Bayesian Computational Approaches to Model Selection

C. Andrieu, A. Doucet, W.J. Fitzgerald and J.-M. Pérez

1 Introduction

A fundamental task in signal and data processing and science is, in general, to develop models for signals which are observed and to determine whether the model function that one is using to describe the data is actually appropriate for the particular problem under investigation. Except in artificial problems, a model is just an approximation that, up to some degree, describes the process which generates a particular signal or set of observations. In this way, one usually entertains several plausible models, realizing that in terms of the real data generation process, the correct model may not be within the set chosen. Multiple model selection, therefore, appears naturally in the analysis of trying to determine which of the entertained models best describes the data at hand. Further, in parametric models, one is also interested in extracting values for the free parameters of the model.

The problems of parameter estimation and model selection can be coherently approached within a Bayesian framework. In Bayesian analysis, the statistical inference is obtained in the form of *posterior distributions*, which incorporate both the scientist's beliefs and the observations, in a well founded probabilistic framework. In particular, the model selection problem can be summarized by the posterior probability of each model. This distribution is meaningful, and certainly easier to interpret than, say, classical P-values.

Bayesian analysis is not without problems, however. In practice, one is forced to establish prior beliefs, in the form of *prior probability distributions*, on the models under consideration and, often more difficult, on the parameters in each model. The latter could prove a daunting task, as the characteristics of many model parameters may well not be known precisely. One other problem one faces in the Bayesian framework is the computation of the quantities that lead to Bayesian model selection – typically integrals of large dimension that do not admit any closed-form analytical solution.

The problem of choosing prior distributions for the parameters is a very delicate one in the framework of model selection as illustrated in Section 2, where we briefly present state-of-the-art approaches to address these difficulties. In Section 3, we describe numerical methods that address the practical computation of the quantities of interest, namely the Bayes Factors and posterior model probabilities. We focus on Markov chain Monte Carlo (MCMC)

methods and especially the reversible jump MCMC method introduced by Green [33]. In Section 4 we present two applications of the methodology to the detection of sinusoids in noise and the determination of the components of Gaussian mixture models.

2 Bayesian model selection

2.1 Bayesian methodology for model selection

Assume that we are analyzing data[1] $\mathbf{y}$ and we believe that the data arise from one of a set of possible models $\mathcal{M}_0, \ldots, \mathcal{M}_{k_{\max}}$ ($k_{\max}$ can be infinite), where under model $\mathcal{M}_i$, $\mathbf{y}$ has density $p_i(\mathbf{y}|\theta_i)$, conditional on $\theta_i \in \Theta_i$. The parameter vectors θ_i are unknown and are typically of different dimension. Let $p_i(\theta_i)$ denote the prior density for θ_i (with respect to a dominating measure, usually Lebesgue), and let p_i denote the prior probability of the model $\mathcal{M}_i$. For the sake of convenience, we introduce a random variable $k \in \{0, \ldots, k_{\max}\}$ such that $\Pr(k = i) = \Pr(\mathcal{M}_i) = p_i$. The prior probability distribution for the random parameters (k, θ) is defined on a space of the form $\Theta \triangleq \bigcup_{i=0}^{k_{\max}} \{i\} \times \Theta_i$ and can be written

$$p(k, d\theta) = \sum_{i=0}^{k_{\max}} p_i(i, d\theta_i)\, \mathbb{I}_{\{i\}\times\Theta_i}(k, \theta), \tag{2.1}$$

where the integrable and distinct functions p_i admit the form

$$p_i(i, d\theta_i) = p_i(\theta_i)\, d\theta_i p_i \tag{2.2}$$

and

$$\mathbb{I}_{\{i\}\times\Theta_i}(k, \theta) = \begin{cases} 1, & \text{if } (k, \theta) \in \{i\} \times \Theta_i, \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

i.e. (k, θ) is in one of the spaces $\{i\} \times \Theta_i$, and the prior probability of k being equal to i and for θ being in an infinitesimal set centered around θ_i is $p_i(i, \theta_i)\, d\theta_i$.

After observing $\mathbf{y}$, one obtains the posterior distribution using Bayes' theorem

$$p(k, d\theta|\mathbf{y}) = \sum_{i=0}^{k_{\max}} p(i|\mathbf{y})\, p_i(d\theta_i|\mathbf{y})\, \mathbb{I}_{\{i\}\times\Theta_i}(k, \theta) \tag{2.4}$$

where $p(i|\mathbf{y})$ is the posterior probability of model $\mathcal{M}_i$ and is given by

$$p(i|\mathbf{y}) \triangleq p(\mathcal{M}_i|\mathbf{y}) = \frac{m_i(\mathbf{y})\, p_i}{\sum_{j=0}^{k_{\max}} m_j(\mathbf{y})\, p_j}, \tag{2.5}$$

[1] We do not distinguish between random variables and their realizations.

where

$$m_i(\mathbf{y}) \triangleq p(\mathbf{y}|i) = \int_{\Theta_i} p(\mathbf{y}|\theta_i)\, p_i(\theta_i)\, d\theta_i \tag{2.6}$$

is called the *marginal* distribution of $\mathbf{y}$ under model $\mathcal{M}_i$. Assuming $\mathcal{M}_i$ is the *true* model, $p(\mathbf{y}|i)$ is the density according to which $\mathbf{y}$ will actually occur. For this reason, $m_i(\mathbf{y})$ is also called the *predictive* density of $\mathbf{y}$. Under a 0-1 loss function, the optimal model is that $\mathcal{M}_k$ which maximizes the posterior model probability $p(k|\mathbf{y})$, $k = 0, \ldots, k_{\max}$.

Note that $p(k|\mathbf{y})$ can be written as

$$p(k|\mathbf{y}) = \left(1 + \sum_{i \neq k} \frac{p_i}{p_k} B_{ik}\right)^{-1}, \tag{2.7}$$

where the factor

$$B_{ik} = \frac{m_i(\mathbf{y})}{m_k(\mathbf{y})} \tag{2.8}$$

is called the *Bayes Factor* of model $\mathcal{M}_i$ against $\mathcal{M}_k$. Intuitively, the Bayes Factor can be interpreted as the odds of $\mathcal{M}_i$ against $\mathcal{M}_k$ given by the observations. Note that Bayes Factors can be used to summarize the analysis independently of the model prior beliefs, p_i.

The Bayesian approach to model selection can be applied to a wide variety of problems, including multiple comparisons and the testing of non-nested hypotheses. The results are easily interpreted (as opposed to frequentist P-values) and automatically penalize overparametrizations [12], [61]. For a detailed discussion of the advantages and applications of Bayes Factors see [7], [36].

We present two classical model selection problems arising in signal processing and statistics.

Example 2.1 *Sinusoids in noise.* We would like to model the data $\mathbf{y} \triangleq \{y_0, \ldots, y_{T-1}\}$ with one of the following models:

$$\begin{array}{ll} \mathcal{M}_0: & y_t = w_{0,t}, \qquad \text{if } k = 0, \\ \mathcal{M}_k: & y_t = \sum_{j=1}^{k} \left(a_{c_{j,k}} \cos[\omega_{j,k} t] + a_{s_{j,k}} \sin[\omega_{j,k} t]\right) + \varepsilon_{k,t} \quad \text{if } k \geq 1, \end{array} \tag{2.9}$$

where $\varepsilon_{k,t} \overset{iid}{\sim} \mathcal{N}(0, \sigma_k^2)$. The model $\mathcal{M}_k$ describes the data in terms of k sinusoids in white Gaussian noise. The unknown parameters for $\mathcal{M}_k$ are $\theta_k = \left(a_{c_{1,k}}, a_{s_{1,k}}, \omega_{1,k}, \ldots, a_{c_{k,k}}, a_{s_{k,k}}, \omega_{k,k}, \sigma_k^2\right)$. We added a subscript k to emphasize that these parameters depend on the model $\mathcal{M}_k$. Bayesian inference is performed on the parameter space $\Theta = \bigcup_{i=0}^{k_{\max}} \{i\} \times \Theta_i$ where $\Theta_i = \left(\mathbb{R}^2 \times (0, \pi)\right)^i \times \mathbb{R}^+$, *i.e.* if $k = i$, the unknown parameters θ_k belongs to $\Theta_{i=k}$. The space **Theta** is a union of disjoint spaces. In this case, one says that the models are nested since $\mathbf{Theta}_{i+1} = \mathbb{R}^2 \times (0, \pi) \times \mathbf{Theta}_i$ for $i = 0, \ldots, k_{\max - 1}$.

Example 2.2 *Mixture of normals.* The data $\mathbf{y} \triangleq \{y_1, \ldots, y_T\}$ are assumed to be distributed according to one of the following models:

$$\mathcal{M}_k : \mathbf{y}_t \overset{iid}{\sim} \sum_{j=1}^{i} w_{j,k} \mathcal{N}(\mu_{j,k}, \Sigma_{j,k}).$$

The model $\mathcal{M}_k$ describes the data in terms of a mixture of k Gaussian distributions, where $k \in \{1, \ldots, k_{\max}\}$ represents the unknown number of components. Conditional on k, the weights of the mixture are given by $\mathbf{w}_k = (w_{1,k}, \ldots, w_{k,k})$, where $w_{j,k}$ is the probability of an observation coming from component j and $\sum_{j=1}^{i} w_{j,k} = 1$. The parameters of the model components are given for $\mathcal{M}_k$ by $\phi_k = (\mu_k, \Sigma_k)$ where $\mu_k = (\mu_{1,k}, \ldots, \mu_{k,k})$ and $\Sigma_k = (\Sigma_{1,k}, \ldots, \Sigma_{k,k})$, with $\mu_{j,k}$ and $\Sigma_{j,k}$ being respectively the mean and covariance matrix of the kth component. The unknown parameters for $\mathcal{M}_k$ are $\theta_k = (\mathbf{w}_k, \phi_k)$. Bayesian inference is performed on the parameter space $\Theta = \bigcup_{i=1}^{k_{\max}} \{i\} \times \Theta_i$. In this case, the models are also nested.

In practice, such as in the two previous examples, Bayesian analysis requires specification of prior distributions on the parameter spaces. Unfortunately, it is usually impossible, or sometimes not desirable, to specify distributions on some or all model parameters. In such cases, one may alternatively[2] consider 'non-informative' or 'default' priors, which are discussed in the next section.

2.2 Specification of default prior distributions

The need for automatic or *default* approaches has been recognized for a long time [35]. In estimation problems, the use of vague or 'non-informative' prior distributions, including sometimes improper prior distributions[3], is typically a satisfactory solution [5], [55]. When performing model selection, however, one has to be much more careful because default priors are typically improper, and, thus, depend on arbitrary multiplicative constants, *i.e.*, $p_i^{\mathrm{N}}(\theta_i) = c_i f_i^{\mathrm{N}}(\theta_i)$. (We use the superscript N to indicate the use of a *non-informative* or *default prior* for the model parameters.) Hence, the resultant Bayes Factor

$$B_{ji}^{N} = \frac{c_j}{c_i} \frac{\int_{\Theta_j} p(\mathbf{y}|\theta_j)\, p_j^N(\theta_j) d\theta_j}{\int_{\Theta_i} p(\mathbf{y}|\theta_i)\, p_i^N(\theta_i) d\theta_i} \tag{2.10}$$

is indeterminate, and cannot be used for model selection.

A number of proposals to overcome this problem have been made. Approaches using conventional priors were studied in [35], [67]. In the case of nested models, proper hierarchical robust prior models have been successfully developed for various applications [2], [53], [58]. Other approaches include the Intrinsic Bayes Factor [7], the Fractional Bayes Factor [45] and the method

[2] Another alternative is given by Robust Bayesian Analysis (see, for example, [4]).
[3] A density $p(\theta)$ is improper iff $\int p(\theta)\, d\theta = +\infty$.

suggested in [61], among others. Most of these later methods deal with the problem by rescaling the Bayes Factor by a *correction factor* in such a way that any normalizing constants would be canceled. We briefly discuss these methods in the following sections as well as a new alternative method called Expected Posterior Prior distributions [47, 48].

2.2.1 Conventional priors

One possible way to obtain a default model selection criterion is to choose a *conventional* prior according to the problem at hand. The term 'conventional prior' is used to refer to a prior which is agreed upon as reasonable for the problem [37].

Conventional prior approaches for tests concerning the mean of a normal model have been proposed in [35] and [67]. For the comparison of models $\mathcal{M}_1 : \mathcal{N}(0, \sigma_1^2)$ vs. $\mathcal{M}_2 : \mathcal{N}(\mu, \sigma_2^2)$, Jeffreys utilizes the priors $p_1^N(\sigma_1) \propto \sigma_1^{-1}$ and $p_2^N(\mu, \sigma_2) = p_1^N(\sigma_2) p_{22}(\mu|\sigma_2)$ for models $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively, where p_{22} is a proper density function. Jeffreys argued that p_{22} should be an even function such that $\int p_{22}(\mu|\sigma_2)\mu^{n-1}d\mu$ diverges for any $n > 1$. He proposed the Cauchy density with scale σ_2 as the simplest density to meet these requirements.

For tests of hypotheses concerning multivariate normal means, Zellner and Siow [67], largely based on Jeffreys' work, suggested using a multivariate Cauchy density in place of p_{22}. The authors applied this idea to typical linear regression testing problems with orthogonal covariates, and provided Laplace approximations for the Bayes Factor in such cases.

2.2.2 Hierarchical models

Another possible solution consists of using a hierarchical model. Indeed, in many cases the prior distribution of the parameter θ_k further depends on a set of hyperprior parameters η. It is possible to add an extra layer in the prior distribution structure, which therefore models the uncertainty on the distribution of θ_k through a prior on η. The model for the prior density now takes the form

$$p(k, \theta_k, \eta) = p(k, \theta_k|\eta)\, p(\eta) \tag{2.11}$$

and the parameters η either become part of the inference problem, or may be integrated out as nuisance parameters, defining the prior for θ_k as

$$p(k, \theta_k) = \int p(k, \theta_k|\eta)\, p(d\eta)\,. \tag{2.12}$$

Typically the prior for η will be a *vague* or *non-informative* (proper) prior so that its influence on the model selection is minimum, leading to a robust estimation. This approach has been applied for example in [53] for Gaussian mixture models, where vague, data dependent hyperpriors are introduced to model the uncertainty on the unknown component variances and locations of the mixture.

The hierarchical prior approach will be further illustrated in the present chapter in Section 4 for the problem of detection of sinusoids in noise.

2.2.3 Spiegelhalter–Smith Bayes Factor

Spiegelhalter and Smith [61] used the device of *imaginary training samples* in the context of linear model comparisons to choose the constant c_j/c_i in (2.10).

Assume that an imaginary training sample, $\mathbf{y}_0^*$, is available such that

1. it is of *minimal size*, *i.e.*, it involves the smallest possible sample size that allows comparing $\mathcal{M}_i$ against $\mathcal{M}_j$ under non-informative priors;
2. it provides a maximum possible support for the simpler model, say, $\mathcal{M}_i$.

The authors argued that, for such a training sample, the Bayes Factor of $\mathcal{M}_j$ against $\mathcal{M}_i$ should be equal to one. Under this assumption, it follows that

$$\frac{c_i}{c_j} = \left\{ \frac{\int_{\Theta_i} p\left(\mathbf{y}_0^* | \theta_i\right) p_i^N\left(\theta_i\right) d\theta_i}{\int_{\Theta_j} p\left(\mathbf{y}_0^* | \theta_j\right) p_j^N\left(\theta_j\right) d\theta_j} \right\}^{-1}. \tag{2.13}$$

The Spiegelhalter–Smith Bayes Factor for observations $\mathbf{y}$ is now given by

$$B_{ji}^{SS} = B_{ji}^N \frac{\int_{\Theta_i} p\left(\mathbf{y}_0^* | \theta_i\right) p_i^N\left(\theta_i\right) d\theta_i}{\int_{\Theta_j} p\left(\mathbf{y}_0^* | \theta_j\right) p_j^N\left(\theta_j\right) d\theta_j}. \tag{2.14}$$

Note, however, that the choice of $\mathbf{y}_0^*$ depends on the models under comparison, and so there is no guarantee that B^{SS} is coherent for multiple model comparisons, *i.e.*, that $B_{ij}^{SS} \neq B_{ik}^{SS} B_{kj}^{SS}$.

2.2.4 Fractional Bayes Factors

The Fractional Bayes Factor, developed by O'Hagan [45], is based on using a fraction b of the likelihood, denoted here $L^b\left(\theta_i\right)$ for θ_i, to update non-informative priors. The Fractional Bayes Factor for model $\mathcal{M}_j$ against $\mathcal{M}_i$ is given by

$$B_{ji}^F = B_{ji}^N\left(\mathbf{y}\right) \frac{\int_{\Theta_i} L^b\left(\theta_i\right) p_i^N\left(\theta_i\right) d\theta_i}{\int_{\Theta_j} L^b\left(\theta_j\right) p_j^N\left(\theta_j\right) d\theta_j}. \tag{2.15}$$

One common choice of b is $b = m/T$, where m is the minimal training sample size and T the number of available data. The Fractional Bayes Factors are typically easier to compute than the Intrinsic Bayes Factors.

2.2.5 Intrinsic Bayes Factors

Consider two models $\mathcal{M}_1$ and $\mathcal{M}_2$ with data $\mathbf{y}$ having density $p\left(\mathbf{y} | \theta_i\right)$ and prior $p_i^N\left(\theta_i\right)$ under model $\mathcal{M}_i$. The *Arithmetic Intrinsic Bayes Factor* developed by Berger and Pericchi [7] is given by

$$B_{21}^{\mathrm{AIBF}} = B_{21}^N \frac{1}{L} \sum_{l=1}^{L} B_{12}^N\left(\mathbf{y}\left(l\right)\right), \tag{2.16}$$

where $B_{21}^N = B_{21}^N(\mathbf{y})$ is the usual Bayes Factor between models $\mathcal{M}_2$ and $\mathcal{M}_1$ and

$$m_i^N(\mathbf{y}) \triangleq \int_{\Theta_i} p(\mathbf{y}|\theta_i)\, p_i^N(\theta_i)\, d\theta_i, \tag{2.17}$$

the marginal under model $\mathcal{M}_i$. The sum is over all possible *minimal training samples*, $\mathbf{y}(l)$, of the observed data $\mathbf{y}$. A minimal training sample is defined as the minimum number of observations such that all marginals $m_i^N(\mathbf{y}(l))$ are positive and finite. Other versions of the IBF can be constructed by using averages rather than the mean in (2.16). For instance, using the geometric mean yields the *Geometric IBF* (B^{GIBF}). Choosing the median of Bayes Factors $B_{12}^N(\mathbf{y}(l))$ gives the *Median IBF* (B^{MIBF}), which appears to be more stable in cases where the sample size is relatively small. For a discussion of the Median IBF, see [10].

For the Arithmetic IBF, it is typically the case that $B_{ij}^{\text{IABF}} \neq 1/B_{ji}^{\text{IABF}}$. Thus, the AIBF is not suitable for multiple comparisons in its pure form. In [7] the authors suggest placing the more complex model, say $\mathcal{M}_j$, in the numerator and defining $B_{ij}^{\text{IABF}} = 1/B_{ji}^{\text{IABF}}$. Alternatively, one might define an *encompassing* model $\mathcal{M}_E$ such that every other model under consideration is nested in it. One can then obtain B_{Ei}^{IABF} for all i, and define the *Encompassing* IBF (based on $\mathcal{M}_E$) by $B_{ji}^{EIBF} = B_{Ei}^{\text{AIBF}}/B_{Ej}^{\text{AIBF}}$.

One appealing property of the arithmetic IBF is its asymptotic equivalence with a 'proper' Bayes Factor arising from *Intrinsic Priors*. By using a Schwartz approximation to the Bayes Factor one obtains

$$B_{21} = B_{21}^N \frac{p_2\left(\hat{\theta}_2\right) p_1^N\left(\hat{\theta}_1\right)}{p_2^N\left(\hat{\theta}_2\right) p_1\left(\hat{\theta}_1\right)} (1 + o(1)). \tag{2.18}$$

Equating this with (2.16) yields

$$\frac{p_2\left(\hat{\theta}_2\right) p_1^N\left(\hat{\theta}_1\right)}{p_2^N\left(\hat{\theta}_2\right) p_1\left(\hat{\theta}_1\right)} (1 + o(1)) = \frac{1}{L}\sum_{l=1}^{L} B_{12}^N(\mathbf{y}(l)) = \tilde{B}_{12}^N. \tag{2.19}$$

Assume

- Under $\mathcal{M}_1$: $\hat{\theta}_1 \to \theta_1$, $\hat{\theta}_2 \to \psi_2(\theta_1)$ and $\tilde{B}_{12}^N \to B_1^*(\theta_1)$;
- Under $\mathcal{M}_2$: $\hat{\theta}_2 \to \theta_2$, $\hat{\theta}_1 \to \psi_1(\theta_2)$ and $\tilde{B}_{12}^N \to B_2^*(\theta_2)$;
- For $i = 1$ or 2, the following exists:

$$B_i^*(\theta_i) = \lim_{L\to\infty} \mathbb{E}_{\mathcal{M}_i}\left[\frac{1}{L}\sum_{l=1}^{L} B_{12}^N(\mathbf{y}(l)) \middle| \theta_i\right], \tag{2.20}$$

where $\mathbb{E}_{\mathcal{M}_i}(\cdot)$ is the expectation with respect to the observations $\mathbf{y}(l)$ under model $\mathcal{M}_i$ and conditional upon θ_i.

If $\mathbf{y}(l)$ is exchangeable then the limit and average over l can be removed. Passing to the limit in (2.19), first under $\mathcal{M}_2$, and then under $\mathcal{M}_1$, results in the *Intrinsic Equations* for the Intrinsic Priors

$$\frac{p_2^I(\theta_2)\, p_1^N(\psi_1(\theta_2))}{p_2^N(\theta_2)\, p_1^I(\psi_1(\theta_2))} = B_2^*(\theta_2), \qquad \frac{p_2^I(\psi_2(\theta_1))\, p_1^N(\theta_1)}{p_2^N(\psi_2(\theta_1))\, p_1^I(\theta_1)} = B_1^*(\theta_1). \quad (2.21)$$

In the case where $\mathcal{M}_1$ is nested in $\mathcal{M}_2$, it can be shown that the Intrinsic Equations (2.21) have a solution of the form

$$\begin{aligned} p_2^I(\theta_2) &= p_2^N(\theta_2)\ \mathbb{E}_{\mathcal{M}_2}\left[\left.\frac{m_1^N(\mathbf{y})}{m_2^N(\mathbf{y})}\right| \theta_2\right], \\ p_1^I(\theta_1) &= p_1^N(\theta_1) \end{aligned} \quad (2.22)$$

(see, for example, [24]).

2.2.6 Expected Posterior Priors

Expected Posterior Priors have recently been proposed by Pérez and Berger [47], [48]. The method consists of adjusting the initial priors for each model by

$$p_i^*(\theta_i) = \int p_i^N(\theta_i|\mathbf{y}^*)\, m^*(\mathbf{y}^*) d\mathbf{y}^*, \quad (2.23)$$

where m^* is a suitable *predictive measure* on the *imaginary training sample* space and, as in the Intrinsic Bayes Factor, $\mathbf{y}^*$ is of minimal size. Several choices for m^* are possible. One choice for m^* that is attractive arises from selecting a *base model* $\mathcal{M}_*$ for the training sample and defining $m^*(\mathbf{y}^*) = m_*^N(\mathbf{y}^*)$. In the case of nested models, the Expected Posterior Priors resulting from this choice of m^* correspond to the Intrinsic Priors for the Arithmetic IBF. An empirical version of m^* can also be considered, where training samples are obtained by resampling from the observations.

The updated prior, p^*, is called the *Expected Posterior Prior* (or EP Prior) under m^*. The Bayes Factor of $\mathcal{M}_i$ against $\mathcal{M}_j$ resulting from the *Expected Posterior Priors* is given by

$$B_{ij}^*(\mathbf{y}) = \frac{m_{p_i^*}(\mathbf{y})}{m_{p_j^*}(\mathbf{y})}, \quad (2.24)$$

where

$$m_{p_i^*}(\mathbf{y}) \triangleq \int_{\Theta_i} p(\mathbf{y}|\theta_i)\, p_i^*(\theta_i)\, d\theta_i, \quad (2.25)$$

which can be written as

$$B_{ij}^*(\mathbf{y}) = \frac{m_{p_i^*}(\mathbf{y})}{m_{p_j^*}(\mathbf{y})} = \frac{\int m_i^N(\mathbf{y}|\mathbf{y}^*)\, m^*(\mathbf{y}^*)\, d\mathbf{y}^*}{\int m_j^N(\mathbf{y}|\mathbf{y}^*)\, m^*(\mathbf{y}^*)\, d\mathbf{y}^*}. \quad (2.26)$$

Therefore, the resulting Bayes Factor does not depend on arbitrary multiplicative constants and, therefore Bayesian model selection based on B_{ij}^* is coherent and allows for multiple comparisons.

The EP Prior approach admits several extensions. The key insight for the EP Prior is the use of $\mathbf{y}^*$ to update model parameters with improper priors. Note, however, that the improper priors p_i^N could be updated with any statistic $\psi_i(\cdot)$ of $\mathbf{y}^*$ and its associated 'likelihood', *i.e.*

$$p_i^N(\theta_i|\psi_i(\mathbf{y}^*)) = \frac{f_i(\psi_i(\mathbf{y}^*)|\theta_i)p_i^N(\theta_i)}{m^N(\psi_i(\mathbf{y}^*))}. \tag{2.27}$$

This gives rise to a more general definition of EP Priors. Suppose there exist statistics, $\psi_i(\mathbf{y}^*)$, for each model $\mathcal{M}_i$, such that $p_i^N(\theta_i|\psi_i(\mathbf{y}^*))$ is a proper density function and such that

$$0 < \mathbb{E}_{\mathcal{M}_i}\left[\frac{m^*(\mathbf{y}^*)}{m^N(\psi_i(\mathbf{y}^*))}\middle|\theta_i\right] < \infty, \tag{2.28}$$

for each model. Then EP Priors can be more generally defined as

$$p_i^*(\theta_i) = \int p_i^N(\theta_i|\psi_i(\mathbf{y}^*))m^*(\mathbf{y}^*)d\mathbf{y}^*. \tag{2.29}$$

Some other interesting properties of this approach are

- The resulting Bayesian inference is coherent and allows for multiple comparisons.
- In many cases, it is possible to find m^* such that, for a sample of minimal size, there is *predictive matching* for the comparisons of model $\mathcal{M}_i$ against $\mathcal{M}_j$, *i.e.*, the Bayes Factor $B_{ij} = 1$.
- In the case of nested models, where $\mathcal{M}_1$ is nested in every other model, choosing $m^*(\mathbf{y}^*)$ to be the marginal of $\mathbf{y}^*$ under $\mathcal{M}_1$ is asymptotically equivalent to the arithmetic Intrinsic Bayes Factor [7].

2.3 Practical issues

The previous subsections have outlined some of the 'theoretical' problems related to the definition of a Bayesian model to perform model selection, specially in the cases where improper, default priors are used. One further problem in the Bayesian framework relates to the computation of integrals, such as the Bayes Factors, which do not admit any closed-form expression, except for simple examples. Although numerical approximations were available, their application was often constrained by the high dimensionality of the integral involved. This problem had severely limited the application of Bayesian inference until the beginning of the 90's. Since the introduction of simulation methods, the Bayesian community has been able to computationally address more complex and large problems that were out of the scope before. The next

section focusses on Monte Carlo and MCMC simulation methods and their uses in solving the numerical problems often encountered in Bayesian analysis. We will discuss the applications of these methods to perform Bayesian detection of sinusoids in noise and Bayesian analysis of finite mixtures of normals in Section 4.

3 MCMC algorithms for Bayesian model selection

This section first presents the principles of Monte Carlo approximations for integrals of the type $\int f(\phi)\, p(d\phi)$. Here, $p(d\phi)$ is some probability distribution from which it is possible to draw i.i.d. samples. This can be difficult in practice. An often easier approach is the MCMC approach. The MCMC methodology for approximation of integrals is shown in detail in subsection 3.2, after recalling some results on the convergence of Markov chains admitting $p(d\phi)$ as invariant distribution.

3.1 Principle of Monte Carlo methods for integration

Consider a probability distribution $p(d\phi)$ defined on a general state space Φ. Assume that $M \gg 1$ samples $\left\{\phi^{(j)}; j = 1, ..., M\right\}$ distributed according to the distribution $p(d\phi)$ are available. Then, a Monte Carlo approximation $\widehat{P}_M(d\phi)$ of this distribution is given by the empirical estimate

$$\widehat{P}_M(d\phi) = \frac{1}{M}\sum_{j=1}^{M} \delta_{\phi^{(j)}}(d\phi), \tag{3.1}$$

where $\delta_{\phi^{(j)}}(d\phi)$ is the Dirac delta measure located at $\phi^{(j)}$. That is, the concentration of the samples in a given zone of the space Φ is assumed sufficiently representative of the probability of this zone under the distribution $p(d\phi)$.

Using this approximation for $p(d\phi)$, one can propose the following estimate f_M of $\mathbb{E}_{p(d\phi)}[f(\phi)]$, where $f : \Phi \rightarrow \mathbb{R}^{n_f}$ is a $p(d\phi)$-integrable function:

$$f_M = \int_{\Phi} f(\phi)\, \widehat{P}_M(d\phi) = \frac{1}{M}\sum_{j=1}^{M} f\left(\phi^{(j)}\right). \tag{3.2}$$

This estimate is unbiased and if the samples $\left\{\phi^{(j)}; j = 1, ..., M\right\}$ are statistically independent, then

$$\lim_{M \rightarrow +\infty} f_M \overset{a.s.}{\rightarrow} \mathbb{E}_{p(d\phi)}(f(\phi)) \tag{3.3}$$

from the strong law of large numbers. Moreover if

$$var[f_M] = \frac{1}{M}\left[\mathbb{E}_{p(d\phi)}(f^2(\phi)) - \mathbb{E}^2_{p(d\phi)}(f(\phi))\right] \triangleq \frac{\sigma_f^2}{M} < +\infty, \tag{3.4}$$