Euro Jnl Phil Sci (2013) 3:101–114
DOI 10.1007/s13194-012-0059-5

ORIGINAL PAPER IN PHILOSOPHY OF PROBABILITY

The role of Bayesian philosophy within Bayesian model selection

Jan Sprenger

Received: 16 January 2012 / Accepted: 1 July 2012 / Published online: 20 September 2012

Abstract Bayesian model selection has frequently been the focus of philosophical inquiry (e.g., Forster, Br J Philos Sci 46:399–424, 1995; Bandyopadhyay and Boik, Philos Sci 66:S390–S402, 1999; Dowe et al., Br J Philos Sci 58:709–754, 2007). This paper argues that Bayesian model selection procedures are very diverse in their inferential target and their justification, and substantiates this claim by means of case studies on three selected procedures: MML, BIC and DIC. Hence, there is no tight link between Bayesian model selection and Bayesian philosophy. Consequently, arguments for or against Bayesian reasoning based on properties of Bayesian model selection procedures should be treated with great caution.

Keywords Model selection · Bayesianism · Probability · BIC · MML

1 Introduction

Model selection is a relatively young subfield of statistics that compares statistical models on the basis of their structural properties and their fit to the data. The goal of model selection consists in comparing and appraising various candidate models on the basis of observed data.[1]

[1]In this paper, I understand "model selection" in a quite broad sense. That is, the statistical analysis need not lead to the *selection* of a particular model. More appropriate might be "model comparison", but I would like to stick with the traditional terminology.

J. Sprenger (✉)
Tilburg Center for Logic and Philosophy of Science (TiLPS),
Tilburg University, PO Box 90153, 5000, LE Tilburg, The Netherlands
e-mail: j.sprenger@uvt.nl
URL: www.laeuferpaar.de

Following up on Forster and Sober's seminal (1994) paper, the problem of model selection attracted much attention in philosophy of science. The properties of various model selection procedures have been used to argue for general theses in philosophy of science, such as the replacement of truth by predictive accuracy as an *achievable* goal of science (Forster 2002), the prediction/accommodation problem (Hitchcock and Sober 2004), the realism/instrumentalism dichotomy (Mikkelson 2006; Sober 2008), and the aptness of Bayesian reasoning for statistical inference (Forster 1995; Bandyopadhyay et al. 1996; Bandyopadhyay and Boik 1999; Dowe et al. 2007).

This paper explores the extent to which Bayesian model selection procedures are anchored within Bayesian philosophy, and in particular their philosophical justification. A model selection procedure is called "Bayesian" when it assigns prior and posterior probabilities to a parameter of interest. These probabilities are interpreted as rational degrees of belief (e.g., Bernardo and Smith 1994).

The classical, subjective view of Bayesian inference consists in reasoning from the prior to the posterior: high posterior probability becomes a measure of the acceptability of a hypothesis.[2] Scientific inference, including model selection, is based on this posterior distribution of beliefs. Accordingly, proponents of the Bayesian view of scientific rationality claim that "scientific reasoning is essentially reasoning in accordance with the formal principles of probability" (Howson and Urbach 1993, xvii)—see also Earman (1992, 142) and Talbott (2008).

However, such an orthodox subjective reading of Bayesianism is seldom put into practice. First, there is a plethora of practical and methodological problems, such as are the computational costs of calculating posterior distributions or handling nested models in a Bayesian framework. Second, when prior probabilities are assigned, reliable expert opinion is usually hard to elicit so that the choice of the prior is often dominated by mathematical convenience. Furthermore, results may be highly sensitive to the prior distribution. Third, even some Bayesian statisticians argue that their work is more guided by a focus on testing model adequacy than by genuinely subjective Bayesian belief revision (Gelman and Shalizi 2012).

Thus, the practice of Bayesian reasoning often differs from eliciting prior degrees of belief and updating them to posterior degrees of beliefs, as one may naïvely imagine. In the following sections, we analyze the foundations of three popular and much-discussed Bayesian model selection procedures—MML, BIC and DIC—in order to uncover the philosophical foundations of Bayesian model selection. As a result of this analysis, we conclude that these procedures are very diverse in the target which they aim at and the justification that they possess. Instead of conforming to the subjective Bayesian rationale,

[2]This is different from *objective* Bayesian inference where the two basic constraints of Bayesian inference—a coherent prior distribution and conditionalization on incoming evidence—are supplemented by further requirements that narrow down the set of rational degrees of belief, often up to uniqueness.

they are *hybrid* procedures: they do not primarily aim at an accurate representation of subjective uncertainty, but use the Bayesian calculus as a convenient mathematical tool for diverse epistemological goals. This has, as I shall argue in the conclusions, substantial repercussions on some bold methodological claims regarding Bayesian reasoning that are made in the literature.

2 MML and the conditionality principle

To avoid equivocations, I begin by fixing some terminology, following Forster (2002, S127). A statistical (point) *hypothesis* is a specific probability distribution from which the data may have been generated, e.g., the standard Normal distribution $N(0, 1)$. A statistical *model* refers, by contrast, to families of hypotheses, e.g. all Normal distributions of the form $N(\theta, \sigma^2)$ with parameter values $\theta \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{\geq 0}$.

For data $x_1, \ldots, x_N$, let us consider a candidate model $M \in \mathcal{M}$ with a respective set of parameters. A model selection criterion is a function of the data that assigns scores to point hypotheses or overarching models. On the basis of that score, the different models or point hypothesis can be compared, ranked or averaged. Quite often, we will identify point hypotheses with *fitted models*: namely when a particular hypothesis has been obtained by fitting parameters to the data. For example, a typical fitted model replaces the parameter values in the general Normal model $\langle N(\theta, \sigma^2), (\theta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^{\geq 0} \rangle$ by their maximum likelihood estimates on the basis of data x: the values $\hat{\theta}$ and $\hat{\sigma}^2$ such that for any other θ and σ^2: $p(x|\hat{\theta}, \hat{\sigma}^2) \geq p(x|\theta, \sigma^2)$, for probability density $p(\cdot)$ and data x. While some model selection procedures evaluate models in terms of expected predictive accuracy (e.g., Akaike 1973), others, typically classified as Bayesian, aim at the model with the highest posterior probability (e.g., Schwarz 1978).

Now we can turn to the Minimum Message Length (MML) principle (Wallace 2005; Dowe 2011). MML is a statistical inference procedure aiming at inferring the hypothesis ("theory")

> that allows the data to be stated in the shortest two-part message, where the first part of the message asserts the theory, and the second part of the message asserts the data under the assumption that the asserted theory is true. (Dowe et al. 2007, 717)

The basic idea is to infer the best explaining hypothesis, which is explicated as the explanation with the shortest expected message length in a probabilistic code. That is, the explanation has to trade off the plausibility of the hypothesis with the likelihood of the data under the hypothesis.

We illustrate this idea by means of an example (cf. Dowe et al. 2007, 721–722). Assume we want to estimate the parameter θ in a Binomial model $B(N, \theta)$, where X quantifies the number of successes in N trials. Then MML partitions the sample space $\mathcal{X} = \{0, \ldots, N\}$ into K interval sets $I_k = \{c_{k-1}, \ldots, c_k - 1\}$ with $c_0 = 0$ and $c_K = N + 1$. Let k_j be a weakly monotonic

Table 1 The optimal MML partitioning of the sample space (= the number of successes) into intervals I_{k_j} and the corresponding parameter estimates $\hat{\theta}_{k_j}$, for the case of the Binomial distribution $B(100, \theta)$ with a uniform prior

I_{k_j}	0	1–6	7–17	18–32	33–49	50–66	67–81	82–93	94–99	100
$\hat{\theta}_{k_j}$	0	0.035	0.12	0.25	0.41	0.58	0.74	0.875	0.965	1

See Wallace (2005, 157–160) and Dowe et al. (2007, 721–722)

sequence such that $j \in I_{k_j}$. Then, for each I_{k_j} we define a corresponding point estimate $\hat{\theta}_{k_j}$ of θ such that any $0 \leq j \leq N$ is mapped to $\hat{\theta}_{k_j}$.

Assuming a uniform prior over θ, the expected message length of estimator $\hat{\theta}$ is measured by the term

$$L := -\left(\sum_{j=0}^{N} p(X = j)\left(\log p(\hat{\theta}_{k_j}) + \log p(X = j|\hat{\theta}_{k_j})\right)\right). \tag{1}$$

In the case of $N = 100$, the optimal partition works with 10 different point estimates, see Table 1. Notably, the "natural" unbiased estimator X/N does not perform well on this count: the low prior probability of the associated intervals, which only consist of a singleton set, diminishes the overall score of X/N.

From a Bayesian point of view, the two components of L correspond to the two core components of Bayesian inference: the (log-)prior of the hypothesis (here: $\hat{\theta}_{k_j}$) and the (log-)likelihood of the data, given that hypothesis (here: $p(X = j|\hat{\theta}_{k_j})$).[3] MML proponents then argue that an inference to the theory that allows for the shortest two-part message will also be an inference to the most probable theory (or model), vindicating the use of Bayesianism in model selection, contra Forster and Sober (1994) and Forster (1995): "Bayes not Bust!" (Dowe et al. 2007).

However, since we measure *expected* total message length, the optimal tradeoff depends on the design of the experiment and in particular the sample size, cf. Eq. 1. This is actually admitted by the inventors of MML:

> The receiver of an explanation message is assumed to have prior knowledge on the set X of possible data, and the message is coded on that assumption. [...] The optimum explanation code requires than one assertion or estimate value serve for a range of distinct but similar possible data values. Hence, it seems inevitable that the assertion [=hypothesis] used to explain the given data will *depend to some extent on what distinct but similar possible data values might have occurred but did not*. (Wallace 2005, 254, my emphasis)

[3] Recall also that $\log P(H|E) \cdot P(E) = \log p(H) + \log P(E|H)$.

What is more, for the entire idea of the "shortest explanation", we have to choose between different conceptualizations of the *hypothesis space*, dependent on the chosen experimental design. This situation is in itself remarkable: while classical Bayesian reasoning considers the set of candidate models as fixed, MML aims at finding the partition of the hypothesis space that allows for the most efficient encoding of hypothesis and data.

These dependencies conflict, however, with subjective Bayesian epistemology, and one of its core principles, the Likelihood Principle:

> All the information about θ obtainable from an experiment is contained in the likelihood function $L_x(\theta) = p(x|\theta)$ for θ given x. Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another (Berger and Wolpert 1984, 19)

To see how closely the Likelihood Principle aligns with Bayesian inference, recall the identity

$$p(H|E) = \left(1 + \frac{p(\neg H)}{p(H)} \frac{p(E|\neg H)}{p(E|H)}\right)^{-1}$$

which is just another way of expressing Bayes's Theorem. From a Bayesian point of view, the likelihood function encompasses all relevant experimental information that is not already contained in the priors.

The Likelihood Principle demands in particular that the inference one draws do not depend on the space of possible outcomes, or on the sampling protocol. Whereas in an MML inference, the same data will lead to different best estimates of θ when obtained from a Binomial design or a Negative Binomial design, respectively.

At this point, one may doubt that the Likelihood Principle is compelling for a Bayesian statistician, so much the more as the wording chosen by Berger and Wolpert is admittedly vague. Therefore it is important to realize that it is actually equivalent to the conjunction of the following two principles:

Sufficiency	Let E be an experiment with a statistical model parametrized by $\theta \in \Theta$ and random variable X. If $T(X)$ is a sufficient statistic for θ, that is, if it satisfies $p(X = x\|T = t, \theta) = p(X = x\|T = t)$, and E_T is the experiment where any outcome x of E is represented by reporting the value $T(x)$, then E and E_T yield the same evidence about θ.
(Strong) Conditionality	If E is any experiment having the form of a mixture of component experiments E_i, then for each outcome (E_i, x_i) of E, [...] the evidential meaning of any outcome x of any mixture experiment E is the same as that of the corresponding outcome x_i of the

> experiment E_i which has actually been performed, ignoring the overall structure of the mixed experiment. (cf. Birnbaum 1962, 270–71)

Since the Sufficiency Principle is unanimously endorsed by both Bayesian and frequentist statisticians, we can focus our inquiry on the (Strong) Conditionality Principle. Informally, Conditionality can be described as asserting the irrelevance of experiments that were actually not performed. From a Bayesian point of view, this is eminently sensible; after all, the entire idea of Bayesian Conditionalization is based on taking into account (only) evidence that has actually occurred. Indeed, a lot of Bayesian arguments in statistical methodology and experimental design (e.g., with respect to optional stopping) are based on the soundness of that principle and explicitly on the irrelevance of which results *might* have been observed (cf. Royall 1997). So Conditionality and MML are directly at odds with each other.

The conflict is, by the way, known from other foundational debates in statistical inference. For instance, reference Bayesians such as Bernardo (2011) determine reference priors for Bayesian inference as a function of the sample space. Seen in that light, MML is somewhat typical for modern Bayesian statistics and its departure from Bayesian orthodoxy. It exemplifies a hybrid approach, where the Bayesian machinery is primarily a mathematical and conceptual toolbox for solving a specific problem whose definition does not depend on the Bayesian framework itself: determining the shortest explanation, the most efficient coding of theory and evidence. This is not meant to doubt that MML shares more with Bayesianism than related model selection criteria (e.g., Grünwald's (2005) Minimum Description Length principle). But it is interesting to see that an identity in formalism, and even an explicit appeal to Bayesian principles can still hide substantial philosophical differences.

3 Bayesianism without model priors: BIC

We now proceed to the next case study: Schwarz's Bayesian Information Criterion (BIC). The BIC is an estimation procedure that aims at the posterior probability of a parametric model M_θ, that is, at the weighted sum of the posterior probabilities of the hypotheses in M_θ that correspond to different values of θ. We will now reconstruct and analyze the motivation of BIC, following Schwarz (1978).

Assume that M_θ is one of our candidate models, whose elements are indexed by a parameter vector θ with model dimension K. We would like to approximate the posterior probability of M_θ. Assume further that all probability densities for data x (with respect to the Lebesgue measure μ) belong to the exponential family and that they can be written as

$$p(x|\theta) = e^{N(A(x)-\lambda|\theta-\hat{\theta}(x)|^2)}. \tag{2}$$

Here, $\hat{\theta}(x)$ denotes the maximum likelihood estimate of the unknown θ, and N the sample size, assuming i.i.d. sampling. This specific form of the likelihood function seems to make a substantial presumption, but in fact, the densities in Eq. 2 comprise the most familiar distributions, such as the Normal, Uniform, Fisher, Poisson and Student's t-distribution. For that reason, the assumption is plausible from a practical point of view.

Then we take a standard Bayesian approach and write the posterior probability of M_θ as proportional to the prior probability $p(M_\theta)$ and the averaged likelihood of the data x under M_θ:

$$\begin{aligned} p(M_\theta|x) \;&\sim\; p(M_\theta) \int_{\theta\in\Theta} e^{N(A(x)-\lambda|\theta-\hat{\theta}(x)|^2)} d\mu(\theta) \\ &= p(M_\theta)\, e^{NA(x)} \int_{\theta\in\Theta} e^{-N\lambda|\theta-\hat{\theta}(x)|^2} d\mu(\theta). \end{aligned}$$

Substituting the integration variable θ by $\theta/\sqrt{N\lambda}$, and realizing that for the maximum likelihood estimate $\hat{\theta}(x)$, $p(x|\hat{\theta}(x)) = e^{NA(x)}$, we obtain

$$\begin{aligned} \log p(M_\theta|x) &\sim \log p(M_\theta) + NA(x) + \log\left(\frac{1}{N\lambda}\right)^{K/2} + \log \int_{\theta\in\Theta} e^{-|\theta-\hat{\theta}(x)|^2} d\mu(\theta) \\ &= \log p(M_\theta) + NA(x) + \frac{1}{2}K \log\left(\frac{1}{N\lambda}\right) + \log \sqrt{\pi}^K \\ &= \log p(M_\theta) + \log p(x|\hat{\theta}(x)) - \frac{1}{2}K \log\left(\frac{N\lambda}{\pi}\right). \end{aligned} \tag{3}$$

Let us take stock. On the left hand side, we have the log-posterior probability, a subjective Bayesian's model comparison criterion. As we see from Eq. 3, this term is proportional to the sum of three terms: log-prior probability, the log-likelihood of the data under the maximum likelihood estimate, and a penalty proportional to the number of model parameters. This derivation, whose assumptions are relaxed subsequently in order to yield more general results, forms the mathematical core of BIC.[4]

In practice, it is difficult to elicit sensible subjective prior probabilities of the candidate models, and the computation of posterior probabilities involves high computational efforts. Therefore, Schwarz suggests to estimate log-posterior probability by a large sample approximation. For large samples, we neglect the terms in Eq. 3 that make only constant contributions and focus on the terms that increase in N: $\log p(M_\theta)$ drops out of the picture. Therefore, in the long run, the model with the highest posterior probability will be the model that minimizes

$$BIC(M_\theta, x) = -2 \log p(x|\hat{\theta}(x)) \;+\; K \log N. \tag{4}$$

[4]The number of parameters K enters the calculations because the expected likelihood of the data depends on the dimension of the model, via the skewness of the likelihood function.

BIC is intended to estimate the model (not the hypothesis) that accumulates, in the long run, the most posterior mass. However, it neglects the contribution of the priors when comparing the models to each other. Keeping in mind the identity

$$\log p(H|E) = \log p(H) + \log\left(p(E|H) \cdot \frac{1}{p(E)}\right) \tag{5}$$

wee see that BIC could as well be described as an approximation to the log-ratio measure of confirmation $\log p(H|E) - \log p(H)$ (up to addition of a constant).

Therefore, BIC should not be described as having a properly Bayesian justification: while (log-ratio) confirmation may be suitable for *comparing* models on the basis of the evidence (e.g., Milne 1996), it is not suitable for Bayesian *inference* since the priors drop out of the picture, as witnessed by the transition from Eqs. 3 to 4.

This finding is, by the way, in agreement with Schwarz' note that BIC extends "beyond the Bayesian context" (1978, 461).[5] Even more, frequentist properties are sometimes invoked in an attempt to justify the practical use of BIC (e.g., Burnham and Anderson 2002).

To further strengthen this conclusion, note that BIC is quite different from a numerical large sample approximation for posterior degrees of belief: the posterior approximated by BIC is detached from subjective prior probability. So BIC is not just a practical approximation to Bayesian coherence. Compare BIC to techniques such as Gibbs sampling or Monte Carlo Markov Chains (Han and Carlin 2001): those techniques aim at numerical approximations of subjective posterior distributions, and offer computational help for tricky multi-dimensional integrals. BIC develops a different philosophical rationale.

Neither does the *statistical consistency* of BIC provide a genuinely Bayesian justification. Here, consistency does not denote logical consistency with another proposition, but a certain long-run property of statistical estimators. That is, as sample size increases, the model favored by BIC converges in probability to the true model as long as the overall model is not misspecified. However, both Bayesians and frequentists regard consistency only as a *necessary* constraint on good estimators, not as a sufficient reason for using a particular method. So neither is consistency in any way peculiar to Bayesian inference, nor is it strong enough to make a case for BIC as opposed to other methods.

Our diagnosis that BIC lacks, in spite of the extensive use of Bayesian formalism, a fully Bayesian rationale, is supported by the variety of purposes to which the criterion is put. Sometimes it is regarded as an approximation to the Bayes factor (Kass and Raftery 1995). Raftery (1995) proposes an interpretation of BIC as an approximation to the integrated likelihood, which

[5]Forster and Sober (1994, 23–24) doubt, for quite different reasons, that Schwarz' Bayesian approach achieves a satisfactory solution to the model selection problem. Notably they also question "that it is securely grounded in the Bayesian framework."

is easily derived on the basis of the above calculations. Romeijn et al. (2012) see different worries with a Bayesian understanding of BIC and propose to anchor it more securely in Bayesian reasoning by taking into account the size of the parameter space. Hence, what is approximated by the asymptotic analysis of BIC is not determined by the mathematics themselves and depends on the general perspective one adopts.

4 Estimating effective complexity: DIC

In reply to the above diagnosis, it is sometimes objected that the distinction between Bayesian and frequentist model selection procedures should be made according to their inferential *targets* (Burnham and Anderson 2002, 2004). According to that proposal, even if the employed inferential strategies are not properly Bayesian at every step, as we have seen for MML and BIC, the target of inference—the posterior probability of a model or a fitted model—can only be formulated within a Bayesian framework. In support of this view, it is sometimes asserted that "Bayesians assess an estimator by determining whether the values it generates are probably true or probably close to truth" (Forster and Sober 2011, 535) or "the model selection literature often errs that AIC and BIC selection are directly comparable, as if they had the same objective target model" (Burnham and Anderson 2004, 299). That is, where frequentist methods, such as AIC, estimate the predictive performance of fitted models, Bayesian methods, such as BIC, estimate the posterior probability of a given model, or construct estimators that minimize mean error with respect to the posterior distribution. To show that this picture is misleading or at least incomplete, I conduct a further case study, namely on the Deviance Information Criterion (DIC).

The DIC is another model selection criterion that is commonly placed in the Bayesian family. Many model selection criteria, such as AIC and BIC, can be written and interpreted as an explicit tradeoff of goodness-of-fit and complexity. This is difficult in a specific context that we often encounter in practice: complex, hierarchical models (cf. Henderson et al. 2010). That is, when we represent the marginal distribution of the data x in a probability model as

$$p(x) = \int_{\theta \in \Theta} p(x|\theta)\, p(\theta)\, d\theta \tag{6}$$

with parameter θ and prior density $p(\theta)$, we may sometimes choose to represent that prior as being governed by a hyperparameter ψ:[6]

$$p(\theta) = \int_{\psi \in \Psi} p(\theta|\psi)\, p(\psi)\, d\psi. \tag{7}$$

[6]The marginal distribution of the data Eq. 6 is not affected by whether we parametrize the prior with hyperparameter ψ according to Eq. 7.

However, it is now unclear what should be considered the likelihood function of the data: $p(x|\theta, \psi)$, $p(x|\theta)$ or $p(x|\psi)$ (Bayarri et al. 1988)? Consequently, it is unclear how complexity of the model should be measured: Should we base our understanding of complexity on the dimension of θ, the dimension of ψ, or an aggregate of both? Apart from this ambiguity, the complexity of a model also depends on the amount of available prior information on the parameter values. The more information we have, the less complex a model is. Straightforwardly measuring complexity as the number of free parameters, as in the case of BIC, is therefore inappropriate as a general procedure.

Therefore, Spiegelhalter et al. (2002) propose to measure complexity by comparing the *expected deviance* in the data (under the posterior distribution) to the deviance in the estimate $\tilde{\theta}(x)$ that we would like to use. In other words, complexity manifests itself in terms of "difficulty in estimation". The authors propose to measure surprise or deviance in the data x relative to a point hypothesis parametrized by $\theta \in \Theta$ by means of the canonical measure $-\log p(x|\theta)$ (Bernardo 1979).[7] The Bayesian twist of DIC, as opposed to frequentist approaches, consists in incorporating prior information on the parameters: "it seems reasonable that a measure of complexity may depend on both the prior information concerning the parameters in focus and the specific data that are observed" (Spiegelhalter et al. 2002, 585).

In particular, $\tilde{\theta}(x)$ denotes the Bayes estimator of the quantity of interest θ, usually the posterior mean of θ. Then, we can compare the expected deviance in the data (conditional on the posterior distribution of θ) to the deviance we observe under our estimate of $\tilde{\theta}(x)$. This quantity p_D indicates how difficult it is to efficiently fit the parameters of a model M_θ:

$$\begin{aligned} p_D(M_\theta, x) &= \mathbb{E}_{\theta|x}[-2\log p(x|\theta)] - 2(-\log p(x|\tilde{\theta}(x))) \\ &= 2\log p(x|\tilde{\theta}(x)) - 2\int_{\theta\in\Theta} \log p(x|\theta)\; p(\theta|x)\, d\theta \end{aligned} \tag{8}$$

where $\mathbb{E}_{\theta|x}$ refers to the posterior expectation with respect to $p(\theta|x)$. Reading Eq. 8 in yet another way, p_D measures the extent to which our estimate $\tilde{\theta}(x)$ is expected to overfit the data and how much deviance we can expect to observe in the future. This interpretation connects p_D to the predictive performance of our estimate.

[7]There are several possible justifications for this particular measure. First of all, this function is inversely related to the probability of x under θ. If x occurs and it was considered to be unlikely, the surprise under the parameter value θ is high. Thus, the hypothesis gets "punished" by being assigned a high deviance $-\log p(x|\theta)$ from the data. Vice versa, if x is likely under θ, the hypothesis is "rewarded" by being assigned a low deviance. Second, if the data x consist of several independent observations $(x_1, \ldots, x_N)$, then we should be able to decompose the overall deviance into the deviance of the single observations. The $-\log p(x|\theta)$ function accounts for that feature in a particularly natural and intuitive way since $\log p(x_1, \ldots, x_N|\theta) = \sum_i \log p(x_i|\theta)$: the overall deviance of independent observations is the sum of the individual deviances.

Indeed, p_D has been used regularly for assigning scores to candidate models, and it serves as the basis of the Deviance Information Criterion (DIC), a model comparison procedure trading off deviance and complexity. DIC is defined as

$$DIC(M_\theta, x) = \mathbb{E}[D(\theta, x)] + p_D(M_\theta, x) \tag{9}$$

where the function $D(\cdot, \cdot)$ is defined as

$$D(\theta, x) = -2 \log p(x|\theta) + 2 \log f(x) \tag{10}$$

for some standardized function of the data $f(x)$. Taking into account that Eq. 10 is mainly a function of the deviance between model M_θ and data x, we can regard the overall DIC score in Eq. 9 as a tradeoff between goodness of fit (the D-term) and the expected overfit (p_D).

The form of DIC already illustrates that its *target* of inference is not particularly Bayesian. The difficulty of accurately fitting a model is relevant for the practitioner (e.g., for checking the adequacy of a model), but not of intrinsic interest for the orthodox Bayesian reasoner. On the other hand, there are many Bayesian elements in DIC: the estimator $\tilde{\theta}(x)$, whose deviance is estimated in Eq. 8, is nothing but the posterior mean of θ, and it is evaluated with respect to the posterior distribution of θ. Also, Spiegelhalter et al. (2002) show how DIC can be understood as an approximate estimator of posterior expected loss.

The inventors of p_D and DIC are actually aware of that tension and clarify that they believe a rigorous Bayesian justification to be neither available nor necessary:

> Our approach here can be considered to be semiformal. Although we believe that it is useful to have measures of fit and complexity, and to combine them into overall criteria that have some theoretical justification, we also feel that an overformal approach to model 'selection' is inappropriate since so many other features of a model should be taken into account before using it as a basis for reporting inferences [...]. (Spiegelhalter et al. 2002, 602)

DIC is thus a formidable example of a hybrid, eclectic approach to inference in model selection: it is inspired by Bayesianism, frequentism and statistical decision theory. Notably, this eclecticism can go either way: For instance, if the amount of prior information is substantial compared to the data set, then the classical, frequentist AIC can be calibrated as to asymptotically approximate the Bayes factor of different models (Kass and Raftery 1995), or it can be represented as a more general Bayesian criterion (Forster and Sober 2011).[8]

From this analysis, we see that the idea to identify Bayesian model selection by means of its inferential targets is not convincing. In particular, DIC clearly

[8]Resampling procedures, cross-validation, provide another benchmark for model selection procedures (Stone 1977; Forster 2007), and it is an empirical question to what extent they can perform this function better or worse than a Bayesian analysis.

demonstrates that the targets of Bayesian model selection procedures are much more nuanced and varied than just posterior probabilities or Bayes estimates. Second, and more generally, target and justification of a model selection procedure are usually intertwined and hard to separate from each other.

5 Conclusions

What do Bayesian model selection procedures teach us about Bayesian philosophy of science? Their explicitly Bayesian formalism suggests that they are supported by a full-fledged Bayesian philosophy of inference. A closer look reveals, however, that this claim is not substantiated. Popular Bayesian model selection procedures, such as MML, BIC and DIC, may only partially conform to Bayesian reasoning, even if they are firmly anchored within the Bayesian formalism. Rather, they should be described as hybrid procedures: the Bayesian calculus may serve a different goal (MML: efficient coding), some crucial elements of Bayesian reasoning may be dropped (BIC: subjective priors), and ideas and techniques from different philosophies (DIC: Bayesianism, decision theory, frequentism) may be mixed. This need not conflict with a general classification as Bayesian model selection procedures, but it highlights differences in target, justification and intended application context.

Accordingly, the question of what *justifies* these procedures cannot be answered in full generality. Neither of them has a general frequentist or Bayesian justification. Consequently, the adequacy of the chosen procedure depends on whether the implicit assumptions in the derivations of the procedures are satisfied. For example, BIC discounts the priors and focuses on asymptotic behavior, whereas DIC is particularly apt in hierarchical models, etc. MML, on the other hand, does not work with a fixed set of candidate models: efficiently partitioning the model space is already an essential part of the inference problem! This is a crucial difference to BIC and DIC. The practitioner faces the non-trival task to ensure that a model selection procedure is adequate for a given application context.

Thus, there is no unified "Bayesian philosophy of model selection" exemplified in MML, BIC and DIC. This has repercussions on attempts to exploit properties of Bayesian model selection procedures for an assessment of Bayesian statistical inference in general. For example, Forster and Sober (1994) write:

> Bayesianism is unable to capture the proper significance of considering *families* of curves [...] Akaike's reconceptualization of statistics does recommend that the foundations of Bayesian statistics require rethinking. (Forster and Sober 1994, 26, original emphasis)

Other authors, on the contrary, promote Bayesianism because of the apparent success of Bayesian model selection in practice. For instance, Dowe et al.

(2007) defend MML on grounds of its generality, efficiency and invariance under transformations of the parameter space. Then they conclude:

> Since MML is a Bayesian technique we should conclude that the best philosophy of science is Bayesian. (Dowe et al. 2007, 712)

However, we have seen that an implicit premise of such arguments—namely that Bayesian model selection is firmly anchored in Bayesian philosophy—is usually not satisfied. Therefore it is hard to draw a general moral from Bayesian model selection for the philosophical dispute between Bayesians and frequentists. Such a negative conclusion may not appeal to everyone, but to me, it seems the most honest answer to the question of what kind of philosophical claims can be supported by the statistical practice of model selection.

Acknowledgements For helpful discussion and feedback, I am indebted to Dawid Dowe, Steve Gardner, Stephan Hartmann, Casey Helgeson, Rogier de Langhe, Jan-Willem Romeijn and all the audiences where this paper was presented. My research was financially supported by Veni grant 016.104.079 "An Objective Guide to Public Policy? Scope and Limits of Data-Driven Methods in Statistical Model Evaluation" of the Netherlands Organisation for Scientific Research (NWO).

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B.N. Petrov, & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Bandyopadhyay, P.S., & Boik, R.J. (1999). The curve fitting problem: a Bayesian rejoinder. *Philosophy of Science, 66*, S390–S402.

Bandyopadhyay, P.S., Boik, R.J., Basu, P. (1996). The curve fitting problem: a Bayesian approach. *Philosophy of Science, 63*, S264–S272.

Bayarri, M., DeGroot, M., Kadane, J. (1988). What is the likelihood function? In S. Gupta, & J. Berger (Eds.), *Statistical decision theory and related topics IV* (pp. 1–27). Springer: New York.

Berger, J.O., & Wolpert, R.L. (1984). *The likelihood principle*. Hayward/CA: Institute of Mathematical Statistics.

Bernardo, J.M. (1979). Expected information as expected utility. *Annals of Statistics, 7*, 686–690.

Bernardo, J.M. (2011). Integrated objective Bayesian estimation and hypothesis testing. In J. Bernardo, et al. (Eds.), *Bayesian statistics 9: Proceedings of the ninth valencia meeting* (pp. 1–68) (with discussion). Oxford: Oxford University Press.

Bernardo, J.M., & Smith, A.F.M. (1994). *Bayesian theory* Chichester: Wiley.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association, 57*, 269–306.

Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Burnham, K.P., & Anderson, D.R. (2004). Multimodel inference. Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261–304.

Dowe, D.L. (2011). MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In Forster, M., & Bandyopadhyay, P.S. (Eds.), *The philosophy of statistics* (pp. 901–982). Dordrecht: Kluwer.

Dowe, D.L., Gardner, S, Oppy, G. (2007). Bayes not bust! Why simplicity is no problem for Bayesians. *The British Journal for Philosophy of Science, 58*, 709–754.

Earman, J. (1992). *Bayes or bust?* Cambridge, MA: The MIT Press

Forster, M. (1995). Bayes or bust: simplicity as a problem for a probabilist's approach to confirmation. *British Journal for the Philosophy of Science, 46*, 399–424.

Forster, M. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science, 69*, S124–S134.
Forster, M. (2007). A philosopher's guide to empirical success. *Philosophy of Science, 74*, 588–600.
Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for Philosophy of Science, 45*, 1–35.
Forster, M., & Sober, E. (2011). AIC scores as evidence—a Bayesian interpretation. In Forster, M., & Bandyopadhyay, P.S. (Eds.), *The philosophy of statistics* (pp. 535–549). Dordrecht: Kluwer.
Grünwald, P. (2005). A tutorial introduction to the minimum description length principle. In P. Grünwald, I.J. Myung, M. Pitt (Eds.), *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.
Gelman, A., & Shalizi, C. (2012). Philosophy and the practice of Bayesian statistics. Manuscript available arXiv:1006.3868v4.
Han, C., & Carlin, B.P. (2001). Markov Chain Monte Carlo methods for computing bayes factors: a comparative review. *Journal of the American Statistical Association, 96*, 1122–1132.
Henderson, L., Goodman, N.D., Tenenbaum, J.B., Woodward, J.F. (2010). The structure and dynamics of scientific theories: a hierarchical Bayesian perspective. *Philosophy of Science, 77*, 172–200.
Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for Philosophy of Science, 55*, 1–34.
Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). La Salle: Open Court.
Kass, R., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–790.
Mikkelson, G.M. (2006). Realism vs. instrumentalism in a new statistical framework. *Philosophy of Science, 73*, 440–447.
Milne, P. (1996). log[P(h|eb)/P(h/b)] is the one true measure of confirmation. *Philosophy of Science, 63*, 21–26.
Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.
Romeijn, J.W., van der Schoot, R, Hoijtink, H. (2012). One size does not fit all: derivation of a prior-adapted BIC. In D. Dieks, W. Gonzales, S. Hartmann, F. Stadler, T. Uebel, M. Weber (Eds.), *Probabilities, laws, and structures* (pp. 87–106). Berlin: Springer .
Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.
Sober, E. (2008). *Evidence and evolution*. Cambridge: Cambridge University Press.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B, 64*, 583–639.
Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B, 39*, 44–47.
Talbott, W. (2008). Bayesian epistemology. Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/epistemology-bayesian/. Accessed 4 April 2012.
Wallace, C. (2005). *Statistical and inductive inference by minimum message length*. New York: Springer.