1 Introduction

As Michael Schultz notes in his very interesting paper (Schultz, 2018), standard model selection criteria such as the Akaike Information criterion (AIC) (Akaike, 1974), the Bayesian Information Criterion (BIC) (Schwarz, 1978), and the Minimum Description Length principle (MDL) (Rissanen, 1978) are purely empirical criteria in the sense that the score a model receives does not depend on how well the model coheres with background theory. This is unsatisfying, since we would like our models to be not just empirically successful, but also theoretically plausible. To remedy this situation, Schultz proposes an amended version of AIC that he calls the “Inferential Information Criterion” (IIC). Mathematically, the IIC score of a model is a trade-off between the model’s AIC score and its conditional probability given a background theory, $P(M|T)$. More precisely, IIC issues the following imperative:

*The Inferential Information Criterion:* Choose the model that has the maximal IIC score, where the IIC score of $M$ is given by the following formula:

$$IIC(M) = - \log P(M|T) + AIC(M) \quad (1.1)$$
From a statistical point of view, the IIC score is a rather odd hybrid, since it involves both the Bayesian-looking $P(M|T)$ and the AIC score, $AIC(M)$, which is typically considered a frequentist construct. The goal of this comment is to analyze IIC from a fully Bayesian point of view. As we will see, IIC has a Bayesian justification. Unfortunately, the Bayesian analysis will also show that there is reason to suspect that IIC will fail as an adequate model selection criterion in precisely the cases that Schultz is most interested in.

## 2 The Bayesian justification of IIC

Let us first have a look at how Schultz derives IIC. Schultz’s derivation has as its starting point the assumption that the optimal model is the model, $M$, that maximizes the log-likelihood of the background theory $T$ given $M$ and data $X$, or, in mathematical notation, $\mathcal{L}(T|X, M)$. Next, Schultz assumes that the theory and data are independent conditional on each model, i.e. that $P(X, T|M) = P(X|M) * P(T|M)$. Given this assumption, Schultz is able to show that the model that maximizes $\mathcal{L}(T|X, M)$ will also maximize $\log P(M|T) + \mathcal{L}(M|X)$. Hence, the optimal model – according to Schultz – is the model that would be selected by what we may call the “Ideal Selection Criterion” (ISC):

**The Ideal Selection Criterion**: Choose the model that has maximal ISC score, where the ISC score of $M$ is given by the following formula:

$$ISC(M) = \log P(M|T) + \mathcal{L}(M|X)$$

($2.1$)

$ISC$ is the model selection criterion we would have preferred to use if we were able to do so. Unfortunately, $\mathcal{L}(M|X)$ is in general not a quantity that

---

$^1$In the appendix, at least, Schultz’s derivation appears to implicitly start from the
we can calculate, so \( ISC \) cannot be applied in practice. However, as long as certain statistical assumptions are satisfied \([\text{Akaike} 1974]\), \( AIC(M) \) is a decent approximation of the mean log-likelihood \( \bar{L}(M|X) \), and by exploiting this fact Schultz is able to derive \( IIC \) from \( ISC \).

So that, in brief, is how Schultz derives \( IIC \). The starting assumption of Schultz’s derivation is puzzling, however. Why should we think that the best model is the model that maximizes the likelihood of the background theory? This assumption does not make sense from either a maximum likelihood or Bayesian point of view. From a maximum likelihood point of view, the best model is the model that has the highest likelihood, given the data and background theory. From a Bayesian perspective, on the other hand, the best model is the model with the highest posterior probability. Thus, from either a Bayesian or frequentist perspective, Schultz’s starting assumption looks unmotivated.

Fortunately, as we will soon see, we can give a Bayesian justification of \( ISC \) that avoids Schultz’s starting assumption. Indeed, \( ISC \) is in effect equivalent to Bayesian model selection, given the assumptions made by Schultz. That is, the model that has the best \( ISC \) score will also – given Schultz’s assumptions – be the model that has the highest posterior probability given the data and background theory, \( P(M|X,T) \). To see why, note that given Schultz’s assumption that the theory and data are independent conditional on the model, we can easily derive the following relation:

\[
P(M|X,T) = \frac{P(M|T)P(X|M)}{P(X|T)}
\]

Next, taking logarithms and using the fact that 
\[
ISC(M) = \log P(M|T) + \mathcal{L}(M|X)
\]

we can write (2.2) in the following form:

\[
P(M|X,T) = \frac{P(M|T)P(X|M)}{P(X|T)}
\]

2 There is an important ambiguity in Schultz’s paper concerning how we are to interpret \( M \). In the main text, Schultz says that \( M \) is to be regarded as a conjunction of assumptions. However, in footnote 7, Schultz says that \( M \) should be regarded as a model that has been
\[
\log P(M|X,T) = ISC(M) - \log P(X|T)
\] (2.3)

Note that for the purposes of model evaluation and selection, \(\log P(X|T)\) is essentially a constant, since it does not depend on \(M\). Hence, from (2.3) it is clear that the model that has the best \(ISC\) score will also be the model that has the highest posterior probability, given both the data and the background theory.

Thus, from a Bayesian point of view, \(ISC\) is a perfectly reasonable model selection criterion since it is equivalent to simply selecting the model that has the highest posterior probability. Moreover, in cases in which the IIC score is a good approximation of the ISC score, it follows that IIC will also be a reasonable model selection criterion.

### 3 Problems with ISC and IIC

In addition to proposing \(IIC\) as an improvement over \(AIC\), Schultz also gives us a practical recipe for how to use \(IIC\) to compare models. Schultz’s recipe involves listing the models in such a way that they form a sequence of nested (and identifiable) models, where each model in the sequence is regarded as a conjunction of simplifying assumptions. These assumptions may, for example, be assumptions about whether two effects interact or whether the relationship between a set of variables has a certain functional form.

However, note that if \(ISC\) is used to compare models that are nested in the way suggested by Schultz, then it will always prefer the first model in the sequence, regardless of whatever the data may say. To see why that is, suppose we have two models in the sequence, \(M_i\) and \(M_{i-1}\), such that \(M_i\) is the result of adding one extra simplifying assumption to \(M_{i-1}\). That is, \(M_i = M_{i-1} \land m_i\). By using equation (2.3), we get:

"fit with MLE parameters. Although I do not believe this ambiguity is harmless, I will ignore it in the rest of my comment."

4
\[ ISC(M_i) - ISC(M_{i-1}) = \log \frac{P(M_{i-1} \land m_i \mid X, T)}{P(M_{i-1} \mid X, T)} \] (3.1)

Now, it is a basic fact about probabilities that the probability of a conjunction cannot be greater than the probability of any of its conjuncts, and from this fact it follows that \( \log \frac{P(M_{i-1} \land m_i \mid X, T)}{P(M_{i-1} \mid X, T)} \leq 0 \). Consequently, we see from (3.1) that \( ISC(M_{i-1}) \geq ISC(M_i) \). In other words, the ISC score of \( M_{i-1} \) is necessarily at least as large as the ISC score of \( M_i \), for every \( i \). It follows that \( M_1 \) – i.e. the model that has the smallest possible number of simplifying assumptions while still being identifiable – will always be favored by ISC.

Hence, ISC is a trivial model selection criterion if it is applied to models that are nested in the way suggested by Schultz. This is not surprising since, as we saw earlier, ISC is equivalent to selecting the model that has the highest posterior probability, and – as [Popper (2002)] initially pointed out and [Forster and Sober (1994) and Gelman and Rubin (1995)] both stress – Bayesians have trouble making sense of model selection when the relevant models are nested.

Interestingly, as Schultz shows in his paper, IIC does not necessarily prefer the most complex model; thus, ISC and IIC sometimes disagree. What is the explanation for this disagreement? Recall that Schultz derives IIC from ISC by using AIC\((M)\) to approximate the mean model log-likelihood, \( \bar{L}(M \mid X) \). Hence, IIC is a reasonable model selection criterion in exactly the cases in which AIC\((M)\) is a good approximation of \( \bar{L}(M \mid X) \). Since ISC and IIC disagree in the examples Schultz considers, that indicates that AIC\((M)\) is not a good approximation of \( \bar{L}(M \mid X) \) in Schultz’s examples. But given that Schultz derives IIC from ISC, we should presumably be suspicious of IIC whenever it disagrees with ISC.

---

3A standard response to this problem ([Henderson et al. (2010); Gelman and Shalizi (2013)]) is to argue that selecting a model is not really a proper Bayesian goal.
4 Conclusion

Insofar as we regard $IIC$ as an approximation of $ISC$ – which Schultz seems to do – we have good reason to suspect that IIC will fail as an adequate model selection criterion in precisely the cases that Schultz is most interested in, since $IIC$ and $ISC$ disagree with each other in those cases.

It is, however, possible that $IIC$ could be given a justification that does not proceed through $ISC$. Analogously, although the Bayesian Information Criterion ($BIC$) has “Bayesian” in its name, its more convincing theoretical justifications have little do with standard Bayesianism, as Sprenger (2013) points out. For example, Raftery (1995) argues that the $BIC$ score of a model should be regarded as an approximation of the integrated likelihood of the model, and that the integrated likelihood, in turn, is an estimate of the model’s out-of-sample predictive accuracy. Thus, although the initial motivation behind $BIC$ was Bayesian, $BIC$ can be given an independent – and arguably more plausible – theoretical justification that is grounded in the goal of predictive accuracy.

Similarly, although Schultz derives $IIC$ from $ISC$, it may be that $IIC$ can be given an independent theoretical justification. Until such an alternative justification is provided, however, the theoretical foundations of $IIC$ remain shaky.

References


