The Weight of Simplicity in Statistical Model Comparison

May 17, 2010

Abstract

The epistemic weight of simplicity in science has, in the last fifteen years, been extensively discussed in the framework of statistical model comparison. This paper defends three theses: First, it juxtaposes Forster and Sober's (1994, 2010) claims regarding the normative force of Akaike's Theorem in statistical model analysis. Second, it elaborates the different targets of the various model comparison criteria, and shows that the weight of simplicity stands orthogonal to the issue of Bayesian vs. frequentist model comparison. Third, it proposes that statistical model comparison cannot be separated from the content of the original scientific problem, rendering elusive a neat division of labor between statisticians and scientists.

1 Introduction

The role of simplicity in scientific inference has always been a subject of vivid discussion, going back to Occam's days. However, the debate has recently undergone a shift of focus, from largely informal investigation of case studies in science to the role of simplicity in statistical model comparison. Especially in the last 15 years, there has been a boom of papers on the virtue of simplicity in statistical inference. This avalanche of contributions was prompted by Forster and Sober's 1994 paper, where they took issue with the traditional view that simpler models are preferred solely on non-empirical, pragmatic grounds, such as mathematical convenience. Instead, they argued that the simplicity of a model (and a specific measure of simplicity, the AIC) was directly linked to the predictive success of the model. Subsequently, the conclusions have been transferred to more general questions in philosophy of science, such as the replacement of truth by predictive accuracy as an *achievable* goal of science (Forster 2000), the prospects for Bayesian statistics (Forster 1995; Dowe et al. 2007), and the realism/instrumentalism dichotomy (Forster and Sober 1994; Mikkelson 2007).

This transfer sounds overambitious since many scientific inferences cannot be described in modern statistical terms. On the other hand, statistical techniques are nowadays a part of nearly every empirical discipline. Quite often, if we want

to make reliable predictions, we have to build statistical models, to fit curves to messy data, and to make a decision which models might be most useful. Comparing a large set of candidate models in order to determine which of them are best suited for future work is nowadays common in many cases of ecological or econometric modeling (Burnham and Anderson 2002; Keuzenkamp 2000). Thus, the specific statistical problem of model comparison – traditionally called model selection[1] –, and the role of simplicity in that procedure, has significant ramifications for science in general.

It is useful to partition the issue of simplicity in statistical model analysis into three questions.

1. The *qualitative* question: Is simplicity more than a purely pragmatic, non-empirical virtue? Does simplicity have epistemic pull, and does statistical model analysis provide the right ground for vindicating simplicity as an epistemic virtue?

This question has to be answered in the affirmative, and it is Forster and Sober's merit to have pointed this out. In section 2, I argue that the proper way to defend that thesis consists in pointing to the difficulty of estimating the parameters of complex models. Forster and Sober, however, go beyond that claim and argue that the epistemic pull of simplicity is established by means of the mathematical properties of a *particular* model comparison criterion, Akaike's information criterion AIC (section 3). This claim (which I deny) brings us to the second question:

2. The *quantitative* question: Is there a definite weight of simplicity in statistical model comparison, or in other words, is there a definite tradeoff rate between simplicity and goodness of fit?

I believe that there cannot be a unique answer. How simplicity should be measured depends to a large extent on exogenous factors, such as the target of the model analysis, the structure of the data, the sample size, and so on. Contrary to what Forster and Sober claim, the mathematics do not generally vindicate the epistemic benefits of simplicity-based model comparison criteria such as AIC (section 4). Rather, these criteria are context-dependent, tentative attempts to quantify the estimation difficulties that arise for complex statistical models. Since comparing their virtues and vices on neutral grounds proves to be very difficult, I conclude that Forster and Sober expect too much from Akaike's Theorem.

[1]Following Sober (2008), I replace the traditional term "model selection" (asking which model we should *select*, while dismissing the rest) by the more adequate term *model comparison* (asking which *evidence* we have for either model).

Conceding that there cannot be a unique simplicity-based model comparison criterion, we still face the question of how the different criteria relate to each other:

3. The *comparative* question: Do the different model comparison criteria pursue the same inferential goals, or are there fundamental differences? Is there at all a common ground for comparing them?

In section 5, I examine the mathematical heuristics behind widespread model comparison criteria, showing that the criteria differ more in terms of their *targets* than in terms of their philosophical underpinnings (e.g. Bayesianism vs. frequentism). Choosing a model comparison criterion presupposes deliberate thinking about the problem which is to be modeled as well as about the goals that are to be achieved. In that respect, I remain pluralist – different criteria suit different goals, and even given a specific goal, it is difficult to compare the criteria meaningfully, apart from idealized or very specific situations (Bandyopadhyay et al. 1996).

Since the weight of simplicity in statistical model comparison is so context- and target-sensitive, I suggest to regard considerations pertaining to simplicity as an element of sensible experimental design. This perspective connects simplicity to questions of scientific understanding, implying that there is no neat division of labor between scientists and statisticians. The final section 6 summarizes the results and concludes.

2 Simplicity Heuristics

Statistical model analysis compares a large set of candidate models to given data, in the hope to select the best model on which predictions and further inferences can be based. Generally, it is unrealistic to assume that the "true model" (i.e. the data-generating process) is found among the candidate models: data sets are often huge and messy, the underlying processes are complex and hard to describe theoretically, and the candidate models are often generated by automatic means (e.g. as combinations of potential predictor variables). This means that the candidate models do not typically provide the most striking mechanism, or the best scientific explanation of the data. Rather, they are constructed from the data, supposed to explain them reasonably well and to be a reliable device for future predictions. This "bottom-up" approach (Sober 2002) to curve-fitting and model-building is complementary to a "top-down" approach where complex models are derived from general theoretical principles (Weisberg 2007).

To explain the role of simplicity in that context, it is useful to consider a simple example. The most common, yet hidden appearance of simplicity in statistical inference occurs in the everyday activity of statistical practice: null hypothesis testing. A null hypothesis H_0, e.g. that the mean μ of a normally distributed population with known variance is equal to μ_0, is tested against the alternative H_1 that the mean is different from μ_0. Certainly, the alternative is more complex than the null since it has one additional degree of freedom: the parameter μ takes no definite value under H_1. This allows H_1 to fit the data better than H_0. On the other hand, if the population mean is not exactly equal to μ_0, but quite close to it, the null hypothesis still does a good job. Therefore, highly significant results are demanded before the null is rejected in favor of the alternative. By imposing high standards before calling observed data significant evidence against H_0, we compensate for the fitting edge that more complex hypotheses possess.

This rationale can be generalized: Complex models have higher power to fit the data than simple models. They can achieve a satisfactory fit for way more data sets than simple models can. Thus, *a priori*, before looking at the data, we already know that very likely, the complex models will achieve a more convincing fit than the simple models (Hitchcock and Sober 2004).

These superior fitting resources can also be a vice as the problem of *overfitting* illustrates: the more degrees of freedom a model has, the more difficult it is to simultaneously estimate all model parameters. We will often fit noise to the data. This is often put in the words that complex models have *high estimation variance*. Notably, complex models can perform worse than simpler models even if they are, in some sense, closer to the data-generating process.

Consider the case of nested polynomials. Assume that we describe the relationship between input variable x and output variable y by a polynomial of degree K plus a noise term ε, with $K+1$ adjustable parameters α_0-α_K, e.g.

$$y = \alpha_0 x^K + \alpha_1 x^{K-1} + \ldots + \alpha_K + \varepsilon_n. \tag{1}$$

Thus, we have a family of $K+1$-dimensional curves that have to be fitted to the observed data. Assume further that factually, the leading coefficients α_0, α_1, etc. are very close to zero, but not identically zero. When we are simulating data with such a model, the data will effectively look like generated from a (simpler) polynomial of lower degree K'. Now, while there is a strong intuition that we should try to find the correct complex model, from a *predictive* perspective, we should go for the wrong simple model: the joint estimation of the coefficients will almost always lead to misleading parameter values.[2] So even when truth is *not* elusive, knowing the true model class need not entail predictive success, due

[2]I am indebted to Jan Magnus for directing my attention to that example.

to the higher estimation variance. Simpler models often do better predictively, even if they are, strictly speaking, wrong.

These problems carry special weight in the context of maximum likelihood estimation. A *maximum likelihood estimator* (MLE) is the parameter value that makes the actual data most likely. MLE is an omnipresent statistical estimator that also figures centrally in regression techniques, such as the ordinary least squares (OLS) estimation.[3] Unfortunately, maximum likelihood estimation is usually overoptimistic with respect to the predictive performance of the chosen model, especially when the number of adjustable parameters is high. An MLE always selects the best-fitting model in a model class, and projecting the current goodness of fit to future predictive accuracy just neglects the problem of high estimation variance and the danger of overfitting.

This informal and intuitive, yet fully empirical understanding of simplicity is the basis of all attempts to introduce an appropriate impact of simplicity into model comparison. It is now natural to ask whether the predictive virtues of simplicity can be derived from the mathematical properties of a particular model comparison criterion. This thesis has been defended with respect to Akaike's model comparison criterion AIC (Forster and Sober 1994) – a criterion that is widely applied in ecological modeling and that has been praised by Forster and Sober in a number of papers. The next section sheds light on the rationale underlying AIC.

3 The AIC Approach to Simplicity

When evaluating the performance of a statistical model we are interested in how good the model will perform in the future. Still, we can only assess it with respect to the past – the fit with a given set of data. That a model which we fitted according to maximum likelihood estimation principles may fail to give reliable future predictions has been pointed out in the previous section. So evaluation criteria other than mere fit between data and candidate model are needed.

An influential idea, stemming from information theory, is to estimate the discrepancy between the candidate model and the unknown true model. A popular metric for this discrepancy is the *Kullback-Leiber divergence*

$$\begin{aligned} KL(f, g_\vartheta) &:= \int f(x)\, \log \frac{f(x)}{g_\vartheta(x)}\, dx \\ &= \int f(x)\, \log f(x)\, dx \,-\, \int f(x)\, \log g_\vartheta(x)\, dx \end{aligned} \tag{2}$$

[3]There, we choose the fitted regression model that makes the squares of the residuals most likely, or least surprising, compared to other candidate models.

where f is the probability density of the unknown true model f, g_ϑ is a class of candidate models indexed by parameter ϑ, and the integral is taken over the sample space (=the set of observable results). Kullback-Leiber divergence (Kullback and Leibler 1951; Shannon and Weaver 1949) is used in information theory to measure the loss of content when estimating the unknown distribution f by an approximating distribution g_ϑ.

Of course, we cannot compute KL-divergence directly for a given candidate model g_ϑ. First, we do not know the true probability density f. This implies that we have to *estimate* KL-divergence. — Second, g_ϑ is no single model, but stands for an entire class of models wherein ϑ is an adjustable fitting parameter. We have to use a particular element of g_ϑ for the estimation procedure. The *maximum likelihood estimator* $g_{\hat{\vartheta}}$ is a particularly natural candidate: it is the model whose parameter values maximize the likelihood of the data, given the model. However, if one used the maximum likelihood estimator to estimate KL-divergence without any corrective terms, one would overestimate the closeness to the true model. — Third, we are not interested in KL-divergence per se, but in predictive success. So we should relate (2) in some way to the predictive performance of a model. Akaike's (1973) famous mathematical result addresses these worries:

> **Akaike's Theorem:** For observed data y and a candidate model class g_ϑ with K adjustable parameters (or an adjustable parameter of dimension K), the model comparison criterion
>
> $$AIC(g_\vartheta, N) := -\log g_{\hat{\vartheta}(y)}(y) + K \tag{3}$$
>
> is an *asymptotically unbiased estimator* of $\mathbb{E}_x\mathbb{E}_y[\log(f(x)/g_{\hat{\vartheta}(y)}(x))]$ – the "expected predictive success of $g_{\hat{\vartheta}}$".

To better understand the double expectation in the last term, note that the maximum likelihood estimate $g_{\hat{\vartheta}}$ is determined with the help of the data set y. Then, $g_{\hat{\vartheta}}$'s KL-divergence to the true model f is evaluated with respect to another set of data x. This justifies the name *predictive* success, and taking the expectation two times – over training data y and test data x – justifies the name *expected* predictive success.

In other words, AIC estimates expected predictive success by subtracting the number of parameters K from the log-likelihood of the data under the maximum likelihood estimate $g_{\hat{\vartheta}}$. It gives an *asymptotically unbiased estimate* of predictive success – an estimate that will, in the long run, center around the true value. The more parameters a model has, the more do we have to correct the MLE estimate in order to obtain an unbiased estimate. We are then to favor the

model which minimizes AIC among all candidate models. According to Forster and Sober,

> "Akaike's theorem shows the relevance of goodness-of-fit *and* simplicity to our estimate of what is true [...]: it shows how the one quantity should be traded off against the other." (Forster and Sober 1994, 11)

Moreover, they use Akaike's theorem to counter the (empiricist) idea that simplicity is a merely pragmatic virtue and that "hypothesis evaluation should be driven by data, not by *a priori* assumptions about what a 'good' hypothesis should look like [such as being simple, the author]" (loc. cit., 27). By means of Akaike's theorem, simplicity is assigned an empirical value and established as an epistemic virtue.

Forster and Sober's praise corresponds well to the successes AIC has celebrated in practice. Scientists have got used to it in order to tackle practical problems. It is recommended by authoritative textbooks on statistical model analysis (Burnham and Anderson 2002). Especially in applied ecological modeling, it is much more popular than competing rationales that penalize model complexity, such as the Bayesian Information Criterion (BIC).

I would like to partition Forster and Sober's claims in two groups. The more general claims they make concern the *qualitative* question from the introduction, namely the empirical value of simplicity-driven considerations. As already shown, they have a valid point there. Their specific claims are more controversial. We will see that the tradeoff between simplicity and goodness of fit given by (3) fails to acquire sufficient normative force from the statistical properties of Akaise's estimator. So the specific way of measuring and trading off simplicity codified in Akaike's theorem fails to answer the *quantitative* question from the introduction – how to tradeoff simplicity against goodness of fit when aiming at accurate predictions. What is more, Akaike's results do not guarantee the epistemic benefits of taking simplicity into considerations – for such a claim, additional assumptions are required. The following section makes that criticism explicit.

4 Akaike's Theorem Revisited

The literature on AIC is large, and a number of papers have found "counterexamples" to the AIC, in the sense of specifying circumstances where use of the AIC systematically fails to choose the right model when sample size increases, or where it is inferior to competing model comparison criteria (Findley 1991; Bandyopadhyay and Boik 1999; Taper 2004; Dowe et al. 2007). One of these

objections concerns the well-known case of nested polynomial models. This is a set of models where each model class corresponds to polynomials of different degrees, such as in equation (1). It has been observed that for increasing sample size, the AIC "overshoots" the true model, i.e. the selected model is more complex than the one that generated the data. This statistical property is called *inconsistency* and seems to be a major threat to the normative force of AIC.

Forster (2002) discusses this objection in detail. He argues that statistical inconsistency is only embarrassing if either of the following two conditions is met: estimating the dimension (complexity) of a model is an important goal in itself, or failure to correctly estimate the dimension yields predictive failure as well:

> "...[the number of adjustable parameters] is an artefact of the representation, and not something that one should be interested in estimating. [...] The fact that AIC overshoots K [=the dimension of the true model] does not exclude the possibility that AIC converges on the true model as $N \to \infty$." (Forster 2002, S130)

Thus, according to Forster, the inconsistency objection need not worry the defender of AIC. Indeed, I am inclined to accept his claim that often, knowing the dimension of the model is important only to the degree that it helps to make better predictions. But the predictive value of getting the dimension of the model right must not be underestimated. Having too many parameters deteriorates predictive performance because when fitting the adjustable parameters, we will typically "explain" effects in the data which are due to random sampling variation. This happens because we have more degrees of freedom than there is complexity in the data. Here, the fact that AIC overshoots the true model *is* a valid objection to the use of the criterion in the above circumstances. Asserting that AIC *might* converge on the correct model, as Forster does, is too weak a rebuttal.

The objection can neither be swayed away by claiming that consistency need not worry us because in practice, all models are wrong (Sober 2008, 91). Convergence to a model whose structure is significantly more complex than the structure of the data-generating process will lead to an inferior predictive performance – the very criterion that Forster and Sober use. The example following equation (1) is telling.

The failure of consistency, and the deficiencies of AIC with respect to other model comparison criteria (Dowe et al. 2007) prompt the question of what kind of assurance Akaike's Theorem gives us (Sprenger 2009). Forster and Sober (1994, 2010) state that AIC is an *unbiased* estimator: as the sample size goes to infinity, there is no systematic tendency in AIC to be either higher or lower

than the value it tries to estimate. Mathematically spoken, an estimator $\tilde{\tau}$ of a statistic τ is unbiased if and only if $E_x[\tilde{\tau}(x)] = \tau$. In other words, if we produce a large number of (independent) data simultaneously and apply AIC to each of the data sets, then the mean AIC value will converge to the target value, as the number of data sets increases. But I shall point out that this is by no means sufficient to lend normative force to AIC.

Note first that unbiasedness is not sufficient to ensure the goodness of an estimator. The goodness of an estimator – usually measured by the mean square error – can be written as the square of the bias (zero in the case of AIC) plus its variance. Unbiasedness does not ensure low variance – an unbiased estimator may dissipate far from the true value and be awfully bad in practice.

This objection may be countered by noting that unbiasedness is an advantage, *ceteris paribus*. Forster and Sober (2010) note that AIC and the Bayesian Information Criterion BIC just differ by a constant. And if estimators differ by a constant, they have the same variance, and the unbiased estimator has the lower mean square error. (This follows straightforwardly from the fact that mean square error = square of bias + variance.) Hence BIC seems to be a worse estimator of predictive accuracy than AIC.

However, this argument is based on an oversight which many authors in the debate (Forster and Sober 1994, 2010; Kieseppä 1997) commit. AIC is *not* an unbiased estimator – it is just *asymptotically* unbiased, in other words, the property of unbiasedness is only realized for very large samples.[4]

To see this with your own eyes, I invite you to have a look at the mathematical details in appendix A. There, the dependence of Akaike's Theorem on the *asymptotical*, and not the actual normality of the maximum likelihood estimator becomes clear. This has substantial consequences, and speaks against a normative interpretation of Akaike's findings. AIC outperforms BIC as an estimator of predictive accuracy only for an infinitely large sample, whereas *actual* applications usually deal with medium-size finite samples. As long as we don't know the speed of convergence – and this varies from data set to data set –, the asymptotic properties are unwarranted.

Besides, the superiority argument in favor of AIC presupposes that predictive accuracy is measured by KL-divergence. But there is a wide choice of disparity measures, and choosing a different measure will lead to different conclusions.

[4]To their excuse it should be said that the pertinent textbook by Sakamoto et al. (1986, 69) sometimes uses this formulation in passing. Several other passages (Sakamoto et al. 1986, 65,77,81) make clear, however, that the word "unbiased", when applied to AIC, is merely used as a shortcut for "asymptotically unbiased".

Some of them, like the Hellinger distance[5] have been proposed as reasonable alternatives to KL-divergence (Lele 2004), partly because they can, unlike KL-divergence, act as a *metric* between different models.

Finally, the contribution of simplicity relative to goodness of fit in Akaike's Theorem diminishes as sample size N increases, as Forster (2002) notes himself (see appendix A for details). The goodness-of-fit term is of the order of N whereas the simplicity-based contribution remains constant. Thus, with increasing sample size, simplicity drops out of the picture, and it becomes increasingly difficult to meaningfully compare different estimators, such as AIC and BIC.

Taking all these observations together, I conclude that asymptotic unbiasedness is at the very best a necessary property for a reasonable estimator (we do not want systematic mistakes in large samples), but by no means sufficient in order to demonstrate the adequacy of that particular estimator. Akaike's Theorem is of high mathematical interest and deserves credit for introducing methods of information theory into statistical model analysis, but mathematically, it is too weak to assign a definite weight of simplicity in statistical model analysis, or to show that AIC usually outperforms other model comparison criteria. With the vanishing normative force of the theorem, there is no real basis for far-reaching methodological claims that Forster and Sober make, e.g. regarding the use of Bayesian methods in statistics (Forster and Sober 1994; Forster 1995). Rather, applications of AIC need to be justified by their *practical* successes. Modern defenders of AIC, such as Burnham and Anderson (2002), have adopted that strategy:

> "it is clear that there are different conditions under which AIC and BIC should outperform the other one in measures such as estimated mean square error." (Burnham and Anderson 2004, 287)

It is surprising that Akaike's Theorem has achieved such a high degree of prominence in philosophy of science, although its methodological ramifications are sparse, and although there is a variety of competing approaches that also take simplicity into account. This section is not meant to push simplicity back into the realm of the non-empirical, nor to argue that AIC should not be applied in practice. It just refutes the claim that AIC provides, *in virtue of Akaike's Theorem*, a vindication of simplicity as an empirical virtue. The next section addresses the comparative question from the introduction and compares AIC to other prominent model comparison criteria.

[5]The Hellinger distance is defined as

$$H^2(f,g) := \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g_\vartheta(x)}\right)^2 dx. \tag{4}$$

5 AIC vs. BIC vs. DIC: A Comparative Analysis

The previous section has suggested that there is more than one reasonable simplicity-based criterion to compare statistical models. Sober (2008) views these different criteria akin to scientific theories whom we assess on the basis of their merits in solving relevant problems. This section explores to what extent they are not only different *approaches*, but also aim at different *targets*, revealing salient methodological differences and the strong context-dependence of the choice of an adequate criterion.

Model comparison criteria are often classified as either Bayesian or frequentist. As the derivation of AIC in the previous section made clear, AIC does not use a Bayesian justification, but draws on frequentist sampling properties. However, it is possible to generalize the result from a Bayesian perspective (Spiegelhalter et al. 2002, 604–605). The motivation for doing so is practical. We often deal with complex, hierarchical models where the parameter ϑ is specified by a given probability distribution that depends on a hyperparameter ψ. Then, it is not clear how to measure the complexity of the model: should we base it on the dimension of ϑ, the dimension of ψ or an aggregate of both? The straightforward measurement of simplicity as the number of free parameters, that was used in the case of AIC, is ambiguous and runs into problems. Therefore, we have to search for alternatives. A natural proposal is to cash out the connection between model complexity and estimation variance that we have elaborated in section 2: complexity is understood as "difficulty in [parameter] estimation" (Spiegelhalter et al. 2002, 585) and manifests itself in the reduction of uncertainty which the estimation process yields. Concretely, model complexity is estimated by

$$p_D = \mathbb{E}_{\vartheta|y}[-2 \log g_{\vartheta}(y)] + 2 \log g_{\hat{\vartheta}(y)}(y), \tag{5}$$

where $-\log g(x)$ gives the score for the surprise or *deviance* in data x relative to model g. Thus, p_D can be understood as the difference between the expected surprise in the data and the surprise we observe under a standard estimate of ϑ (typically the mean of the posterior distribution). Put another way, p_D measures the extent to which our estimate $\hat{\vartheta}(y)$ is expected to overfit the data. The Deviance Information Criterion (DIC) – the model comparison criterion in which p_D figures – does not depend on the number of model parameters, but on the model's fitting properties, which makes it tailor-made for applications in hierarchically structured modeling problems! We see that the context and the target of the modeling problem are crucial to determining how to understand simplicity and which criterion to apply – especially given the failure of rigorous mathematical superiority results.

It is noteworthy that DIC contains Bayesian elements: expectations in (5) are taken with respect to the posterior distribution of ϑ, and the parameter estimate is based on the posterior as well. Another model comparison criterion motivated by Bayesian considerations the Bayesian Information Criterion (BIC) (Schwarz 1978). BIC aims at the *class of models* with the highest posterior probability, not a single best *fitted model*. Schwarz develops an asymptotic approximation of the posterior probability of a model class g_ϑ. To that end, it is assumed that all probability densities belong to a particular subclass of the exponential family, and that the density with respect to the Lebesgue measure μ can be written as

$$L_x(\vartheta) = e^{N(A(x)-\lambda|\vartheta-\hat{\vartheta}|)^2}.$$

Then the posterior probability of the model class (not the fitted model!) g_ϑ can be written as the product of prior probability and likelihood of the data x^*:

$$\begin{aligned}\mathbb{P}(g_\vartheta|x^*) &= \mathbb{P}(g_\vartheta)\int_{\vartheta\in\Theta} e^{N(A(x^*)-\lambda|\vartheta-\hat{\vartheta}|)^2} d\mu(\vartheta) \\ &= \mathbb{P}(g_\vartheta)e^{NA(x^*)}\int_{\vartheta\in\Theta} e^{-N\lambda|\vartheta-\hat{\vartheta}|^2} d\mu(\vartheta).\end{aligned}$$

Substituting the integration variable ϑ by $\vartheta/\sqrt{n\lambda}$, and realizing that for the maximum likelihood estimate $\hat{\vartheta}$, $L_x(\hat{\vartheta}) = e^{NA(x)}$, we obtain a formula that is already quite close to BIC (see Lemma 1 in Schwarz (1978)):

$$\begin{aligned}\log\mathbb{P}(g_\vartheta|x^*) &= \log\mathbb{P}(g_\vartheta) + NA(x^*) + \log\left(\frac{1}{N\lambda}\right)^{K/2} + \log\int_{\vartheta\in\Theta} e^{-|\vartheta-\hat{\vartheta}|^2} d\mu(\vartheta) \\ &= \log\mathbb{P}(g_\vartheta) + NA(x^*) + \frac{1}{2}K\log\left(\frac{1}{N\lambda}\right) + \log\sqrt{\pi}^K \\ &= \log\mathbb{P}(g_\vartheta) + \log\mathbb{P}(x^*|\hat{\vartheta}) - \frac{1}{2}K\log\left(\frac{N\lambda}{\pi}\right) \qquad (6)\end{aligned}$$

On the left hand side, we have the the log-posterior probability, a Bayesian's standard model comparison criterion. As we see from (6), 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 parameters. This derivation, whose assumptions are relaxed subsequently in order to yield more general results, forms the mathematical core of BIC. Similar to AIC, the number of parameters K enters the calculations because the expected likelihood depends on the model dimension (via the skewness of the likelihood function).

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 that make only constant contributions and focus on the terms that increase in N. In the long run, the model with the highest posterior probability will be the model that minimizes

$$BIC(g_\vartheta, N) = -2 \log g_{\hat{\vartheta}}(x^*) + K \log N. \tag{7}$$

Thus, BIC neglects the contribution of the priors ($\log \mathbb{P}(g_\vartheta)$), when comparing the models to each other. It is then questionable whether BIC should be described as a *subjective* Bayesian technique, as opposed an allegedly objective, non-Bayesian AIC.

Let's make the critique of that dichotomy explicit. First, there are striking methodological parallels: Both criteria study the sampling properties of the maximum likelihood estimator. Both are specific large-sample approximations where the goodness of the approximation depends on the nature of the asymptotics. For both criteria, the relationship between sample size and model dimension, and specific assumptions on the nature of the probability densities, enter the mathematical motivations. Second, subjective Bayesian positions typically take prior probabilities seriously, and consider them to be inferentially relevant. However, the prior probabilities in BIC are *not* understood as subjective prior beliefs that need to be elicited: rather they drop out of the estimation procedure (see the transition from (6) to (7)). Neither is the derivation of BIC committed to the true model being in the set of candidate models which is a standard premise of Bayesian convergence theorems. All this shows that for BIC, Bayesianism constitutes no philosophical underpinning (e.g. as a logic of belief revision), but only a convenient framework which motivates that use of specific estimators of log-posterior probability. The Bayesianism involved in the derivation of BIC might be characterized as an *instrumental Bayesianism* – an approach which uses Bayes's Theorem to conveniently model a scientific problem and to obtain an attractive mathematical result, but without taking the Bayesian elements as seriously and literally as a real, subjective Bayesian would do. Forster's (1995) criticism of Bayesian model comparison that scientists are not interested in posterior probabilities, but in predictive success, is therefore an idle wheel.

This instrumental conception of Bayesianism is rather widespread in statistics. Even the work of many famous Bayesian statisticians, such as James O. Berger or José Bernardo, often contradicts (subjective) Bayesian principles, because they frequently use "objective" or improper reference priors (i.e. priors that do not sum up to one, or that are stipulated instead of elicited). Therefore, the debate between AIC and BIC should not be recast as a debate between Bayesians and frequentists. Utterances such as

> "what fundamentally distinguishes AIC versus BIC model selection

is their different philosophies" (Burnham and Anderson 2004, 284)

must not be misread as suggesting a dichotomy between Bayesian and non-Bayesian methods. The intermediate case of DIC that aims at the performance of fitted models (like AIC), while being embedded in a Bayesian framework, makes clear that such a dichotomy is elusive. Neither is it impossible to restate AIC in a non-parametric framework, or as a Bayesian model comparison with specific priors (Stone 1977; Bandyopadhyay and Boik 1999). Rather, what we should stress are the different *targets* of AIC and BIC. Where one criterion (AIC) tries to estimate a quantity that may be interpreted as expected predictive success under a specific divergence measure, the other one (BIC) tries to elicit which class of models will, in the long run, be favored by the data. In other words, BIC estimates the *average* performance of different model classes whereas AIC compares *representatives* of each model class, namely the maximum likelihood estimates. So does DIC for the special case of a hierarchical model with focus on uncertainty on a medium level.

Apart from the different targets, the chosen model comparison procedure is also affected by expectations on the structure of future data. This can be seen most forcefully when considering the issue of sample size. The more the data tell us, the smaller the role of simplicity-based considerations in the model comparison process. This insight stands behind quotes such as "we *cannot* ignore the degree of resolution of the experiment when choosing our prior," (Williams 2001, 235), when determining the weight of simplicity. It is also part of the mathematics of our model comparison criteria: the log-likelihood term is of the order N whereas the simplicity term is of the order $\log N$ (BIC) or 1 (AIC). Moreover, if the data structure is likely to change over time, we will not be interested in a specific fitted model for predictive purposes – the target of AIC. Sample size, tapering effects, presence of nested models, overdispersed data, etc. are further important considerations that affect the weight of simplicity, and the choice of a model comparison criterion.

Thus, it is misleading to see the contest of different model comparison criteria as a contest between different philosophical schools of inductive inference. It is much more plausible, and does more justice to scientific practice, to let pluralism reign. Statisticians have recognized that "overformal" approaches which apply simplicity mechanically, are not suited to the multitude of inferential contexts and goals we have to deal with (Spiegelhalter et al. 2002, 602).[6] The comparative question from the introduction cannot be answered generally: the adequacy of a

[6] A genuinely Bayesian approach (Kass and Raftery 1995) is not immune to these worries either. As demonstrated by Han and Carlin (2001), model comparison based on Bayes factors is usually computationally expensive and based on the assumption that the set of candidate models remains fixed with increasing sample size. This assumption need not be realistic.

criterion is a function of the objective of the model analysis, the structure of the data, and further problem specifics. In this way, choosing an appropriate model comparison procedure can be regarded as a matter of scientific judgment and sound experimental design, in the same way that the choice of a stopping rule in sequential trials calibrates a statistical cost-benefit analysis with the specifics of a given scientific problem.

Summing up, we see that statistical model analysis requires a synthetic perspective: scientific understanding determines the kind of assumptions and projections that we should reasonably make, statistical sophistication develops the adequate mathematical tools for comparing models in those circumstances. This implies that scientists and statisticians have to collaborate in statistical model comparison, and that a clear-cut division of labor is impossible.

6 Summary

The paper started with three questions: a qualitative question about the empirical significance of simplicity, a quantitative question about its weight in model comparison, and a comparative question about how the different simplicit-based model comparison criteria relate to each other. These questions have been investigated in the context of statistical model analysis.

Our analysis yields that neither AIC, BIC nor DIC directly substantiate an understanding of simplicity as an empirical, truth- or prediction-conducive virtue by means of some waterproof mathematical results. Rather, they are approximative estimators of relevant quantities, and the results that link them to their targets are not general enough to secure universal, mechanical application. In particular, we reject Forster and Sober's claim that Akaike's theorem vindicates simplicity as an empirical virtue, and provides the right rationale to measure it. The asymptotic results of that theorem are neither strong enough to establish how much weight simplicity should carry, nor are they based on uncontentious assumptions.

However, this does not mean that simplicity is pushed back again into the realm of the pragmatic, non-empirical – there is a neat rationale for avoiding complex and preferring simple models, ceteris paribus, namely the high estimation variance that complex models possess. A unified analysis is, however, hampered by the lack of compelling mathematical results, the observation that different procedures will succeed in different circumstances, and the lack of sufficiently general, goal-independent benchmarks. These differences are much more important than whether the criteria are motivated from a Bayesian or a frequentist perspective: Since Bayesianism is employed instrumentally, i.e. as a convenient mathematical framework, but without commitments to substantial

philosophical claims, the emphasis of past debates on AIC vs. BIC, on Bayesian vs. frequentist model comparison, is misplaced.

Thus, while there is no definite answer to the quantitative question, the qualitative question is answered in the affirmative. Finally, an answer to the comparative question sensitively depends on context, target and specifics of the modeled problem. An experimental design perspective is, to my mind, the best way to compare different ways of taking simplicity into account. The down-to-earth story which this paper tells makes it difficult to extract ramifications of statistical model analysis to general issues in philosophy of science, disappointing some hopes that flourished in the past. But such a sober conclusion is perhaps just what we can and should expect.

A Sketch of the derivation of Akaike's information criterion

The goal of the estimation problem is the estimation of the "expected predictive success"

$$\mathbb{E}_x\mathbb{E}_y\left[\log\frac{f(x)}{g_{\hat{\vartheta}(y)}(x)}\right] \quad = \quad \mathbb{E}_x\mathbb{E}_y\left[\log f(x)\right] - \mathbb{E}_x\mathbb{E}_y\left[\log g_{\hat{\vartheta}(y)}(x)\right]. \tag{8}$$

The first term on the right hand side of (8) is equal for all candidate models. When comparing them, it drops out as a constant. Hence we can neglect it in the remainder and focus on the second term in (8).

The standard derivation of AIC proceeds by a double Taylor expansion of the log-likelihood function that relates our maximum likelihood estimate $\hat{\vartheta}$ to the "optimal" parameter value ϑ_0[7] – the parameter value that minimizes Kullback-Leibler divergence to the true model. The term $\log g_{\hat{\vartheta}(x)}(y)$ is expanded around ϑ_0 and the expansion is trunctated at the second order term,[8] yielding

$$\begin{aligned}\log g_{\hat{\vartheta}(y)}(x) \quad \approx \quad & \log g_{\vartheta_0}(x) + N\left(\left(\frac{\partial}{\partial\vartheta}\log g_{\vartheta}(x)\right)(\vartheta_0)\right)(\hat{\vartheta}(y)-\vartheta_0) \\ & + \frac{1}{2}N(\hat{\vartheta}(y)-\vartheta_0)^T\left(\left(\frac{\partial^2}{\partial\vartheta^2}\log g_{\vartheta}(x)\right)(\vartheta_0)\right)(\hat{\vartheta}(y)-\vartheta_0). \end{aligned} \tag{9}$$

The matrix

$$J := -\frac{\partial^2}{\partial\vartheta^2}\log g_{\vartheta}(x)(\vartheta_0) \tag{10}$$

[7] See Chapter 4.3 in Sakamoto et al. (1986) and Chapter 7.2 in Burnham and Anderson (2002).

[8] The general formula of Taylor expansion for analytic real-valued functions f is

$$f(x) = \sum_{k=0}^{\infty} f^{(k)}(x_0)(x-x_0)^k.$$

that also occurs in (9) is called the *Fisher information matrix* of the data. It plays a crucial role in an asymptotic approximation of the maximum likelihood estimator that holds under plausible regularity conditions:

$$\sqrt{N}(\hat{\vartheta}(y) - \vartheta_0) \quad \to \quad \mathcal{N}(0, J^{-1}). \tag{11}$$

This asymptotic normality of the maximum likelihood estimator can be used to simplify (9). The term

$$\sqrt{N}(\hat{\vartheta}(y) - \vartheta_0)^T \, (-J) \, \sqrt{N}(\hat{\vartheta}(y) - \vartheta_0) \tag{12}$$

is asymptotically χ^2-distributed with K degrees of freedom. Hence, the expectation of (12) is K. By taking a double expectation over x and y, we thus obtain that

$$\mathbb{E}_x \mathbb{E}_y \left[\frac{1}{2} N (\hat{\vartheta}(y) - \vartheta_0)^T \left(\left(\frac{\partial^2}{\partial \vartheta^2} \log g_\vartheta(x) \right) (\vartheta_0) \right) (\hat{\vartheta}(y) - \vartheta_0) \right] \quad \approx \quad \frac{K}{2} \tag{13}$$

Moreover, the linear term in (9) vanishes because the maximum likelihood estimate is an extremal point of the log-likelihood function. Thus, the mean of the first derivative is also zero:

$$\mathbb{E}_x \mathbb{E}_y \left[N \left(\left(\frac{\partial}{\partial \vartheta} \log g_\vartheta(x) \right) (\vartheta_0) \right) \right] \quad = \quad 0 \tag{14}$$

Combining (9) with (13) and (14), we obtain for large samples that

$$\mathbb{E}_x \mathbb{E}_y \left[\log g_{\hat{\vartheta}(y)}(x) \right] \quad \approx \quad \mathbb{E}_x \left[\log g_{\vartheta_0}(x) \right] - \frac{K}{2}. \tag{15}$$

Repeating the Taylor expansion around the maximum-likelihood estimate and applying the same arguments once more gives us

$$\mathbb{E}_x \mathbb{E}_y \left[\log g_{\vartheta_0}(x) \right] \quad \approx \quad \mathbb{E}_y \left[\log g_{\hat{\vartheta}(y)}(y) \right] - \frac{K}{2}. \tag{16}$$

Finally, by combining (15) and (16) we obtain AIC as an estimate of "expected predictive accuracy":

$$\mathbb{E}_x \mathbb{E}_y \left[\log g_{\hat{\vartheta}(y)}(x) \right] \quad \approx \quad \mathbb{E}_y \left[\log g_{\hat{\vartheta}(y)}(y) \right] - K. \tag{17}$$

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. Akademiai Kiado, Budapest.

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.

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

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., 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.

Findley, David F. (1991): "Akaike's Information Criterion and Recent Developments in Information Complexity", *Annals of the Institute of Statistical Mathematics* 43, 505–514.

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 (2000): "Key Concepts in Model Selection: Performance and Generalizability", *Journal of Mathematical Psychology* 44, 205–231.

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

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 (2010): "AIC Scores as Evidence – A Bayesian Interpretation", forthcoming in Malcolm Forster and Prasanta S. Bandyopadhyay (eds.): *The Philosophy of Statistics*. Kluwer, Dordrecht.

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.

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

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

Keuzenkamp, Hugo (2000): *Probability, Econometrics and Truth*. Cambridge University Press, Cambridge.

Kieseppä, Ilkka A. (1997): "Akaike Information Criterion, Curve-fitting and the Philosophical Problem of Simplicity", *British Journal for the Philosophy of Science* 48, 21-48.

Kullback, S., and R. A. Leibler (1951): "On information and sufficiency", *Annals of Mathematical Statistics* 22, 79–86.

Lele, Subhash (2004): "Evidence Functions and the Optimality of the Law of Likelihood", in: Mark Taper, Subhash Lele (ed.), *The Nature of Scientific Evidence*, 191-216 (with discussion). The University of Chicago Press, Chicago & London.

Mikkelson, Gregory M. (2007): "Realism vs. Instrumentalism in a New Statistical Framework", *Philosophy of Science* 73: 440–447.

Sakamoto, Y., M. Ishiguro and G. Kitigawa (1986): *Akaike Information Criterion Statistics*. Reidel, Dordrecht.

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

Shannon, Claude, and Warren Weaver (1949): *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Sober, Elliott (2002): "Instrumentalism, Parsimony, and the Akaike Framework", *Philosophy of Science* 69: S112–S123.

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

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.

Sprenger, Jan (2009): "Statistics between Inductive Logic and Empirical Science", *Journal of Applied Logic* 7, 239–250.

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.

Taper, Mark (2004): "Model Identification from Many Candidates", in: Mark Taper, Subhash Lele (ed.), *The Nature of Scientific Evidence*, 488–524 (with discussion). The University of Chicago Press, Chicago & London.

Weisberg, Michael (2007): "Who is a Modeler?", *British Journal for the Philosophy of Science* 58, 207–233.

Williams, David (2001): *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press, Cambridge.