What is a Bayesian Model Selection Procedure?

Jan Sprenger

March 10, 2011

Abstract

The debate about the role of Bayesianism in model selection, and about differences to frequentist methods, usually relies on a tacit identification of Bayesianism with orthodox, properly subjective Bayesian inference. We argue that such an assumption is neither descriptively appropriate nor normatively helpful if Bayesianism is supposed to be a model of scientific reasoning. First, instead of the orthodox Bayesian approach, instrumental Bayesianism prevails in practice: the Bayesian framework is understood as a convenient mathematical machinery and conceptual toolbox, not as a philosophy of inductive inference opposed to frequentist reasoning. Second, Bayesian inferences need (and should) not always be based on high posterior probabilities, as demonstrated by embedding Bayesian belief revision into a decision-oriented framework. These findings do not only clarify the various functions of Bayesian reasoning in model selection, but also explain the conceptual eclecticism in the current statistical literature, and open novel prospects for a Bayesian philosophy of science. We defend our claims by examining three prominent Bayesian model selection criteria: BIC, DIC and BRC.

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]

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 model selection (Forster 1995; Bandyopadhyay et al. 1996; Bandyopadhyay and Boik 1999; Dowe et al. 2007).

[1] In this paper, we 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 we would like to stick with the traditional terminology.

This last point is the focus of the article. Of the many model selection criteria that have been invented in the last decades, some have been described as Bayesian, while others have been advanced as a distinctly non-Bayesian, frequentist approaches to model selection (e.g., AIC in Forster and Sober's 1994 paper). Dependent on the author's standpoint, the label "Bayesian" is understood as either an asset or a drawback of a procedure. To give two quotes characteristic of both ends of the spectrum:

> 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)

And vice versa, with reference to a particular model selection procedure, the Minimum Message Length (MML) principle:

> The MML principle provides a much superior formalization [...] Since MML is a Bayesian technique we should conclude that the best philosophy of science is Bayesian. (Dowe et al. 2007, 712)

In that debate, Bayesianism is by default identified with orthodox, subjective Bayesianism. The basic principles of this account state that agents entertain subjective degrees of belief in a hypotheses H, that these degrees of belief can be represented by a probability function $p(\cdot)$, and that we learn from experience by means of conditionalizing on data D according to Bayes' Theorem:

$$p_{\text{new}}(H) := p(H|D) = p(H)\frac{p(D|H)}{p(D)}. \tag{1}$$

The question at stake is not whether Bayes' Theorem is a valid belief revision rule, but whether orthodox Bayesianism is a viable model of *scientific* rationality. Proponents claim that "scientific reasoning is essentially reasoning in accordance with the formal principles of probability" (Howson and Urbach 1993, xvii). Accordingly, high posterior probability becomes a measure of the acceptability of a hypothesis, and scientific inference, including model selection, is based on this posterior distribution of beliefs. This variety of Bayesianism is the most common one in philosophy of science (e.g., Earman (1992, 142); Talbott (2008)), and associated with the dissent between Bayesians and frequentists on the foundations of statistics.

As we shall argue, orthodox Bayesian inference is at the heart of just a minority of Bayesian model selection procedures, and moreover an inadequately narrow framework. Therefore, identifying Bayesianism with that particular variety is misleading. We set out to give a more comprehensive and balanced picture, arguing for the prevalence of an *instrumental* view of Bayesianism within model

selection. On the instrumental view, Bayesianism is a convenient mathematical and conceptual toolbox – assigning probabilities to unknown quantities facilitates various inferential problems – but one need not regard model probabilities as an accurate, honest representation of subjective uncertainty. In particular, we argue that quite often, Bayesian model selection strategies cannot be justified by recourse to foundational principles, and that Bayesian inferences in model selection need not be based on the most probable models. Consequently, we can explain the conceptual and philosophical eclecticism that dominates in the model selection literature, and the lack of a sharp distinction between Bayesian and frequentist procedures, either in terms of inferential targets, or in terms of reasoning strategies.

The paper rests on three main arguments: First, we show that properly subjective, orthodox Bayesian approaches to model selection face serious theoretical and practical difficulties that prevent them, as things stand now, from being a general Bayesian solution to the problem of model selection (section 2). Second, we argue that Bayesian procedures are typically motivated by arguments that fit into the conceptual and mathematical Bayesian framework, but without possessing rigorous justifications. This claim is defended by an analysis of two of the most prominent Bayesian criteria, BIC and DIC (section 3 and 4). Third, we oppose the view that Bayesian procedures are, unlike their frequentist counterparts, always interested in models with high posterior probability. While this may be a fair characterization of many available procedures, a decision-oriented Bayesian approach, such as Bernardo's BRC, demonstrates that this need not be a conceptual necessity. Like frequentists, Bayesians can rank fitted models on the basis of expected predictive accuracy (section 5). Based on these insights, we make a novel proposal for a "Bayesian philosophy of science": as a decision-oriented modeling tool for statistical inference that is more flexible, noncommittal and realistic than an exclusively probabilistic approach (section 6).

2. The Dilemma of the Orthodox Bayesian

To avoid equivocations, we fix 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 contrary, 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 $x_1, \ldots, x_N$, we consider candidate models $M \in \mathcal{M}$ 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 $\hat{\theta}(x)$ and $\hat{\sigma}^2(x)$ such that for any other θ and σ^2: $p(x|\hat{\theta}, \hat{\sigma}^2) \geq p(x|\theta, \sigma^2)$. The classical model selection literature focused on selecting a fitted model and evaluating its predictive accuracy (e.g., Akaike 1973), but there are others that aim at the model with the highest posterior probability (e.g., Schwarz 1978).

From a naïve orthodox Bayesian point of view, one would tackle a model selection problem by assigning truly subjective prior probabilities to candidate models and the hypotheses therein, and conditionalizing on the data by means of equation (1). Then, we can select a model or hypothesis on the basis of its posterior probability.

But even in one of the very simplest cases of model selection, this is problematic. Assume that we are comparing a specific hypothesis, namely that the data follow a $N(\theta_0, 1)$-distribution, to the more general model $N(\theta, 1)$ with unknown parameter (vector) θ. Any reasonable subjective prior density over the unknown θ will assign probability zero to any point value of θ, including $\theta = \theta_0$, because we do not have a special reason to prefer this hypothesis over, say, $\theta = \theta_0 \pm \epsilon$, for an arbitrary $\epsilon \in \mathbb{R}$. Then, the posterior probability of the general model will always be greater than the posterior probability of the point hypothesis. So it seems that a Bayesian would never select the simpler model, although that one can be much more informative and useful (Forster and Sober 1994; Forster 1995). More generally, proper Bayesian model selection based on posterior probabilities is not able to appreciate the specifics of having to choose between nested models.

This problem therefore demands a different approach. Model selection based on *Bayes factors* is a natural candidate. Bayes factors are a measure of evidence where the evidence for M_i vis-à-vis M_j is written as B_{ij} and defined as the ratio of prior and posterior odds:

$$B_{ij}(D) := \frac{p(M_i|D)}{p(M_j|D)} \cdot \frac{p(M_j)}{p(M_i)} = \frac{p(D|M_i)}{p(D|M_j)}, \tag{2}$$

which can alternatively be interpreted as an averaged likelihood ratio of M_i vs. M_j.

For a analysis based on Bayes factors it is no problem that the encompassing general model is always more "probable" than the nested model, since we are comparing how well each of them explain the data, and not whether they are

more likely true or false. Kass and Raftery (1995) have argued in a much-cited review paper that Bayes factors can act successfully as a model selection criterion. In particular, the authors argue that they act as a natural Ockham's razor in the case of comparing nested models, because there are more poorly fitting hypotheses in the general models than in the simple models. This argument is echoed in Henderson et al. (2010).

However, this approach faces some substantial theoretical and practical difficulties. First of all, some typical assumptions of a Bayes factors analysis, e.g., that the true model is included in the set of candidate models, or that the set of models does not vary with the sample size, can be easily violated in practice (Han and Carlin 2001). Second, calculating Bayes factors requires big computational efforts, making it often impracticable. But the most obvious and serious criticism concerns the assignment of subjective priors. In general, the results of a Bayes factors analysis are sensitive to the choice of a prior distribution over the elements of a model. If there is no reliable, commonly accepted source for constructing priors, one may find oneself in a dilemma: the more we rely on distinctly subjective beliefs as an input for calculating Bayes factors, the less persuasive force does our analysis possess. Notably, it does not help to use a conventional procedure (e.g., assigning uniform priors over parameters in a model), because the results would then vary with our specific parametrization. For instance, if we decide to adopt a uniform distribution as an expression of our ignorance, and the hypotheses are parametrized by $\theta \in [0, 1]$, then our inferences will be different if we switch to the parametrization $\theta^2 \in [0, 1]$. But since such parametrizations are just an artefact of the mathematical language we choose, and do not convey any information about our *beliefs*, they should not affect a subjective Bayesian *inference* (Forster 1995, 409–410).

The aforementioned problems demonstrate that orthodox Bayesianism does not provide a straightforward solution to the complex and intricate problem of model selection. To take up the challenge, the orthodox Bayesian can develop a methodology to specify reasonable priors for subjective inference. This can mean to develop feasible procedures for eliciting subjective priors, or to defend sensible reference priors. If successful, that approach would take away one of the most pressing problem of a properly Bayesian approach, namely the alleged arbitrariness of an analysis based on subjective priors. Another option is to change the conceptual framework for eliciting prior probabilities. This option has been pursued by Dowe et al. (2007): they defend the Minimum Message Length (MML) principle as a properly Bayesian approach that can alternatively be regarded as a solution to data compression problems. This move also addresses criticisms pertaining to model misspecification (Dowe 2011, 944).

While these attempts maintain the principles of orthodox Bayesian inference, namely inference based on high posterior probabilities, they still have to struggle against the theoretical and practical difficulties mentioned above. The jury on this research program is still out – what matters for the purposes of this article is that trying to save the Bayesian orthodoxy is not the only way to be a Bayesian in model selection. Quite to the contrary, identifying Bayesianism in model selection with orthodox subjective Bayesianism would misrepresent the state of the art. This does *not* mean that practitioners resort to Jeffreys' old idea of putting model complexity considerations into the prior. Instead, most Bayesian model selection procedures pursue the instrumental Bayesian approach: probability statements about parameters of interest need not be an accurate expression of subjective uncertainty, but may be made *ad hoc* to construct sensible estimators and selection procedures (where "sensible" may at times be interpreted in a frequentist sense). On that view, we do not share the philosophical commitments of orthodox Bayesianism, but take advantage of the flexibility of the Bayesian framework. In other words, the formal differences – Bayesians, but not frequentists assign probabilities to unknown parameter values – need not lead to completely different inferential strategies. We argue for this thesis by investigating the derivations of three prominent model selection criteria: the Bayesian Information Criterion (BIC), the Deviance Information Criterion (DIC), and the Bayesian Reference Criterion (BRC).

3. Bayesianism Without Model Priors: 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 corresponding 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, and that 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, that is, they can be written as

$$p(x|\theta) = e^{N(A(x)-\lambda|\theta-\hat{\theta}(x)|)^2}. \tag{3}$$

Here, $\hat{\theta}(x)$ denotes the maximum likelihood estimate of the unknown θ, and N the sample size, assuming independent sampling. This specific form of the likelihood function seems to make a substantial presumption, but in fact, the densities in (3) comprise the most familiar distributions, such as the Normal, Fisher, Poisson, Student's $t-$, and the uniform 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 the formula

$$\begin{aligned} \log p(M_\theta|x) &\sim \log p(M_\theta) + NA(x) + \log\left(\frac{1}{N\lambda}\right)^{K/2} + \log \int_{\theta\in\Theta} e^{-|\theta-\hat{\theta}(x)|^2} d\mu(\theta) \\ &= \log p(M_\theta) + NA(x) + \frac{1}{2}K \log\left(\frac{1}{N\lambda}\right) + \log \sqrt{\pi}^K \\ &= \log p(M_\theta) + \log p(x|\hat{\theta}(x)) - \frac{1}{2}K \log\left(\frac{N\lambda}{\pi}\right). \end{aligned} \tag{4}$$

Let us take stock. On the left hand side, we have the the log-posterior probability, a Bayesian's standard model comparison criterion. As we see from (4), this term can be split up into 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.[2]

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 (4) 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

$$BIC(M_\theta, x) = -2 \log p(x|\hat{\theta}(x)) \; + \; K \; \log N. \tag{5}$$

BIC is evidently an estimator of the model that would, in the long run, accumulate the most posterior mass. However, it neglects the contribution of the model priors when comparing the models to each other. Therefore, it is questionable whether it should be described as a *subjective* Bayesian technique.[3] Where subjective Bayesian positions typically consider prior probabilities to be relevant

[2]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.

[3]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."

for inference and see a need to elicit them, they drop out of the picture in BIC, as witnessed by the transition from (4) to (5).

Similarly, it has been observed (e.g., Kass and Raftery 1995) that BIC can be used as an approximation to the Bayes factor, the Bayesian's measure of evidence. Taking the difference of the BIC score of two models $M^{(1)}$ and $M^{(2)}$ with dimension K_1 and K_2 and maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, we obtain, involving some abuse of notation:

$$\begin{aligned} & BIC(M^{(1)}, x) - BIC(M^{(2)}, x) \\ = \quad & -2 \log p(x|\hat{\theta}_1(x)) \, + \, K_1 \, \log N \, + \, 2 \log p(x|\hat{\theta}_2(x)) \, - \, K_2 \, \log N \quad (6) \\ = \quad & 2 \log \Big(p(x|\hat{\theta}_2(x)) / p(x|\hat{\theta}_1(x)) \Big) - (K_2 - K_1) \log N \\ \stackrel{N \to \infty}{\longrightarrow} \quad & 2 \, \log B_{21}, \end{aligned}$$

since for large N, $(K_2 - K_1) \log N / \log B_{21} \, \to \, 0$. Thus, the BIC difference approximates twice the log-Bayes factor of $M^{(2)}$ with respect to $M^{(1)}$ so that it can be related, at least for large samples, to a standard Bayesian measure of evidence.

Again, we stress that calculations such as (6) should not me misunderstood as an ultimate Bayesian *justification* of BIC. Rather, they show the *compatibility* of BIC and the Bayes factor approach, in the sense that under specific conditions (regular priors, large samples, etc.), the results delivered by BIC would agree with those obtained from an orthodox, truly subjective Bayesian analysis. This is as Bayesian as it gets with BIC. Moreover, one of the standard arguments in favor of BIC focuses on its *statistical consistency*, 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.[4] On this criterion, the BIC tends to outperform frequentist alternatives, such as Akaike's (1973) classical AIC. So although the above justifications of BIC are entirely couched into Bayesian terms, we diverge conceptually from orthodox Bayesian inference, as evident from the neglect of model priors, and the evaluation in terms of long-run (sampling) properties, which is typical of frequentist statistics. All this underlines that for BIC, Bayesianism constitutes no philosophical underpinning as a logic of belief change, but only a *convenient framework* for motivating the use of a specific estimator of log-posterior probability.

[4]To repeat, consistency is not meant as a logical consistency with another proposition, but as a characterization of the long-run properties of a statistical estimator. It should also be mentioned that if the amount of data per (nuisance) parameter is bounded, then all maximum-likelihood-based methods will turn out to be inconsistent, including BIC and Akaike's AIC (Neyman-Scott problem).

4. Estimating Effective Complexity: DIC

Another important model selection criterion that figures as Bayesian is the Deviance Information Criterion (DIC). Most 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. That is, when we represent the marginal distribution of the data x in a certain probability model as

$$p(x) = \int_{\theta \in \Theta} p(x|\theta)\, p(\theta)\, d\theta \tag{7}$$

with parameter θ and prior density $p(\theta)$, we may sometimes choose to represent that prior as being governed by a hyperparameter ψ:[5]

$$p(\theta) = \int_{\psi \in \Psi} p(\theta|\psi)\, p(\psi)\, d\psi. \tag{8}$$

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)? Likewise, 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. The straightforward measurement of complexity as the number of free parameters, that was used in the case of BIC, is therefore inappropriate as a general procedure.

The inventors of DIC propose to measure model complexity in terms of its estimation properties. Such an understanding is also known as the *effective* number of parameters of a model. 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).

To make this explicit, the authors propose to measure complexity by comparing the expected deviance in the data (under our posterior distribution) to the deviance in the estimate $\tilde{\theta}(x)$ that we would like to use. This would give us some idea of the "difficulty in estimation." Thus, we need to measure the surprise or *deviance* in the data x relative to a point hypothesis θ.

The canonical measure of deviance between data x and a model is $-\log p(x|\theta)$ (Good 1952; Bernardo 1979). There are several possible justifications for this particular measure; we give the one that we find most simple and appealing.

[5]The marginal distribution of the data (7) is not affected by whether we parametrize the prior with hyperparameter ψ according to (8).

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, our 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 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.

Having chosen a way of measuring deviance, we return to comparing expected to actual deviance. $\tilde{\theta}(x)$ denotes a standard Bayesian estimator of our quantity of interest θ, namely 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 standard estimate of θ. This quantity p_D indicates how difficult it is to efficiently fit the parameters of a model M_θ:

$$\begin{aligned} p_D(M_\theta, x) &= \mathbb{E}_{\theta|x}[-2\log p(x|\theta)] - 2(-\log p(x|\tilde{\theta}(x))) \\ &= 2\log p(x|\tilde{\theta}(x)) - 2\int_{\theta\in\Theta} \log p(x|\theta)\, p(\theta|x)\, d\theta \end{aligned} \tag{9}$$

Reading (9) in 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.

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 trading off deviance and complexity. DIC is defined as

$$DIC(M_\theta, x) = \mathbb{E}[D(\theta, x)] + p_D(M_\theta, x) \tag{10}$$

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

$$D(\theta, x) = -2\log p(x|\theta) + 2\log f(x) \tag{11}$$

for some standardized function of the data $f(x)$. Taking into account that (11) is mainly a function of the deviance between model M_θ and data x, or viewed the other way round, of the fit between model and data, we can regard the overall DIC score in (10) as a tradeoff between goodness of fit (the D-term) and the model's complexity (p_D).

Does this procedure exemplify the orthodox or the instrumental Bayesian approach? First, Spiegelhalter et al. (2002) show how DIC can be understood as an approximate estimator of posterior expected loss. Unfortunately, like in

the case of BIC, that derivation is not rigorous, based on various contentious assumptions, and can easily break down. Second, and presumably more importantly, our estimator $\tilde{\theta}(x)$ in (9) is nothing but the posterior mean of θ, and evaluated again with respect to the posterior distribution of θ. So indeed, DIC makes substantial use of prior information and posterior distributions in its understanding and measurement of model complexity, ostensibly in line with the orthodox understanding.

However, rather than estimating the most probable model or hypothesis, DIC and p_D estimate the complexity of a model, a quantity that is not of intrinsic interest for the orthodox Bayesian. Moreover, the multiple use of posterior distributions is, from an orthodox perspective, also a weakness. The data are used several times, to obtain an estimate $\tilde{\theta}(x)$ of parameter θ, and to calculate a posterior density over θ which figures in our evaluation of this estimate. From an orthodox Bayesian point of view, this double-counting amounts to inferential bias (Berger and Wolpert 1984). Data should enter an inference only once, namely via the likelihood function, and the resulting posterior distribution is all that we need for inferential purposes.

Thus, a rigorous Bayesian justification of DIC is unavailable. The inventors of p_D and DIC are actually quite aware of that and clarify that they consider a rigorous Bayesian justification of these techniques neither available of necessary:

> There has been a long and continuing debate about whether the issue of selecting a model as a basis for inferences is amenable to a strict mathematical analysis using, for example, a decision theoretic paradigm [...]. Our approach here can be considered to be semi-formal. 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)

So again, we feel it is adequate to characterize the underlying rationale as instrumentally Bayesian: the Bayesian machinery is used to solve a specific estimation problems about inference in hierarchical models, but not by means of foundational Bayesian arguments. As the ensuing discussion in Spiegelhalter et al. (2002) makes clear, the assessment of DIC crucially depends on the kind of solutions it gives to canonical statistical inference problems, such as model selection in the case of nested models. The theoretical worries we outlined above are, in practice, a minor issue. Calling DIC a "Bayesian measure of model complexity and fit" – the title of the paper where it was proposed – is

appropriate for an instrumental understanding, but not for an orthodox view on Bayesian inference.

DIC is thus a formidable example of a hybrid philosophy of inference in model selection, and of the eclectic, engineering-like approach that dominates much of the current model selection literature. Bayesian machinery is borrowed for pursuing a specific inferential goal. 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, like the Bayesian BIC (Akaike 1983; Kass and Raftery 1995). This reveals an interesting function of Bayes factors: the different conditions under which AIC or BIC succeed in approximating them can be used to characterize the properties of these criteria, the implicit assumptions they make, and the circumstances where they work.[6]

5. Predictive Accuracy, Not Probability: BRC

If Bayesian model selection criteria do not always have a foundationally sound derivation, what remains distinctive of Bayesian inference in model selection? Often, a distinction is suggested in terms of *inferential targets*. Characteristic are quotes such as "Bayesians assess an estimator by determining whether the values it generates are probably true or probably close to truth" (Forster and Sober 2011) 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). The last quote appears to invoke the old Bayesians vs. frequentists distinction, but actually rephrases this distinction in terms of inferential targets, rather than in terms of epistemic justifications. 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. Put slightly differently, frequentists aim at representative targets, and Bayesians at aggregate targets. These differences appear to be the best game in town for classifying model selection criteria in a

[6]Thus, frequentist procedures may turn out to be compatible with a Bayesian analysis, in the sense of approximating Bayes factors, posterior distributions, or the like, under well-defined circumstances. These findings are not too surprising: in regular circumstances, good Bayesian procedures are expected to have good frequentist properties, and vice versa. Thus, a Bayesian framework can be used for better understanding and effectively gauging various model selection procedures, e.g., by analyzing which kind of prior assumptions are implicit in a specific criterion (Burnham and Anderson 2002), or by comparing model averaging techniques to predictive posterior distributions. This comparative function of Bayesian reasoning might be the valid core of the often-quoted folklore that Bayesians factor simplicity into the priors (Jeffreys 1939; Earman 1992). 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.

meaningful way, and for identifying a common element in Bayesian procedures.

However, taking posterior probability as a model selection criterion is just a special case of Bayesian decision models that contain, besides the probability component, an equally important utility function. To illustrate this claim, consider the simple case of two (supposedly not misspecified) models, M and M' with posterior probabilities 0.75 and 0.25. The natural inference, namely preferring M to M', is decision-theoretically justified if we assume a naïve zero-one utility function, that is, either correct decision has a constant value (one), and vice versa for both incorrect decisions (zero). For different utility functions, e.g., if the cost of erroneously selecting M were much higher than the cost of erroneously selecting M', we might opt for the less probable model M'. There is a wide range of decision models that correspond to a particular posterior probability distribution. Hence,

> "the more traditional Bayesian approaches to model comparison, such as determining the posterior probabilities of competing models or computing the relevant Bayes factors, can be obtained as particular cases [...] by using appropriately chosen, stylised utility functions."(Bernardo and Smith 1994, 420)

Unfortunately, primers on Bayesian philosophy of science, as well as its philosophical critics, tend to reinforce the misleading "traditional" picture, by focusing exclusively on the probabilistic, epistemic dimension of Bayesianism. The neglect of the decision-oriented dimension sometimes goes so far to state explicitly that "issues in Bayesian decision theory will be ignored" (Earman 1992, 33). This is, first of all, ironical since the orthodox approach relies on the decision-theoretic Dutch Book Argument to justify the probabilistic coherence of degrees of belief. More importantly, however, it narrows down the wide decision-theoretic scope of Bayesianism – at all, model selection is ultimately about making inferences, making decisions – to a very special reduct, namely orthodox Bayesianism and inference based (exclusively) on posterior distributions and Bayes factors. Of course, authors such as Earman (1992) or Howson and Urbach (1993) know as well that Bayesian belief models can be extended to Bayesian decision models. However, by neglecting the latter and focussing on the epistemic dimension only, they deprive Bayesianism of the ability to be a viable model of *scientific* reasoning, instead of just another logic of belief revision.

Indeed, the structure of a real model selection problem may easily violate the zero-one utility assumption. To show that this is more than a remote theoretical possibility, consider a nested testing problem: we test a null model $M_0 : \theta = \theta_0$ against the more general alternative $M_1 : \theta \in \Theta$. Practical examples are a two-

sided test of the mean of a Normal distribution, curve-fitting with polynomials of different degrees, etc.

From an orthodox Bayesian point of view, the more specific and informative model is always less probable, and therefore less likely. There is no way to favor the more specific model (that may end up to be more appropriate) on grounds of its *probability*. Therefore, these nested model selection problems should be conceived of as aiming at the *predictive performance* of the candidates (cf. Forster 2002). How can a Bayesian model that?

Answers to this question have, apparently independently, been developed by Bernardo and Smith (1994), Bernardo (1999), and Dupuis and Robert (2003). For reasons of simplicity and space, we focus on Bernardo's (1999) approach and make some notational simplifications. We understand nested model selection as a proper decision problem where the criterion for selecting or rejecting the null model $M_0 : \theta = \theta_0$ consists in trading off expected predictive accuracy with some context-dependent criteria, such as simplicity or informativity. This perspective provides a utility structure that is very different from an orthodox, probability-oriented Bayesian treatment. Briefly, the simpler model can and should be used as a proxy for the more general one if and only if its expected predictive accuracy exceeds a certain threshold.

To set up such a decision model we first need a scoring rule for evaluating models/hypotheses with respect to a set of data. This is again a problem of measuring deviance between parameter value θ and data y, so that we can fall back on the logarithmic scoring rule $\log p(y|\theta)$ that we have motivated in section 4.

A generalization of this utility function describes the score of data y under parameter value θ as $q(\theta, y) = \alpha \log p(y|\theta) + \beta(y)$, where α is a scalar term, and $\beta(y)$ is a function that depends on the data only. Informally speaking, $q(\cdot, \cdot)$ is decomposed into a prediction-term and a term that only depends on the desirability of an outcome (the latter will turn out to be irrelevant). This is a useful generalization of the logarithmic score. Consequently, if θ is the true parameter value, the utility of taking M_0 as a proxy for the more general model M_1 is

$$\int q(M_0, Y)\, dP_{Y|\theta} \quad = \quad \alpha \int \log p(y|\theta_0)\, p(y|\theta)\, dy + \int \beta(y)\, p(y|\theta)\, dy.$$

The overall utility U of a decision, however, should not only depend on the predictive score, as captured in q, but also on the cost c_j of selecting a specific model M_j. The more general M_1 is more difficult to handle than the simpler M_0 because it has one additional degree of freedom, making it less informative and more prone to the risk of overfitting. Therefore it is fair to set $c_1 > c_0$.

Writing $U(M_j, \theta) = \int q(M_j, Y)\, dP_{Y|\theta} - c_j$, we then obtain

$$U(M_0, \theta) = \alpha \int \log p(y|\theta_0)\, p(y|\theta)\, dy + \int \beta(y) p(y|\theta) dy - c_0$$

$$U(M_1, \theta) = \alpha \int \log p(y|\theta)\, p(y|\theta)\, dy + \int \beta(y) p(y|\theta) dy - c_1.$$

Note that the utility of selecting M_0 is evaluated against the true parameter value θ, and that we evaluate the general model M_1 not with respect to a probabilistic average (e.g., the posterior mean), but with respect to is optimal representative, the true value θ. Consequently, the difference in *expected utility*, conditional on the posterior density of θ, can be written as

$$\begin{aligned}
& \int_{\theta \in \Theta} \left(U(M_1, \theta) - U(M_0, \theta)\right)\, p(\theta|x)\, d\theta \\
= \;& \alpha \int_{\theta \in \Theta} \int \log \frac{p(y|\theta_0)}{p(y|\theta)}\, p(y|\theta)\, p(\theta|x) dy\, d\theta + \int \beta(y)\, p(y|\theta)\, dy - \int \beta(y)\, p(y|\theta)\, dy + c_0 - c_1 \\
= \;& \alpha \int_{\theta \in \Theta} \left(\int \log \frac{p(y|\theta_0)}{p(y|\theta)}\, p(y|\theta)\, dy \right) p(\theta|x)\, d\theta + c_0 - c_1.
\end{aligned}$$

This means that the expected utility difference between inferring to the null model and keeping the general model is essentially a function of the expected log-likelihood ratio between the null model and the true model. Simplifying notation, we will reject the null if and only if $\mathbb{E}_\theta[U(M_1, \theta)] > \mathbb{E}_\theta[U(M_0, \theta)]$, that is

$$\int_{\theta \in \Theta} \left(\int \log \frac{p(y|\theta_0)}{p(y|\theta)}\, p(y|\theta)\, dy \right) p(\theta|x)\, d\theta > d^*, \tag{12}$$

for some context-dependent value d^*. The model selection/hypothesis testing criterion based on (12) is called the *Bayesian Reference Criterion* (BRC).

So BRC selects the simpler model if the loss in information incurred by using M_0 as a proxy for the general model M_1 is compensated by the context-dependent advantage of working with a simpler null. In other words, we accept the more specific model if its mean prediction error is small enough to be offset by the gain in informativity, resistance to overfitting, understanding, convenience, etc. Notably, the derivation of BRC did not suggest a particular interpretation of the posterior probability $p(\theta|x)$: this can be filled in by subjective or objective, conventional priors.[7]

The use of the BRC in nested model selection problems shows convincingly that Bayesian model selection is much more encompassing and flexible than inference based on high Bayes factors or posterior probabilities. The orthodox

[7] If there is reason to suspect that the models are seriously misspecified, the BRC procedure may be inappropriate since it relies on the true parameter being included in (or at least well approximated by) the general model. For this case, other procedures have to be devised, see Bernardo and Smith (1994, ch. 6).

reading of Bayesianism neglects the significance of utility functions in contexts where, such as in model selection problems, simplicity, informativity and vulnerability to overfitting of a model are important inferential criteria. In particular, Bayesians can construct decision models that need not favor the most probable models, and that trade off expected predictive accuracy with other, context-dependent considerations. After debunking the thesis that frequentist and Bayesian model selection procedures substantially differ in their *justifications*, we have now demonstrated that they need not differ in their *targets* either.

6. Conclusions

What is a Bayesian model selection procedure? We have argued at length that the orthodox answer, namely basing model selection on posterior probabilities or Bayes factors, is neither normatively nor descriptively appropriate. On the descriptive side, the practically prevailing *instrumental Bayesianism* uses the Bayesian framework as a convenient mathematical machinery and conceptual toolbox, but not necessarily as a philosophy of induction that bases inferences on a honestly subjective posterior distribution. On the normative side, apart from well-known difficulties (e.g., the choice of the priors), orthodox Bayesian model selection is just a very special reduct of a more flexible decision-oriented Bayesian approach. These findings shed a skeptical light on attempts to establish orthodox Bayesian inference as the philosophy of scientific reasoning.

More precisely, the distinctive features of Bayesianism in model selection can be summarized as follows: First, the justifications of Bayesian model selection procedures are semiformal rather than rigorous. Some details may deviate significantly from orthodox Bayesian reasoning, as shown in our analysis of BIC and DIC. Characteristic of these instrumental procedures is also the tendency to incorporate (frequentist) considerations about the sampling distributions of (Bayesian) estimators. Second, Bayesians need not aim at models or hypotheses with the highest posterior probability: by extending the Bayesian model to a full decision model, they can target models with, e.g., the highest expected predictive accuracy, as the case of BRC has made clear. This decision-theoretic dimension is often neglected in discussions of Bayesian philosophy of science.

Thus, in spite of the conceptual and notational differences, there is no clear-cut distinction between Bayesian and frequentist model selection procedures in terms of the underlying epistemic justifications, or in terms of their inferential targets. There is a residual formal difference about whether probability statements on model parameters are meaningful, but in practice, this seems to lead to remarkably little conflict.

Overall, the results of our analysis demonstrate that Bayesian model selec-

tion procedures have surprisingly wide scope. Is there any hope left for a unified and securely grounded Bayesian approach, that is, a Bayesian philosophy of science with serious normative ambitions? It would be beyond the scope of the paper to answer this question, but to our mind, the most sensible proposal is a decision-oriented approach: by combining utility- and probability-related considerations, the Bayesian can defend specific procedures as rational, and retains maximal modeling flexibility. On the other hand, this approach is not yet part of the mainstream so that its descriptive accuracy may be questioned. Be this as it may, Forster and Sober's call for rethinking Bayesian inference, quoted in the introduction, is answered by making Bayesianism more instrumental and more decision-oriented, vindicating it as a fruitful, flexible and philosophically noncommittal tool in model selection.

References

Akaike, Hirotugu (1973): "Information Theory as an Extension of the Maximum Likelihood Principle", in: B. N. Petrov, F. Csaki (ed.), *Second International Symposium on Information Theory*, 267–281. Budapest: Akademiai Kiado.

Akaike, Hirotugu (1983): "Information Measures and Model Selection", *Bulletin of the International Statistical Institute* 50: 277–290.

Bandyopadhyay, Prasanta S., Robert J. Boik, and Prasun Basu (1996): "The Curve Fitting Problem: A Bayesian Approach", *Philosophy of Science* 63: S264-S272.

Bandyopadhyay, Prasanta S., and Robert J. Boik (1999): "The Curve Fitting Problem: A Bayesian Rejoinder", *Philosophy of Science* 66: S390-S402.

Bayarri, M., DeGroot, M., and Kadane, J. (1988): "What is the Likelihood Function?", in: S. Gupta and J. Berger (ed.), *Statistical Decision Theory and Related Topics IV*, 1–27. Springer: New York.

Berger, James O., and Robert L. Wolpert (1984): *The Likelihood Principle*. Hayward/CA: Institute of Mathematical Statistics.

Bernardo, José M. (1979): "Expected Information as Expected Utility", *Annals of Statistics* 7: 686–690.

Bernardo, José M. (1999): "Nested Hypothesis Testing: The Bayesian Reference Criterion" (with discussion), in J. Bernardo et al. (eds.): *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting*, 101–130 Oxford: Oxford University Press.

Bernardo, José M., and Adrian F. M. Smith (1994): *Bayesian Theory*. Chichester: Wiley.

Burnham, Kenneth P., and David R. Anderson (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Second Edition. New York: Springer.

Burnham, Kenneth P., and David R. Anderson (2004): "Multimodel Inference. Understanding AIC and BIC in Model Selection", *Sociological Methods and Research* 33: 261–304.

Dowe, David L. (2011): "MML, Hybrid Bayesian Network Graphical Models, Statistical Consistency, Invariance and Uniqueness", in: Malcolm Forster and Prasanta S. Bandyopadhyay (eds.) (ed.), *The Philosophy of Statistics*, 901–982. Dordrecht: Kluwer.

Dowe, David L., Steve Gardner and Graham Oppy (2007): "Bayes not Bust! Why Simplicity is no Problem for Bayesians", *The British Journal for Philosophy of Science* 58: 709–754.

Dupuis, Jerôme A., and Christian P. Robert (2003): "Variable selection in qualitative models via an entropic explanatory power", *Journal of Statistical Planning and Inference* 111: 77–94.

Earman, John (1992): *Bayes or Bust?*. Cambridge/MA: The MIT Press.

Forster, Malcolm (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, Malcolm (2002): "Predictive Accuracy as an Achievable Goal of Science", *Philosophy of Science* 69: S124–S134.

Forster, Malcolm (2007): "A Philosopher's Guide to Empirical Success", *Philosophy of Science* 74: 588–600.

Forster, Malcolm, and Elliott Sober (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, Malcolm, and Elliott Sober (2011): "AIC Scores as Evidence – A Bayesian Interpretation", forthcoming in Malcolm Forster and Prasanta S. Bandyopadhyay (eds.): *The Philosophy of Statistics*. Dordrecht: Kluwer.

Good, I. J. (1952): "Rational Decisions", *Journal of the Royal Statistical Society B* 14: 107–114.

Han, Cong, and Bradley P. Carlin (2001): "Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review", *Journal of the American Statistical Association* 96: 1122–1132.

Henderson, Leah, Noah D. Goodman, Joshua B. Tenenbaum, and James F. Woodward (2010): "The Structure and Dynamics of Scientific Theories: A Hierarchical Bayesian Perspective", *Philosophy of Science* 77: 172–200.

Hitchcock, Christopher, and Elliott Sober (2004): "Prediction versus Accommodation and the Risk of Overfitting", *The British Journal for Philosophy of Science* 55: 1–34.

Howson, Colin, and Peter Urbach (1993): *Scientific Reasoning: The Bayesian Approach*. Second Edition. La Salle: Open Court.

Jeffreys, Harold (1939): *Theory of Probability*. Oxford: Clarendon Press.

Kass, Robert, and Adrian Raftery (1995): "Bayes Factors", *Journal of the American Statistical Association* 90: 773–790.

Mikkelson, Gregory M. (2006): "Realism vs. Instrumentalism in a New Statistical Framework", *Philosophy of Science* 73: 440–447.

Schwarz, Gideon (1978): "Estimating the Dimension of a Model", *Annals of Statistics* 6: 461–464.

Sober, Elliott (2008): *Evidence and Evolution*. Cambridge: Cambridge University Press.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde (2002): "Bayesian measures of model complexity and fit (with discussion)", *Journal of the Royal Statistical Society B* 64: 583–639.

Stone, Michael (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, William (2008): "Bayesian Epistemology", *Stanford Encyclopedia of Philosophy*, accessed on March 4, 2010 at `http://plato.stanford.edu/entries/epistemology-bayesian/`.