used models, subjective priors may be difficult to specify precisely, and therefore, several automated learning procedures have been suggested in the literature. Here we introduce a novel Bayesian learning method based on the predictive entropy of a probability model, that can combine both subjective and objective probabilistic assessment of uncertain quantities in putative models. It is shown that our approach can avoid some of the limitations of the earlier suggested objective Bayesian methods.

Key words: Bayesian inference, entropy, information theoretic criteria, objective model learning

1. Introduction

Consider a situation where judgement of the relative plausibilities of probability models in a finite class $\mathcal{M} = \{M_j : j \in J\}$, is sought in the light of some observed data. Plausibility should be understood here as a general term extending over concepts like *evidence*, *utility*, *degree of belief*, *predictive power* etc. Among generally acknowledged principles that attempt to provide an answer to such a question, there are the information theoretic approach pioneered by Akaike (1974, 1978, 1979), the prequential approach (Dawid, 1984), the coding theoretic approach (Rissanen, 1987, 1995), and the normative Bayesian framework, which is discussed in detail in Bernardo and Smith (1994).

The above mentioned methods are from the theoretical perspective suitable for ordering the elements of $\mathcal{M}$ with respect to their ability to explain observed data. In typical applications, models in the class $\mathcal{M}$ have varying parametric dimensionality, and therefore, the maximum likelihood (ML) principle is not applicable as such to the assessment of the relative plausibility of any particular M_j. Even advocates

of the ML principle have clearly stated that its use is often unfortunate in the model comparison context (Lindsey, 1996), a fact which has partly stimulated the emergence of some of the alternative approaches, such as Akaike (1974). For a discussion of the problematic issues of measuring evidence using the law of likelihood, see the discussion of Royall (1992).

Using the theory developed in Bernardo and Smith (1994) as a benchmark, we discuss the requirements of the normative Bayesian framework to be met in order to obtain a coherent solution to the stated problem. Many prominent Bayesians, see, e.g. Lindley (1991, 1992), have claimed that the only satisfactory measures of support for scientific hypotheses presented in terms of parametric models are probability based. The Bayesian framework requires that a scientist seeking answers to questions involving model uncertainty expresses strictly probabilistic, *subjective* beliefs about both observables and unobservables. By doing so, coherent information processing is guaranteed with the normative tools of inference. This framework could thus be considered as an ideal, which tells us how we should do inference if we are able to meet its requirements.

From the pragmatic point of view, it cannot be ignored that specification of strictly probabilistic beliefs can be a daunting task even in problems of moderate dimensionality, and that the computations involved therein may be intractable in practice. Therefore, a considerable theoretical activity has been devoted to the development of automated methods for assessment of the elements of $\mathcal{M}$ in the light of data. Some of these methods yield consistent approximations to Bayesian decision rules, whereas others fall outside the normative framework even asymptotically.

Generally, the automated methods have the common aim of being *objective*, i.e. the results should not be dependent on the opinion of any particular applier of the method, but solely be a function of the observed data and the class of putative models $\mathcal{M}$. Here we discuss to which extent such an aim is attainable, and investigate assessment of relative plausibilities of models using the concept of predictive entropy, which enables combination of both subjective and objective probabilistic assessment of uncertain quantities in putative models. The structure of the paper is as follows. In Section 2, the Bayesian model learning problem is formally defined, and in Section 3 we derive the entropy based approach. Illustrative examples of its application are provided in the final section, together with some concluding remarks.

2. Bayesian Model Learning

A tradition put forth by Bernardo and Smith (1994), see also the discussion in Aitkin (1991), is to consider the comparison of model plausibilities from three different perspectives; $\mathcal{M}$-closed, $\mathcal{M}$-completed and $\mathcal{M}$-open. Within these perspectives, the answers obtained depend on whether or not we wish to make some inference, conditional on that we have actually chosen a model from $\mathcal{M}$. To formalize the problem, the comparisons are embedded into a decision situation, and under the

principles of quantitative coherence, the solution is to order the models with respect to their expected utilities.

Let $\mathbf{x}$ denote generically the data available for judging plausibilities of the models in $\mathcal{M}$. Each model M_j, $j \in J$, is assumed to be labeled by a finite-dimensional parameter θ_j, taking values in a generic space Θ_j. For some models the parametric dimension of the model can be specified explicitly, and will be denoted by d_j.

Following the notation of Bernardo and Smith (1994), within the $\mathcal{M}$-closed perspective it is assumed that one of the models in $\mathcal{M}$ is the true data generating mechanism, and that the (subjective) belief model for $\mathbf{x}$ is defined as

$$p(\mathbf{x}) = \sum_{j \in J} P(M_j) p(\mathbf{x} \mid M_j), \tag{1}$$

where $P(M_j)$ is *a priori* degree of belief (or probability) of M_j, and $p(\mathbf{x} \mid M_j)$ is the marginal data density given M_j, which equals

$$\int_{\Theta_j} p(\mathbf{x} \mid \theta_j) \pi(\theta_j) \, d\theta_j, \tag{2}$$

where $p(\mathbf{x} \mid \theta_j)$ is the conditional data density given θ_j, and $\pi(\theta_j)$ is the prior density of θ_j.

If the problem at hand is formally treated as a decision problem in which a choice is to be made from $\mathcal{M}$, ω is an unknown of interest, and $u(M_j, \omega)$ is a utility function, the plausibilities of the models are specified by the expected utilities

$$\bar{u}(M_j \mid \mathbf{x}) = \int_{\Omega} u(M_j, \omega) p(\omega \mid \mathbf{x}) \, d\omega, \tag{3}$$

where

$$p(\omega \mid \mathbf{x}) = \sum_{j \in J} p_j(\omega) P(M_j \mid \mathbf{x}), \tag{4}$$

and

$$P(M_j \mid \mathbf{x}) = \frac{P(M_j) p(\mathbf{x} \mid M_j)}{\sum_{j \in J} P(M_j) p(\mathbf{x} \mid M_j)}. \tag{5}$$

The quantity $P(M_j \mid \mathbf{x})$ can be interpreted as conditional or posterior probability of M_j being the true data generating mechanism, or simply be regarded as a weight in the predictive mixture distribution. The most challenging practical issues related to this strategy are the choice of the priors $\pi(\theta_j)$, $j \in J$, and the computation of (5). Numerous methods have been suggested for derivation of the priors as a function of the model structure, see e.g. Kass and Wasserman (1996). However, these methods

typically yield improper reference priors for which $\int_{\Theta_j} \pi(\theta_j) d\theta_j \neq 1$, leading to undetermined probabilities (5).

Bernardo and Smith (1994) claim that, when literally taken, the above perspective appears to be quite unconvincing in reality, where the truth of a single model available in $\mathcal{M}$ is seldom to be expected. In contrast, within the $\mathcal{M}$-completed perspective, the truth of any M_j is not assumed, but there exists a separate belief model, with respect to which the models in $\mathcal{M}$ are to be evaluated. Generally, we are then interested in how well the elements of $\mathcal{M}$ work as approximations to the separate belief model.

Finally, in the $\mathcal{M}$-open perspective, the models are simply available for comparison, no overall actual belief model $p(\mathbf{x})$ is assumed to be available. According to us, this perspective captures an important situation appearing in scientific modeling of real-world phenomena, in which a scientist is in lack of competence to proceed with the strictly normative Bayesian framework. Gutiérrez-Peña and Walker (2001) discuss the $\mathcal{M}$-closed and $\mathcal{M}$-open perspectives, and develop a strategy to model selection using a nonparametric approach.

Paradoxically, to be able to utilize the arguments of the *normative* Bayesian framework for assessment of relative plausibilities of the models in $\mathcal{M}$, it would inevitably be necessary to proceed within the formalism defined under the $\mathcal{M}$-closed perspective, as is clearly stated by de Finetti (1974). However, it is important to notice that in the specification of the subjective belief model as in (1), it is really not necessary to believe in the existence of any true model. The prior quantities $P(M_j)$ can simply be regarded as predictive weights, measuring our opinion about the expected relative predictive performance of the models. If the subjective Bayesian approach is not taken, then whatever procedures we use for model assessment, they need to be theoretically motivated either as (i) approximations to Bayesian decision rules under quantitative coherence, or (ii) as formal rules for model learning based on some alternative inference theory. The lack of solid normative theory behind the $\mathcal{M}$-completed and $\mathcal{M}$-open perspectives, has in fact not been clearly put forth in Bernardo and Smith (1994).

The challenge related to the specification of $p(\mathbf{x})$ in (1) through the priors $\pi(\theta_j)$, $P(M_j)$, $j \in J$, for a given model class $\mathcal{M}$ and a data set, has inspired the development of objective Bayesian methods for model learning. Also, the demonstrated sensitivity of subjectively determined posterior probabilities (5) for various common modelling situations, illustrates the need for an objective approach. A considerable number of Bayesian automated methods for ranking the elements of $\mathcal{M}$, given the observed data, have been introduced in the literature, see e.g. Schwarz (1978), Aitkin (1991), O'Hagan (1995), Berger and Pericchi (1996), Key et al. (1999), Bayarri and Berger (1998), Berger and Mortera (1999) and Perez and Berger (2002). Such methods do not typically necessitate the explicit specification of proper prior distributions for the model parameters, however, some of them rely on the reference type of priors (Kass and Wasserman, 1996). Examples of successful performance of the automated methods have been generally demonstrated, however, a

variety of acknowledged deficiencies have also been identified (see, e.g. discussions in the papers), including statistical inconsistency, poor small sample behavior, and inapplicability to even extremely simple, commonly used discrete model families.

3. Predictive Entropy Based Model Comparison

In Bernardo (1999) and Bernardo and Rueda (2002), arguments for considering nested model comparison in a Bayesian decision theoretic framework utilizing information theoretic concepts were put forth. In particular, these papers discussed the problem of providing meaningful probabilities (both prior and posterior) as measures of the degree of belief to a particular probability model. Rather elegantly, the authors then derived a reference criterion for determining the plausibility of a model as a description of an observed data set. However, as such, the approach can only be applied to pairwise comparisons of probability models which are nested in the standard parametric manner, and does not yield an unambiguous ordering of the elements of $\mathcal{M}$ in cases where there are more than two candidate models. In Corander (2003a,b), a modification of the reference criterion was introduced for the structural learning of undirected graphical models (see, e.g. Jordan, 2004). Here, we consider a further generalization of the Bayesian information and decision theoretic approach to model comparison. There exist also several related Bayesian learning rules derived from the coding theoretic principle (see, e.g. Meir and Merhav, 1995; Engel et al., 2003; Weissman and Merhav, 2003).

Assume that the utility of a density $p_j(\cdot \mid \theta_j)$, $j \in J$, to describe the behavior of a random quantity $\mathbf{y}$ is measured with a logarithmic score function (see Bernardo and Smith, 1994) $\alpha \log p_j(\mathbf{y} \mid \theta_j) + \beta(\mathbf{y})$, such that $\alpha > 0$, and $\beta(\cdot)$ is an arbitrary, real valued function in the joint sample space $\mathcal{X}$. Since $\beta(\cdot)$ does not depend on the actual density used, it will be ignored in the sequel by setting $\beta(\mathbf{y}) = 0$, for all $\mathbf{y} \in \mathcal{X}$, as in Bernardo (1999). To simplify the notation, we may abbreviate $p_j(\cdot \mid \theta_j)$ as p_j. Given that we have the data $\mathbf{x}$ comprising n observations currently available, the expected utility of p_j for the prediction of a future data set $\mathbf{y}$ also comprising n observations, can be written as

$$\bar{u}(p_j \mid \mathbf{x}) = \int_{\Theta_j} u(p_j, \theta_j)\pi(\theta_j \mid \mathbf{x})\, d\theta_j, \tag{6}$$

where

$$u(p_j, \theta_j) = \alpha \int_{\mathcal{X}} p_j(\mathbf{y} \mid \theta_j) \log p_j(\mathbf{y} \mid \theta_j)\, d\mathbf{y}, \tag{7}$$

is the negative entropy of $p_j(\cdot \mid \theta_j)$, and $\pi(\theta_j \mid \mathbf{x})$ is the posterior of θ_j,

$$\pi(\theta_j \mid \mathbf{x}) = \frac{p_j(\mathbf{x} \mid \theta_j)\pi(\theta_j)}{\int_{\Theta_j} p_j(\mathbf{x} \mid \theta_j)\pi(\theta_j)\, d\theta_j}, \tag{8}$$

where $\pi(\theta_j)$ is a prior deemed suitable for the model M_j. The most appropriate objective choice of the prior seems to be provided by the method introduced in Bernardo (1979), and developed further in Berger and Bernardo (1989, 1992). In the subjective Bayesian approach defined in (1), the priors $\pi(\theta_j)$, $j \in J$, are required to be coherent about features that are common to different models. Here, no such claims are made, but the prior $\pi(\theta_j)$ (either subjectively or objectively determined) is a function of the characteristics of M_j only.

In Bernardo (1999), the expectation in (7) is instead taken with respect to a true sampling model in which $p_j(\cdot \mid \theta_j)$ is nested, leading to a measurement of the expected performance of the approximation provided by $p_j(\cdot \mid \theta_j)$. To avoid the necessity of using such an encompassing model approach, (6) is introduced to enable characterization of the expected predictive performance of $p_j(\cdot \mid \theta_j)$, were it an appropriate sampling model for the data. Notice that the current uncertainty about θ_j is taken into account, since it is quantified by the posterior $\pi(\theta_j \mid \mathbf{x})$. Also, for this particular utility measure, the precise value of α is irrelevant for ranking the models in $\mathcal{M}$, since it does not depend on the actual model used. In the sequel, the utility structure is simplified by setting $\alpha = 1$.

As discussed in Bernardo and Smith (1994) and Bernardo (1999), the above type utility function measures the terminal utility of using p_j to explain the data structure. However, in some situations, we would wish to extend the utility function additively to include a measure of the cost c_j expected from using p_j, which measures the relative simplicity of the model, or reflects some other scientific implications considered relevant in the particular application. Then, the expected utility equals

$$\bar{u}(p_j \mid \mathbf{x}) = \int_{\Theta_j} u(p_j, \theta_j)\pi(\theta_j \mid \mathbf{x})\, d\theta_j - c_j. \tag{9}$$

This slightly more general form of the utility measure will be referred to as the Bayesian entropy criterion (BEC).

Certain characteristics of the suggested utility measure are important to be acknowledged. Firstly, the utility is not invariant with respect to monotone reparametrizations of the models in $\mathcal{M}$. Surely, invariance in this sense is one of the fundamental properties sought for objective prior distributions, which is discussed in detail in Kass and Wasserman (1996). Several of the earlier mentioned automated model learning methods are invariant, or at least, invariant under limited classes of monotone transformations. However, for model learning problems, invariance property has generally its price as a limited applicability, since it is typically achieved through either an encompassing model (Bernardo, 1999), existence of a training sample of the data $\mathbf{x}$ associated with minimal information (O'Hagan, 1995; Berger and Pericchi, 1996), or asymptotic approximations (Schwarz, 1978). For instance, none of these arguments are applicable to the common problem of learning graphical models for sparse high-dimensional contingency tables (see the discussion of O'Hagan, 1995).

Secondly, the utility measure should not be applied to model learning problems where certain candidates in $\mathcal{M}$ limit directly the entropy of p_j in terms of restrictions to θ_j. This is evident from the fact that the predictive entropy can be made arbitrarily small by fixing certain elements of θ_j, related to the variance of the observed variables. Fortunately, model learning applications related to such hypotheses are generally of limited interested, which makes this restriction less important.

Spiegelhalter et al. (2002) considered the quantification of the expected future performance of a Bayesian model through the expected log-likelihood for the current observations, corrected for the effective parametric dimension of the proposed model. Here, as in Bernardo (1999), the expectation of the utility function taken with respect to the posterior distribution, has a similar role of providing a quantification of the increase in uncertainty when the complexity of the model structure is increased.

When the statistical consistency of the sought solution to model ranking problem is of concern, it should be reflected in the choice of the cost c_j. In regular cases where the parametric dimension of M_j is straightforward to quantify, the cost can be defined analogously to consistent penalized ML criteria, such as those introduced by Schwarz (1978), or Hannan and Quinn (1979). As noted by Corander (2003a,b), a careless choice of the cost function would lead to an asymptotically satisfactory behavior of the model assessment, while even worsening the performance for relatively small observed samples, as compared to asymptotic approximations to the logarithm of the marginal data distribution (2), such as the one introduced by Schwarz (1978). As a cautious strategy, slowly increasing cost functions, such as the one introduced in Hannan and Quinn (1979), and utilized by Corander (2003a,b), should be advocated as a choice of c_j.

The objective Bayesian approach to inference under a fixed model structure has been shown to often produce even very acceptable frequentist properties (e.g. Bernardo and Smith, 1994). Such handling of uncertainty can be seen as a general agreement concerning how to proceed under lack of knowledge about the parameters of a statistical model in any particular application. However, no generally applicable objective Bayesian method seem yet have appeared for the model learning situation considered here. This is not entirely surprising, as it may be argued, that a successful application of the Occam's razor principle (for a statistical introduction, see Madigan and Raftery, 1994) requires in general some subjective assessment, which is effectively intrinsic to the normative Bayesian approach. Such assessment may also be incorporated to the BEC criterion, most reasonably through the cost function, which can reflect the expected behavior of the model.

Since learning about the relative merits of the models in $\mathcal{M}$ is a very complicated issue under any realistic and interesting data analysis scenario, it is not feasible to assume that the utility structure could be *universally* calibrated, without any reference to the particular application at hand. Bernardo (1999) obtains such a calibration by restricting the attention to the simple case with only two nested models in $\mathcal{M}$ (although, for criticism of his calibration, see the discussion of the

paper). Similarly, there are several other methods (e.g. O'Hagan, 1995; Berger and Pericchi, 1996) where the calibration is based on the existence of a minimal training sample with respect to the largest model in $\mathcal{M}$. Such samples do not even exist for many regularly used models for discrete data (e.g. multinomial distribution), and unfortunately can exceed the size of the complete observed data set (an example of this is provided in the next section).

The decision theoretic framework defines the optimal model in $\mathcal{M}$ as the M_j obtained by solving

$$\arg\max_{j\in J} \bar{u}(p_j \mid \mathbf{x}), \tag{10}$$

thus, the model learning problem corresponds to the minimization of the predictive entropy within $\mathcal{M}$. Interpretation of the differences in the model utilities is also facilitated by comparing the relative utilities

$$\frac{\exp(\bar{u}(p_j \mid \mathbf{x}))}{\sum_{j\in J} \exp(\bar{u}(p_j \mid \mathbf{x}))}. \tag{11}$$

The values of the BEC criterion can in general be computed by using Markov Chain Monte Carlo (MCMC) simulation (see Robert and Casella, 1999). If the observations are independently and identically distributed (iid) given θ_j, we have

$$u(p_j, \theta_j) = n \int_{\mathcal{X}^*} p_j(y \mid \theta_j) \log p_j(y \mid \theta_j)\, dy, \tag{12}$$

where $\mathcal{X}^*$ is the sample space of a single observation y. By the strong law of large numbers we get

$$\frac{n}{m^*} \sum_{l=1}^{m^*} \log p_j\big(y^{(l)} \,\big|\, \theta_j\big) \underset{a.s.}{\to} u(p_j, \theta_j), \quad \text{when } m^* \to \infty, \tag{13}$$

where $y^{(l)}, l = 1, \ldots, m^*$ are iid samples from M_j with the fixed θ_j. Also, under the standard ergodicity assumption for MCMC,

$$\frac{1}{m} \sum_{i=1}^{m} u\big(p_j, \theta_j^{(i)}\big) \to \int_{\Theta_j} u(p_j, \theta_j) \pi(\theta_j \mid \mathbf{x})\, d\theta_j \tag{14}$$

where $\theta_j^{(i)}, i = 1, \ldots, m$ is a sequence from $\pi(\theta_j \mid \mathbf{x})$, obtained from an MCMC simulation. Thus we can approximate the key value $\bar{u}(p_j \mid \mathbf{x})$ as

$$\bar{u}(p_j \mid \mathbf{x})^* = \frac{1}{m} \sum_{i=1}^{m} \frac{n}{m^*} \sum_{l=1}^{m^*} \log p_j\big(y^{(l)} \,\big|\, \theta_j^{(i)}\big), \tag{15}$$

where $y^{(l)}, l = 1, \ldots, m^*$ and $\theta_j^{(i)}, i = 1, \ldots, m$ are as above. Clearly, as $m, m^* \to \infty$, $\bar{u}(p_j \mid \mathbf{x})^* \to \bar{u}(p_j \mid \mathbf{x})$. Notice that the above double sum need not be numerically prohibitive in practice, since generation of $y^{(l)}$ is often computationally inexpensive given fixed values of the model parameters.

It is illuminating to consider the behavior of the utility function under a less general setting, while assuming the conditional independence. Let the likelihood $p_j(y \mid \theta_j)$ belong to the exponential family, such that the density of the sufficient statistic $t = (T_1(y), \ldots, T_k(y))$ can be written as

$$c(\theta_j) \exp\left(\theta_j^T t\right), \tag{16}$$

where θ_j is the natural parameter. Then, using Theorem (2.64) in Schervish (1995), it follows that the integral in (12) can be written as:

$$u(p_j, \theta_j) = \log c(\theta_j) - \sum_{l=1}^{k} \theta_j^{(l)} \frac{\partial}{\partial \theta_j^{(l)}} \log c(\theta_j), \tag{17}$$

where $\theta_j^{(l)}$ is now the lth element of θ_j. For any value of θ_j, the above is the expected log-likelihood for a future set of n exchangeable observations, and in particular, when θ_j equals the ML-estimate $\hat{\theta}_j$ given $\mathbf{x}$, the utility equals the maximized log-likelihood for the *current* data.

For a wide range of models in the exponential family, the posterior expectation of the utility function can be calculated analytically under certain reference type of priors. The asymptotic behavior of the BEC criterion is tractable for such cases, and in the next section we provide some illuminating examples.

4. Examples and Discussion

Perhaps one the simplest possible examples of model comparison within the exponential family is concerned with two univariate normal distributions: $\mathcal{M} = \{p_1 = N(0, \sigma^2), p_2 = N(\mu, \sigma^2)\}$. Using the canonical parametrization, the components of θ for p_2 are $\theta_1 = \mu/\sigma^2, \theta_2 = -1/(2\sigma^2)$. Given n observations, the relevant statistics for the two models are $n\tilde{s}^2 = \sum_{i=1}^n x_i^2, ns^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, respectively. The cumulant generating function for the latter model is

$$n\left(-\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log\left(-\frac{\theta_2}{\pi}\right)\right) \tag{18}$$

The negative entropies for the two models are equal, and given by

$$\frac{n}{2}\left(\log\left(-\frac{\theta_2}{\pi}\right)\right) - \frac{n}{2} \tag{19}$$

Let $\lambda = -2\theta_2$. The reference prior for λ given in Bernardo and Smith (1994) is $\pi(\lambda) \propto \lambda^{-1}$ for both models, and the reference posteriors are the following Gamma distributions $Ga(n/2, n\tilde{s}^2/2)$ and $Ga((n-1)/2, ns^2/2)$, respectively. Standard probability calculus yields the expectation (9) in a closed form, and it is illuminating to consider the difference between the negative expected entropies, which equals

$$\frac{n}{2}\left(\log(\tilde{s}^2) - \log(s^2) + \psi\left(\frac{n-1}{2}\right) - \psi\left(\frac{n}{2}\right)\right) \tag{20}$$

where $\psi(\cdot)$ is the digamma function. The first part is recognized as the maximized log-likelihood ratio, and the strictly negative difference $\psi((n-1)/2) - \psi(n/2)$ reflects the increased uncertainty in the posterior caused by the inclusion of the mean parameter.

In Figure 1, we illustrate the error rate of the BEC criterion with particular choices of c_j as a function of n ($n = 2, \ldots, 30$), assuming p_1 to be the true model. For comparison, the AIC (Akaike, 1974) and SBC (Schwarz, 1978) criteria have also been included. It is seen, that the error rate for BEC with $c_j = d_j/2$ is nearly constant over n, reflecting the fact that the posterior expectation takes the curvature in the likelihood into account. Indeed, from the frequentist point of

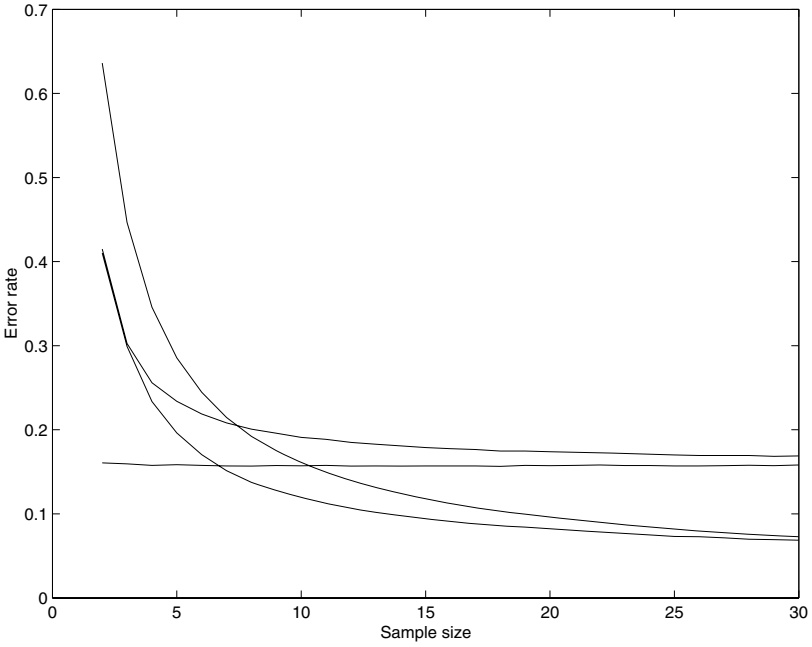

Figure 1. Empirical error rates (over 1000000 replicates) for the univariate normal example as function of n. From top to bottom (at right vertical axis) curves correspond to: (1) AIC, (2) BEC with $c_j = d_j/2$, (3) BEC with $c_j = d_j \log\log n$, (4) SBC.

view, our approach behaves similarly to a small sample correction to the asymptotic likelihood ratio test as in Porteous (1985). As expected, error rate for the choice $c_j = d_j \log\log n$ (Hannan and Quinn, 1979; Corander, 2003a,b) tends to zero as n increases, and approaches that of the SBC criterion.

Simply choosing the best model from $\mathcal{M}$ is not as such in general a comprehensive strategy for the model learning problem, when the utility of the best model is relatively close to that of any other model. The relative expected utilities can therefore be pursued to obtain an intuitive interpretation of the plausibility of the best model. In the normal example, we may consider the empirical probability that the relative utility would yield strong evidence for the incorrect model, by investigating the probability that (11) exceeds .1 for the model p_2. This is illustrated in Figure 2 for BEC with $c_j = d_j/2$ and $c_j = d_j \log\log n$, respectively. Again, as expected, it is seen that the probability of this event is fairly small.

As a more realistic example, consider now the multivariate linear regression model (see, e.g. Mardia et al., 1979) defined by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \tag{21}$$

where $\mathbf{Y}(n \times p)$ is an observed matrix of p response variables in each of n observations, $\mathbf{X}(n \times q)$ is a known matrix, $\mathbf{B}(q \times p)$ is a matrix of unknown regression parameters and $\mathbf{U}$ is a matrix of unobserved random disturbances whose rows for

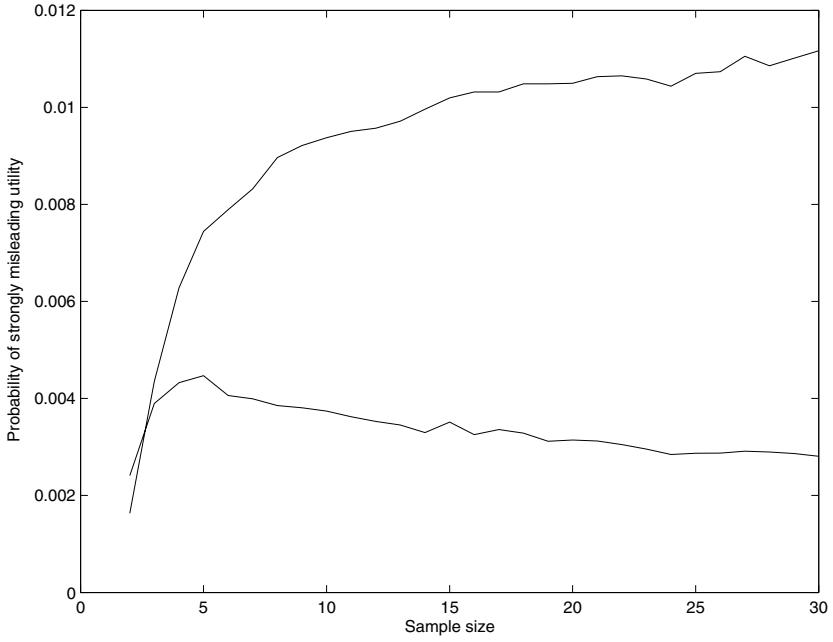

Figure 2. Empirical probability (over 1000000 replicates) that (11) exceeds .1 for the model p_2. Upper curve: BEC with $c_j = d_j/2$; lower curve: BEC with $d_j = d_j \log\log n$.

given $\mathbf{X}$ are uncorrelated, each with mean 0 and common covariance matrix Σ. We further assume that rows of $\mathbf{U}$ are from the normal distribution $N_p(0, \mathbf{\Sigma})$. Now we have the parameter $\theta = (B, \mathbf{\Sigma})$, and the negative entropy (7) for n future observations is given as

$$u(p_j, \theta_j) = n \int_{\mathbb{R}^p} \Pr(Y = y \mid \mathbf{B}, \mathbf{\Sigma}) \log \Pr(Y = y \mid \mathbf{B}, \mathbf{\Sigma})\, dy \tag{22}$$

$$= n \int_{\mathbb{R}^p} \Pr(U = y - x\mathbf{B} \mid \mathbf{B}, \mathbf{\Sigma}) \log \Pr(U = y - x\mathbf{B} \mid \mathbf{B}, \mathbf{\Sigma})\, dy. \tag{23}$$

From the standard probability calculus, it follows that (23) is simply n times the negative entropy of a variable U with distribution $N_p(0, \mathbf{\Sigma})$. Thus, we can write

$$u(p_j, \theta_j) = \frac{n}{2}(p + p \log(2\pi) + \log |\mathbf{\Sigma}|). \tag{24}$$

Next we must specify the joint prior distribution π for the parameters $\mathbf{B}$ and $\mathbf{\Sigma}$. We assume that the elements of $\mathbf{B}$ and those of $\mathbf{\Sigma}$ are independently distributed; that is,

$$p(\mathbf{B}, \Sigma) = p(\mathbf{B}) p(\mathbf{\Sigma}). \tag{25}$$

In (25) we take the Jeffreys' prior (Bernardo and Smith, 1994)

$$p(\mathbf{B}) \propto \text{constant} \tag{26}$$

and

$$p(\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(p+1)/2}. \tag{27}$$

With the specified prior the marginal posterior distribution of $\mathbf{\Sigma}$ is computed in Zellner (1971) and is equal to the inverted Wishart distribution $W_p^{-1}(S^{-1}, \upsilon)$, where υ is the degrees of freedom, and

$$S = (Y - X\hat{B})^T (Y - X\hat{B}), \tag{28}$$

and

$$\hat{B} = (X^T X)^{-1} X^T Y, \tag{29}$$

is a matrix of least squares quantities. Given Lemma 5.1 and Proposition 5.2 in Corander (2003a), we can now integrate (24) with respect to the posterior, which gives

$$\bar{u}(p_j \mid \mathbf{x}) = \frac{n}{2}\left(p + p \log \pi + \log |S^{-1}| - \sum_{i=0}^{p-1} \psi((\upsilon - i)/2) \right). \tag{30}$$

It is thus seen that BEC is available in a closed form for the class of normal linear models, which have been extensively considered in the papers concerned with automated Bayesian model selection.

Using the Proposition 5.2 in Corander (2003a), the expectation (9) may be derived for any multivariate linear model with Gaussian iid errors, where the reference inverse Wishart posterior is available in a closed form. Such modeling setting includes the predictor choice in the standard linear multivariate regression model (even with several response variables), order determination for vector autoregressive models, and learning of graphical Gaussian models (Giudici and Green, 1999). Analogously to the criterion of Corander (2003a,b), BEC can be applied to learning of graphical log-linear models for multinomial data. Also, BEC can be derived in a closed form for unsupervised Bayes classifiers based on multivariate Gaussian mixtures, using the data partition model formulation of Corander et al. (2005).

We now turn to an example of a model learning scenario, for which the subjective Bayesian approach is extremely complicated, and for which the earlier referred automated methods may easily fail.

The likelihood for a decomposable graphical Gaussian model G for a finite set V of variables, may be written in the factored form

$$\frac{\prod_{c\in\mathcal{C}} p(x_c \mid \theta_c)}{\prod_{s\in\mathcal{S}} p(x_s \mid \theta_s)}, \tag{31}$$

where $\mathcal{C}$ and $\mathcal{S}$ are the sets of cliques and separators in G, respectively (for details, see Lauritzen, 1996). Under (31), the posterior factorizes, and consequently, the expectation of (9) may be calculated separately for each term. Letting $a \subseteq V$ be an arbitrary subset of the variables, the graphical Gaussian model specifies $p(x_a \mid \theta_a)$ as $N(\mathbf{0}, \mathbf{\Sigma}_a)$, where $\mathbf{\Sigma}_a$ is the covariance matrix of the corresponding variables. Given the reference prior $\pi(\mathbf{\Sigma}_a) \propto |\mathbf{\Sigma}_a|^{-(|a|+1)/2}$, where $|a|$ is the cardinality of a, the reference posterior is the inverse Wishart distribution $W^{-1}_{|a|}(\mathbf{S}_a, n)$, where $\mathbf{S}_a$ is the ML-estimate of $\mathbf{\Sigma}_a$. Proposition 5.2 in Corander (2003a) shows that the expected negative entropy for the density term corresponding to a can be written as

$$\begin{aligned}\bar{u}(p_a|\mathbf{x}) = {} & \frac{n|a|}{2}[1 + \log(2\pi) + \log n] \\ & + \frac{n|a|}{2}\left[\log|\mathbf{S}_a| - \log 2 - \sum_{i=1}^{|a|-1} \psi((n-1-i)/2)\right]\end{aligned} \tag{32}$$

which is similar to the univariate Gaussian case. Thus, the expected utility of graphical Gaussian model G equals

$$\sum_{c\in\mathcal{C}} \bar{u}(p_c \mid \mathbf{x}) - \sum_{s\in\mathcal{S}} \bar{u}(p_s \mid \mathbf{x}) - c_G, \tag{33}$$

where the cost c_G is a function of the parametric dimensionality of the model,

$$d_G = \sum_{c \in \mathcal{C}} \binom{|c|}{2} + |c| - \sum_{s \in \mathcal{S}} \binom{|s|}{2} + |s|. \tag{34}$$

The simplicity of the BEC approach in this particular case can be compared to the computationally intensive MCMC method, which was considered in Giudici and Green (1999). In this application, a minimal training sample (O'Hagan, 1995; Berger and Pericchi, 1996) must be defined with respect to the complete graph G, which easily leads to a training sample with size larger than n for extensive sets of variables V. Also, it is well known that the asymptotic model learning criteria lead to coarse approximations for this type of situations.

For BEC, the expectation above is with respect to the posterior for any actual graphical model, not for the complete model as in Corander (2003a). Since for the inverse Wishart parametrization a change in the degrees of freedom is induced by marginalization, the unbiasedness of the Bayesian entropy estimate given by Proposition 5.2 in Corander (2003a), is in fact satisfied by BEC for each term in (33), but not by the approach of Corander (2003a) (except for the complete graph).

For the above type of exponential models, it follows from the behavior of the digamma function that the difference between the expected utility for two candidate models M_j, M_k, with $d_j > d_k$ tends to

$$\log p_j(\mathbf{x} \mid \hat{\theta}_j) - \log p_k(\mathbf{x} \mid \hat{\theta}_k) - \frac{d_j - d_k}{2} - (c_j - c_k), \tag{35}$$

as $n \to \infty$. Depending on the choice of the model cost c_j, the criterion can now be made asymptotically equivalent to several model selection criteria introduced in the literature. For the particular choices $c_j = \{d_j/2, d_j \log\log n, (d_j/2)\log n\}$, BEC tends to the Akaike's, Hannan and Quinn's, and Schwarz criteria, respectively. However, for small values of n, value of BEC may differ considerably from (35), since (9) takes into account the expected curvature in the log-likelihood function for the future data, and the curvature in the posterior distribution of θ_j. Clearly, the criteria based on the asymptotic approximation of the above type ignore the curvature and are linear in the penalty term with respect to the model dimension for any fixed sample size n. Notice that, under typical improper reference priors, if the model dimension d_j would be too extensive with respect to n, the corresponding posterior expectation of (7) would tend to minus infinity, thus automatically preventing the use of models for which there is not enough data to estimate the parameters.

As discussed in the previous section, calibration of the cost function c_j necessarily requires some subjective argument related to the characteristics of $\mathcal{M}$, however, for many standard model classes $c_j = d_j \log\log n$ seems to be a reasonable choice when statistical consistency is sought.

In the examples belonging to the regular exponential family which were considered above, the predictive negative entropy was expressible as the sum of the

maximized log-likelihood for the currently observed data, and a non-linear correction term depending on the sample size and on the number of parameters in the model. Our aim in future is to investigate properties of BEC further in important general model classes, such as the curved and stratified exponential families, and to obtain more explicit results concerning the calibration of the utility structure.

Acknowledgement

Research in this paper was financially supported by the funds of University of Helsinki, Finland. The authors would like to thank the anonymous referees for comments and suggestions that led to a significant improvement of the original manuscript.

References

Aitkin, M., 1991, "Posterior Bayes factors," *J. Roy. Statist. Soc.* **B53**, 111–142 (with discussion).

Akaike, H., 1974, "A new look at the statistical model identification," *IEEE Trans. Autom. Control* **19**, 716–723.

Akaike, H., 1978, "A new look at the Bayes procedure," *Biometrika* **65**, 53–59.

Akaike, H., 1979, "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika* **66**, 237–242.

Bayarri, M. J. and Berger, J., 1998, "Robust Bayesian analysis of selection models," *Ann. Statist.* **26**, 645–659.

Berger, J.O. and Pericchi, L.R., 1996, "The intrinsic Bayes factor for model selection and prediction," *J. Amer. Stat. Assoc.* **91**, 109–122.

Berger, J.O. and Bernardo, J.M., 1989, "Estimating a product of means: Bayesian analysis with reference priors," *J. Amer. Stat. Assoc.* **84**, 200–207.

Berger, J.O. and Bernardo, J.M., 1992, "On the development of reference priors," in J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), *Bayesian Statistics 4*, Oxford: Oxford University Press, pp. 35–60 (with discussion).

Berger, J.O. and Mortera, J., 1999, "Default Bayes factors for nonnested hypothesis testing," *J. Amer. Stat. Assoc.* **94**, 542–554.

Bernardo, J.M., 1979, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc.* B **41**, 113–147 (with discussion).

Bernardo, J.M., 1999, "Nested hypothesis testing: The Bayesian reference criterion," in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 6*, Oxford: Oxford University Press, pp. 101–130 (with discussion).

Bernardo, J.M. and Rueda, R., 2002, "Bayesian hypothesis testing: A reference approach," *Int. Stat. Review* **70**, 351–372.

Bernardo, J.M. and Smith, A.F.M., 1994, *Bayesian Theory*, Chichester: Wiley.

Corander, J., 2003a, "Bayesian graphical model determination using decision theory," *J. Multiv. Analysis* **85**, 253–266.

Corander, J., 2003b, "Labeled graphical models," *Scand. J. Stat.* **30**, 493–508.

Corander, J., Gyllenberg, M., and Koski, T., 2005, "Bayesian unsupervised classification algorithms based on parallel search strategy," *Patt. Recog.* (under revision).

Dawid, A.P., 1984, "Present position and potential developments: Some personal views. Statistical theory. The prequential approach," *J. Roy. Statist. Soc.* **A47**, 278–292 (with discussion).

Engel, Y., Mannor, S., and Meir, R., 2003, "Bayes meets Bellman: The Gaussian process approach to temporal difference learning," in T. Fawcett and N. Mishra (eds.), *Proceedings of the 20th International Conference on Machine Learning*, Washington D.C.: AAAI Press.

de Finetti, B., 1974, *Theory of Probability I*, Chichester: Wiley.

Giudici, P. and Green, P.J., 1999, "Decomposable graphical Gaussian model determination," *Biometrika* **86**, 785–801.

Gutiérrez-Peña, E. and Walker, S.G., 2001, "A Bayesian predictive approach to model selection," *J. Statist. Planning Inference* **93**, 259–276.

Hannan, E.J. and Quinn, B.G., 1979, "The determination of the order of an autoregression", *J. Roy. Statist. Soc.* **B41**, 190–195.

Jordan, M., 2004, Graphical models," *Stat. Sci.* **19**, 140–155.

Kass, R. and Wasserman, L., 1996, "The selection of prior distributions by formal rules," *J. Amer. Stat. Assoc.* **91**, 1343–1370.

Key, J.T, Pericchi, L.R., and Smith, A.F.M., 1999, "Bayesian model choice: What and why?" in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 6*, Oxford: Oxford University Press, pp. 343–370 (with discussion).

Lauritzen, S.L., 1996, *Graphical Models*, Oxford: Oxford University Press.

Lindley, D., 1991, "Discussion of paper by M. Aitkin," *J. Roy. Statist. Soc.* **B53**, 111–142 (with discussion).

Lindley, D., 1992, "Discussion of paper by R. Royall," in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 4*, Oxford: Oxford University Press, pp. 405–418 (with discussion).

Lindsey, J.K., 1996, *Parametric Statistical Inference*, Oxford: Oxford University Press.

Madigan, D. and Raftery, A.E., 1994, "Model selection and accounting for model uncertainty in graphical models using Occam's window," *J. Amer. Stat. Assoc.* **89**, 1535–1546.

Mardia, K.V., Kent, J.T. and Bibby, J.M., 1979, *Multivariate Analysis*, London: Academic Press.

Meir, R. and Merhav, N., 1995, "On the stochastic complexity of learning realizable and unrealizable rules," *Machine Learning* **19**, 241–261.

O'Hagan, A., 1995, "Fractional Bayes factors for model comparison," *J. Roy. Statist. Soc.* **B57**, 99–138 (with discussion).

Perez, J.M. and Berger, J., 2002, "Expected posterior prior distributions for model selection," *Biometrika* **89**, 491–512.

Porteous, B.T., 1985, "Improved likelihood ratio statistics for covariance selection models," *Biometrika* **72**, 97–101.

Rissanen, J., 1987, "Stochastic complexity," *J. Roy. Statist. Soc.* **B49**, 223–239.

Rissanen, J., 1995, "Fisher information and stochastic complexity," *IEEE Trans. Inf. Theory* **42**, 40–47.

Robert, C.P. and Casella, G., 1999, *Monte Carlo Statistical Methods*, New York: Springer.

Royall, R., 1992, "The elusive concept of statistical evidence," in J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), *Bayesian Statistics 4*, Oxford: Oxford University Press, pp. 405–418 (with discussion).

Schervish, M.J., 1995, *Theory of Statistics*, New York: Springer-Verlag.

Schwarz, G., 1978, "Estimating the dimension of a model," *Ann. Stat.* **6**, 461–464.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A., 2002, "Bayesian measures of model complexity and fit," *J. Roy. Statist. Soc.* **B64**, 583–640 (with discussion).

Weissman, T. and Merhav, N., 2003, "On competitive predictability and its relation to rate-distortion theory and to channel capacity theory," *IEEE Trans. Inform. Theory* **49**, 3185–3194.

Zellner, A., 1971, *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.