I. A. KIESEPPÄ:

STATISTICAL MODEL SELECTION CRITERIA AND BAYESIANISM

DEPARTMENT OF PHILOSOPHY

P. O. BOX 24 (UNIONINKATU 24)

00014 UNIVERSITY OF HELSINKI

E-MAIL: KIESEPPA@CC.HELSINKI.FI

TEL. (OFFICE): +358 9 191 23805

TEL. (MOBILE): +358 40 748 8991

FAX: +358 9 191 7627

I. A. Kieseppä:

Statistical Model Selection Criteria and Bayesianism†

ABSTRACT

Two Bayesian approaches to choosing between statistical models are contrasted. One of these is an approach which Bayesian statisticians regularly use for motivating the use of AIC, BIC, and other similar model selection criteria, and the other one is a new approach which has recently been proposed by Bandyopadhayay, Boik, and Basu. The latter approach is criticized, and the basic ideas of the former approach are presented in a way which makes them accessible to a philosophical audience. It is also pointed out that the former approach establishes a new, philosophically interesting connection between the notions of simplicity and informativeness.

1. Introduction

  In recent years philosophers of science have shown some interest for statistical model selection criteria. There are several reasons for such interest. Perhaps the most obvious of these is the fact that in many of the typical applications of model selection criteria the methodological rules which are

associated with them instruct us to behave in accordance with the traditional idea that *simple models should preferred to more complicated ones, other things being equal*. Hence, the theoretical justifications of the model selection criteria seem to justify this traditional methodological doctrine in some important special cases.

Statistical model selection criteria are quantities which depend on the available measurement results and on the properties of the considered model, and they are associated with methodological rules which state that one should choose the model for which the criterion receives its smallest value. The two most popular model selection criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[1] The use of AIC has originally been motivated by the wish to maximize the value of a quantity which philosophers have called the *predictive accuracy* of the model (see Forster and Sober, 1994; cf. Kieseppä, 1997), and the use of BIC is motivated by the Bayesian idea that one should choose the model with the largest *posterior probability*.

In Bandyopadhayay, Boik and Basu (1996) it was claimed that one could use this Bayesian idea for defending the use of a large variety of different ways of choosing between models. The way in which Bandyopadhayay, Boik and Basu arrive at this conclusion has subsequently been criticized by the statistician Jouni Kuha in Kuha (*submitted*). Kuha also points out in *ibid.* that there is a more standard Bayesian argument with which one can arrive at a similar conclusion. This argument can be used for motivating not only the use of BIC, but also the use of AIC and various other information criteria. It is also quite interesting philosophically, among other reasons because it establishes a connection of a new kind between the informativeness and the simplicity of the

---

[1] For standard introductions to the statistical methods based on AIC see e.g. Sakamoto et al. (1986) and Burnham–Anderson (1998). The use of BIC is discussed at length in e.g. Raftery (1995).

considered models.

However, Kuha's presentation of this argument is very dense, rather technical and, accordingly, for most philosophers quite difficult to follow. It is badly in need of being complemented by a discussion of the same topics which is more accessible to a philosophical audience and which emphasizes the philosophically relevant aspects of its subject matter. Such a complementation will be presented below.
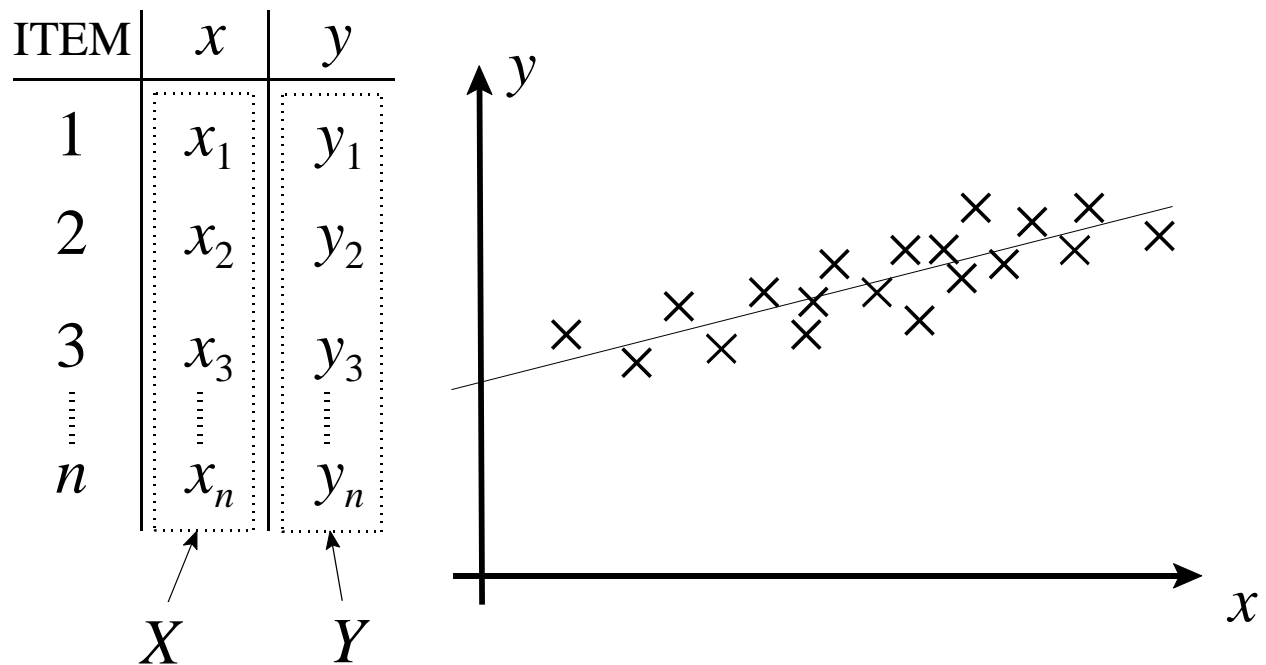

Figure 1.

2. AIC, BIC and Curve-Fitting

Until now philosophers have discussed the use of statistical model selection criteria mostly in the context of *curve-fitting problems*, and below I shall follow this practice. Of course, these problems

constitute only a minor subset of the set of all the problems to which the criteria can be applied.

Curve-fitting problems are problems of finding a curve which expresses the connection between two quantities – which we shall call x and y – on the basis of a finite sample within which one has measured their values. Such measurement results can be represented by tabulating the observed x values and the observed y values for each item in the sample in the way which is illustrated by Figure 1. As the figure indicates, we shall below use the capital letters X and Y for referring respectively to the list of the observed x values and the list of the observed y values that would appear in such a table.

The measurement results can, of course, also be represented as points in a coordinate system. The task of picking up a curve – like e.g. the straight line in Figure 1 – which one supposes to express the connection between x and y can be divided into two parts: first one chooses a *statistical model* for the connection of x and y, and then one picks up its best-fitting curve. Each of such models specifies *a family of curves* and claims that one of its curves is the "true curve" in the sense that Y – *i.e.* the observed combination of y values – has a conditional probability distribution which is centred around this curve. The family of curves in question might e.g. be *the family of all straight lines*, or *the family of all parabolas*. The curves which a model allows for are normally identified by a list of quantities, the *parameters* of the model. As a rule, the larger the number of parameters of a family of curves is, the larger is the variety of different curves that it allows for. For example, the model which claims that the true curve is a straight line has in its standard representation two parameters, but the model which claims that the true curve is a parabola has three of them.

When the parameters of the model **M** are denoted by $\alpha_1, \alpha_2, ..., \alpha_k$, each combination of the values of $\alpha_1, \alpha_2, ..., \alpha_k$ corresponds to a curve, which further corresponds to a probability distribution of the y

values in *Y* which is conditional on the *x* values in *X*.[2] This distribution can be denoted by[3]

   prob(*Y given X*, **M**, *and* α_1, α_2,..., α_k).

The above quantity is at the same time the *likelihood* of the curve which corresponds to the parameter

values α_1, α_2,..., α_k. The curve which has the largest likelihood is normally taken to be the best-fitting

curve within its model. The parameter values which correspond to this curve are often denoted by

$\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$. Using this notation, the definitions of the AIC and BIC values of an arbitrary model **M**

can be expressed as follows (see e.g. Burnham and Anderson, 1998, 46 and 68):[4]

(1)   AIC(**M**) = $^{-}$2 (*logarithm of prob*(*Y given X*, **M**, *and* $\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$))

            +2(*number of parameters of* **M**)

(2)   BIC(**M**) = $^{-}$2 (*logarithm of prob*(*Y given X*, **M**, *and* $\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$))

            + (*logarithm of n*)(*number of parameters of* **M**)

---

[2] To keep things simple, we are focusing our attention on the case in which according to each considered model *the error distribution* – *i.e.* the *probability distribution* of the difference of the observed *y* value and the *y* value which corresponds to the observed *x* value on the true curve – is a fixed normal distribution with some known variance σ^2. In this case each of the curves of each of the models corresponds to precisely one probability distribution of the *y* values.

[3] To be quite precise, the quantity *prob*(*Y given X*, **M**, *and* α_1, α_2,..., α_k) is not the *probability* of *Y*, when *X*, **M**, and α_1, α_2,..., α_k have been given. Rather, it expresses the *probability density* of Y when *X*, **M**, and α_1, α_2,..., α_k have been given. An analogous remark applies to also all the other quantities of the form *prob*(... *given* ...) which are mentioned in this paper.

[4] In these formulas, and also elsewhere in this paper, the word 'logarithm' refers to the *natural logarithm*.

Here *n* is the number of items in the sample. As already stated, the criteria are associated with methodological rules which state that one should choose the model with the smallest criterion value:

(Rule‒AIC)   Among the considered models choose the model **M** for which AIC(**M**) receives its smallest value!

(Rule‒BIC)   Among the considered models choose the model **M** for which BIC(**M**) receives its smallest value!

The logarithm which occurs in the first term of the expressions of AIC(**M**) and BIC(**M**) is the *log likelihood* of the best-fitting curve of the considered model, and it can be viewed as a measure of the *fit* between the model and evidence. Since this measure of fit is *multiplied by a negative number* in the two formulas, (Rule‒AIC) and (Rule‒BIC) both instruct us to prefer models which fit the evidence well. The latter terms of the two expressions are different, but each of them is the product of the *number of the parameters of the model* and a quantity which has the same value for all models. The number of the parameters of a model can be viewed as a measure of its complicatedness, since a model with a small number of parameters is, as a rule, simple in the sense that it allows a smaller variety of curves than a model with a large number of parameters. If one thinks about the number of parameters in this way, both AIC and BIC will be seen to give an advantage to simple models over complicated ones, since according to both of the two criteria, when two models fit the evidence approximately equally well, the simpler model should be preferred. However, the amount of such advantage is different in the two criteria.

3. The Bayesian Approach to Model Choice

   A Bayesian approach to model choice is based on the idea that the model with the *largest posterior probability* should be chosen. The posterior probability of a model **M** is its probability *after* the the evidence has become known, and it depends on the one hand on how well the available evidence fits **M**, and on the other hand on the *prior probability* of **M**. This probability, which we shall denote by *prior*(**M**), is the probability that **M** had *before* the evidence became available.

   In our current context the posterior probability of a model **M** is *the probability of* **M,** *given that the observed y and x values are the ones in Y and X*, respectively. When this probability is denoted by *probability*(**M** *given X and Y*), the basic idea of a Bayesian approach to model selection can be formulated as the following methodological recommendation:

(B)   Choose the model for which *probability*(**M** *given X and Y*) is largest among the considered models **M**!

Below we shall denote the probability distribution of *Y* – i.e. of the observed *y* values – given that the observed *x* values are the ones in *X* and given that **M** is correct, by *prob*(*Y given X and* **M**). The Bayes formula implies that *probability*(**M** *given X and Y*) is proportional to the product of the *prior*(**M**) and *prob*(*Y given X and* **M**), so that the rule (B) can be reformulated as follows:

(B′)   Choose the model for which *prior*(**M**)×*prob*(*Y given X and* **M**) is largest among the considered models **M**!

In order to be able to apply this methodological recommendation one has to fix the prior probabilities of the considered models, and one has to calculate *prob*(*Y given X and* **M**). This quantity is different from the quantities

$$prob(Y\ given\ X,\ \mathbf{M},\ and\ \alpha_1,\alpha_2,...,\alpha_k)$$

which we discussed earlier and which were associated with the *individual curves* of the model **M**, rather than with the whole model **M**. In Bayesian statistics the probabilities of the former type are calculated by introducing separately for each model a *prior probability distribution of the parameter values*. Such a distribution can be denoted by

$$prior(\alpha_1,\ \alpha_2,...,\ \alpha_k\ given\ \mathbf{M}).$$

Together with the known distribution

$$prob(Y\ given\ X,\ \mathbf{M},\ and\ \alpha_1,\ \alpha_2,...,\ \alpha_k),$$

the prior distribution of the parameters suffices to determine the value of the quantity

$$prob(Y\ given\ X\ and\ \mathbf{M})$$

which occurs in (B′). Of course, different choices of the prior distribution lead to different values of *prob*(*Y given X and* **M**) and, accordingly, to different choices between models when the rule (B′) is applied. Next we shall have a quick look at the way Bandyopadhayay. Boik, and Basu choose this distribution, and we then turn to a discussion of a more standard way of choosing it.

4. The Approach of Bandyopadhayay, Boik, and Basu

In order to provide a unified treatment for the argument of Bandyopadhayay *et al.* and its more customary alternative we shall put the rule (B′) into a form in which its connection with the statistical

model selection criteria is easy to see. First we observe that since the *logarithm function* is an increasing function, the rule (B′) can be formulated also by saying that one should choose the model **M** for which

$$logarithm\ of\ [\ prior(\mathbf{M}){\times}prob(Y\ given\ X\ and\ \mathbf{M})]$$

is largest. Secondly, since the logarithm function "converts products into sums" in the sense that

$$logarithm\ of\ [A{\times}B] = [logarithm\ of\ A] + [logarithm\ of\ B]$$

for any numbers *A* and *B*, the rule (B′) is further equivalent with a rule which instructs us to choose the model for which the value of

$$[logarithm\ of\ prior(\mathbf{M})] + [logarithm\ of\ prob(Y\ given\ X\ and\ \mathbf{M})]$$

is largest. This recommendation is clearly equivalent with the following rule:

(B″)   Choose the model for which

$$(-2)[logarithm\ of\ prob(Y\ given\ X\ and\ \mathbf{M})] + (-2)[logarithm\ of\ prior(\mathbf{M})]$$

is smallest among the considered models **M**!

A comparison of (B″) and the definitions of BIC(**M**) and AIC(**M**) suggests an obvious way in which one could give a Bayesian defence to the use of these and other information criteria. The rule (B″) would be equivalent with e.g. (Rule–BIC) if the prior distributions of parameters were such that, firstly, the difference of

$$logarithm\ of\ prob(Y\ given\ X\ and\ \mathbf{M}),$$

which occurs in (B″), and

$$logarithm\ of\ prob(Y\ given\ X,\ \mathbf{M},\ and\ \hat{\alpha}_1,\ \hat{\alpha}_2,...,\ \hat{\alpha}_k\ ),$$

which occurs in the definition of BIC(**M**), was a constant which had the same value for all the considered models, and if, secondly, the difference of

$$(-2)[\text{ logarithm of prior}(\mathbf{M}) \text{ }]$$

and

$$(\text{logarithm of } n)(\text{number of parameters of } \mathbf{M})$$

was a constant which had the same value for all the considered models.

In Bandyopadhayay et al. (1996) one considers a family of prior distributions which depends on a parameter τ. The parameter has been chosen in such a way that the former of the above conditions becomes valid in the limit in which τ approaches infinity.[5] Hence, the approach of Bandyopadhayay et al. seems to enable one to give a new Bayesian defence for BIC. Of course, if this defence is acceptable, it can easily be modified so that it turns into an acceptable defence of AIC or of some other similar criterion: one just has to keep the prior distributions of the parameters within each model the way they are, but change the prior probabilities of the models so that they produce the criterion in question.

However, Kuha (*submitted*) contains several criticisms of the approach of Bandyopadhayay et al. The most obvious of these is that their results are concerned with the case in which $\tau=\infty$, but the value ∞ of the parameter τ does not correspond to legitimate prior distibution. My refusal to give this criticism the same weight that Kuha gives to it is based on the fact that it is easy to modify the argument of Bandyopadhayay et al. in such a way that this criticism no longer applies to it: one can

---

[5] The validity of the former condition follows from the fact, which is stated in Bandyopadhayay et al. (1996, 268), that with their choice of priors the posterior probability of each model is in the limit in which $\tau \to \infty$ proportional to the product of its prior probability and its maximum likelihood.

choose a prior which corresponds to some very large, finite value of τ and observe that their results

must be *almost exactly* valid for this legitimate prior. Kuha makes also another criticism which is,

in my view, more serious: Bandyopadhayay et al. assume that there is much more prior information

concerning the values of the parameters of the large models than concerning the values of the

parameters of the small models. There seems to be no other reason for making this assumption than

the fact that, if Bandyopadhayay et al. did not make it, their conclusions would not follow.

   Hence, the new approach to calculating the posterior probabilities which occur in (B′) can be

rejected for rather obvious reason, because of the unreasonable and *ad hoc* way in which the prior

distributions of the parameters are chosen in this approach. However, the more traditional approach

to calculating them is worthy of more attention than philosophers have until now given to it.

5. The Standard Bayesian Construction

   The argument which we shall now consider can be used for motivating the use of (Rule–BIC),

(Rule–AIC) and other similar rules when all the considered models are taken to have *the same prior*

probability. Whenever this is the case the rule (B″) becomes equivalent with the following simpler

rule:

(B‴)   Choose the model for which

$$(-2)[logarithm\ of\ prob(Y\ given\ X\ and\ \mathbf{M})]$$

         is smallest among the considered models **M**!

The value of *prob*(*Y given X and* **M**) depends on the prior probabilities that the parameters have within the model **M**. Unlike in the construction of Bandyopadhayay et al., in the argument we are currently considering this prior distribution has not been chosen in an *ad hoc* manner, with the aim giving some particular value to *prob*(*Y given X and* **M**) in sight. Rather, its choice is motivated by the use of a *quantitative measure of informativeness* of the prior distribution.

There is a natural measure for the amount of information that the *available observations* contain concerning the values of the parameters. If one assumes that the model **M** is true, so that the true curve is the curve which corresponds to some combination $\alpha_1{}^*, \alpha_2{}^*, ..., \alpha_k{}^*$ of the values of the parameters of this model, it becomes possible to ask how the estimates $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_k$ of these values would be distributed if one *repeatedly* measured the values of *y* for those *x* values for which measurements are currently available. As a matter of fact, it is fairly easy to explicitly calculate the probability distribution which the estimates $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_k$ would have in this case. This distribution is a *many-dimensional normal distribution* – it is *k*-dimensional when there are *k* parameters – and it is centred around the true values $\alpha_1{}^*, \alpha_2{}^*, ..., \alpha_k{}^*$ of the parameters (see e.g. Wetherill, 1986, 7-8).

Just like the familiar one-dimensional normal distribution specifies a distribution for some *single* quantity *z*, this *k*-dimensional normal distribution gives a probability distribution for the *combination* of the values of the parameter estimates $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_k$. However, the one-dimensional and the many-dimensional distributions differ in so far that the one-dimensional normal distribution can be represented graphically by the familiar bell-shaped curve, but the many-dimensional distribution does not have any analogous graphical representation.

The extent to which a one-dimensional normal distribution of a quantity *z* is *uninformative* can naturally be measured by its *variance*. To see why this is the case, one can imagine that a quantity

z has a true value *z**, and that is a method of measuring this value which is such that the probability distribution of the measurement result is a normal distribution which is centred around *z**. The bell-shaped curve which represents this normal distribution is the broader the larger is the variance of the distribution, which means that when the variance of this distribution becomes larger, the judgements that one can reasonably make on the basis of a measured value of *z* concerning the its true value (i.e. *z**) become less and less informative. This makes it natural to take the *inverse of the variance* of *z*, i.e. the quantity $1/(variance\ of\ z)$, to be a measure of *informativeness* of the one-dimensional normal distribution of the quantity *z*.

This measure of informativeness has an analogy also in the *k*-dimensional case. A *k*-dimensional probability distribution does not have any single number as its variance; rather, the *k*-dimensional analogy of the variance is the *covariance matrix*, which is an array of $(k \times k)$ numbers. The covariance matrix of the estimates $\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$ specifies *both* the variances that the different estimates $\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$ have when they are considered separately *and* their mutual *covariances* which measure the extent to which their values depend on each other. Similarly with the variance of a one-dimensional normal distribution, the numbers which appear in this matrix are the larger the smaller is the amount of information that the probability distribution gives about $\alpha_1^*, \alpha_2^*, ..., \alpha_k^*$, i.e. about the unknown true values of $\alpha_1, \alpha_2, ..., \alpha_k$. Analogously with the one-dimensional case, it is natural to take the inverse of this matrix to be a measure of the informativeness of the probability distribution concerning $\alpha_1^*, \alpha_2^*, ..., \alpha_k^*$. This measure of informativeness is sometimes called the *observed information matrix*:[6]

---

[6] See e.g. Kuha (*submitted*), section 3; *cf.* Wetherill (1986), formula (1.5) on p. 7.

(3)   [*Observed information matrix*] =

      inverse of [*covariance matrix of the parameter estimates* $\hat{\alpha}_1$, $\hat{\alpha}_2$,..., $\hat{\alpha}_k$.]

Here we shall not explain in a detailed manner what one precisely speaking means by the inverse of a matrix: for our purposes the only essential feature of such an inverse is its being analogous with the inverse of a number in so far that, the larger are the numbers in a matrix, the smaller are, as a rule, the numbers which appear in its inverse.

  The observed information matrix which is defined by (3) is approximately proportional to the number of the available observations: if the number of the available observations *e.g.* rises from 100 to 200, the numbers which appear in this matrix will be approximately doubled as well.[7] This makes it natural to think of the information that each single observation brings as approximately given by the matrix

      $(1/n)$[*Observed information matrix*],

when n is the number of the available observations. Here the operation of multiplying a matrix by $(1/n)$ should be understood in the obvious way, as yielding a matrix each of whose elements has been obtained from the corresponding element of the original matrix by multiplying it by $(1/n)$. Similarly, it is natural to think that the amount of information that n_0 observations would bring as given by

(4)   [*Information matrix for n_0 observations*] = (n_0/n)[*Observed information matrix*]

---

[7] More precisely, this is typically the case when the x values of the new measurements have the same order of magnitude with the x values that have been measured earlier, like when they e.g. are between some of the x values that have been measured previously.

Now, also the prior distribution that the true values of the parameters $\alpha_1, \alpha_2, ..., \alpha_k$ are supposed to have in a Bayesian approach is a *k*-dimensional probability distribution, and this is often assumed to be normal. Reasoning by analogy, one can measure the informativeness of also this distribution by the inverse of its covariance matrix. In the Bayesian construction that we are currently considering this inverse is taken to be the matrix which corresponds, in the sense that was explained above, to n_0 imagined observations for some n_0. In other words, one chooses the prior distribution in such a way that

(5)   *Inverse of* [*Covariance matrix of prior distribution*] = [*Information matrix for n_0 observations*]

When the prior distribution of the parameters $\alpha_1, \alpha_2, ..., \alpha_k$ of the model **M** have been fixed in this way, it becomes possible to calculate the value *prob*(*Y given X and* **M**). When the result of this calculation is substituted into the expression which occurs in (B′′′), the value of this expression turns out to be[8]

(8) (-2)[*logarithm of prob*(*Y given X and* **M**)] $\approx$

              (-2)[*logarithm of prob*(*Y given X,* **M**, *and* $\hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_k$)

              $+$ *log* (n/n_0)(*number of parameters of* **M**) $+$ *constant*

Comparison of this formula with the definition of BIC(**M**) shows that this construction leads to BIC when n_0 is chosen to be 1, i.e. when the prior distribution is thought of as containing as much

---

[8] This follows immediately from the formula (6) of Kuha (*submitted*, Section 3), and somewhat less directly from formula (7) in Smith–Spiegelhalter (1980, 215). *Cf.* also Raftery (1995, 131).

information concerning the value of the considered quantity as a single observation would contain. Similarly, the construction leads to AIC when n_0 is chosen in such a way that *log* (n/n_0) =2, which is equivalent with $n_0 = (1/e^2)n$.

 This means that the case in which this Bayesian construction leads to AIC and the one in which it leads to BIC differ in so far that when one arrives at AIC the prior distribution of the parameters *contains more information* concerning their values than when one arrives at BIC. In the former case the information in the prior distribution corresponds to a number n_0 of imagined observations which grows as the number of actual observations, n, grows. It is also clear that the same construction can be used for yielding any information criterion of the form

$$(-2)[\textit{logarithm of prob}(Y \textit{ given } X, \textbf{ M}, \textit{ and } \hat{\alpha}_1, \hat{\alpha}_2, ..., \hat{\alpha}_k)$$

$$+ f(n)(\textit{number of parameters of } n),$$

where $f(n)$ is an arbitrary function of n, by adjusting the information in the prior distribution of the parameters suitably.

6. Concluding Remarks

 We have seen that a Bayesian approach to model choice is more flexible than the Akaikean approach, which is motivated by the wish to maximize predictive accuracy. The latter approach leads to the particular information criterion AIC, but the Bayesian approach produces a whole family of different information criteria which correspond to different assumptions concerning the amount of available prior information concerning the values of the parameters. Such flexibility could be viewed as a positive aspect of the Bayesian approach, but it could also be used as a criticism of Bayesianism.

After all, such flexibility shows that Bayesianism fails to answer the question how one should choose between models when it is not clear how the prior probabilities of their parameters should be fixed.

Despite of this obvious criticism, the Bayesian construction which we discussed above is quite interesting philosophically. This is because it builds a connection between the important methodological concepts of *informativeness* and *simplicity*. In its typical applications the Bayesian method of justifying the use of various model choice criteria which we described above leads to the conclusion that, the more prior information concerning the values of the parameters the researchers are assumed to have, the less weight they should give to the simplicity of the model that they choose. In this short paper we have not been able to analyse the significance of this conclusion in a detailed manner. However, already at this stage we can state where, also more generally, the philosophically interesting aspects of Bayesian approaches to model selection are to be found. They are associated with ways in which the different ways of fixing *the prior distributions that the parameters have within each model* make researchers give different amounts of advantage to simpler models, and not with the more or less trivial observation that, if one fixes *the prior probabilities that the models themselves have* in such a way that the simpler models have larger prior probabilities, Bayesian methodological rules will instruct us to prefer simple models to more complicated ones.

REFERENCES

Bandyopadhayay, P. S., R. J. Boik,and P. Basu (1996), "The Curve-Fitting Problem: A Bayesian Approach", *Philosophy of Science* 63 *(Proceedings)*: S264-S272.

Burnham, K. P. and D. R. Andersson (1998), *Model Selection and Inference. A Practical Information-Theoretic Approach.* New York: Springer.

Forster, M. and E. Sober (1994), "How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions", *British Journal for the Philosophy of Science* 45: 1-35.

Forster, M. (*forthcoming*), "The New Science of Simplicity", forthcoming in H. Keuzenkamp, M. McAleer, and A. Zellner (*eds.*), *Simplicity, Inference and Econometric Modelling*, Cambridge University Press.

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

Kuha, J. (*submitted*), "Simplicity and Model Fit: Implications of a Bayesian Approach", submitted for publication in the *Philosophy of Science*.

Raftery, A. E. (1995), "Bayesian Model Selection in Social Research", in P. V. Marsden (*ed.*), *Sociological Methodology 1995*, Washington DC: Blackwell Publishers, pp. 111-163.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa, (1986), *Akaike Information Criterion Statistics*, Tokyo: KTK Scientific Publishers.

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

Smith, A. F. M. and D. J. Spiegelhalter (1980), "Bayes Factors and Choice Criteria for Linear Models", *Journal of the Royal Statistical Society, Series B* 42: 213-220.

Wetherill, G. B. (1986), *Regression Analysis with Applications*, London: Chapman and Hall Ltd.