# Discussion of the Papers by Rissanen, and by Wallace and Dowe

B. CLARKE[1]

*Department of Statistics, UBC, 6356 Agricultural Road, Room 333, Vancouver, BC, Canada, U6T 1Z2*
*Email: bertrand@stat.ubc.ca*

**Commenting on the papers by Dr Rissanen and Professors Wallace and Dowe is a daunting sort of pleasure. Superficially, the papers are not closely related: Dr Rissanen has focused on the Shtarkov optimality criterion in a minimum description length (MDL) context giving some new implications from normalization (see his equation (1)) for model selection and hypothesis testing. By contrast, Wallace and Dowe have given a first, and welcome, effort to axiomatize a setting in which it is reasonable to hope that the algorithmic complexity approaches of Kolmogorov and Solomonoff—streams one and two, using UTMs—may coincide with the Shannon theory approaches—stream three in its minimum message length (MML) and MDL versions. Despite these differences, there are several senses in which these two papers are closely related.**

## 1. MODEL SELECTION TECHNIQUES

First, note that a common feature of all the techniques mentioned in both papers is that they are intended to identify general features of the data-generating mechanism, ideally providing us with a model. Second, if the unification envisaged by Wallace and Dowe can be formally obtained, any implications derived from one technique, such as Rissanen's latest implications, will obtain for the other techniques that are equivalent to it.

Third, the techniques elaborated by Rissanen and unified into a class by Wallace and Dowe are themselves examples within a broader class of model selection principles (MSPs) that have been studied by statisticians, engineers and computer scientists (amongst others) for decades. Recent work due to Yang and Barron [1], and earlier work due to Barron and Cover [2], Bethel and Shumway [3], are efforts to provide general results for collections of classes that are recognized to have common properties—typically the dependence of the penalty term on $n$, the sample size.

In particular, one may consider an Akaike information criterion or AIC-class, see [4], of MSPs which contains the AIC and its equivalent formulations such as Mallow's $C_p$ and cross-validation, see [5]. Equivalently, the AIC class of MSPs can be defined as the MSPs that satisfy the same optimality criterion from prediction as the AIC itself does, see [6, 7].

Second, one may consider a Bayes information criterion, or BIC-class of MSPs which contains the BIC and other equivalent formulations such as the posterior quantities it approximates, see [8]. Perhaps the optimality criterion satisfied by the BIC, see [9], or the posterior probabilities, can be used to define this class. In some cases, the MDL is similar to the BIC, see [2] for the interpretation of the MDL in terms of penalized likelihood. However, this overlap is not representative of the relationship between MDL and BIC because the MDL has such different properties outside the finite-dimensional parameter context.

Accepting the main point of Wallace and Dowe leads one to suggest that Kolmogorov's approach, Solomonoff's approach, the MDL and the MML may form a third class. The defining feature of this third class appears to be its interpretation in terms of two-stage coding. This is clear for the MML and some versions of MDL, such as Barron and Cover [2]. It is explicitly in the axiomatization offered by Wallace and Dowe and implicitly in their discussion about converting from one UTM to another.

Now we can consider three classes of MSPs: the AIC and equivalent class; the BIC and equivalent class, and the two-stage coding class. The AIC class has a penalty term given by the number of parameters. The BIC class has a penalty term given by $\log n$ times the number of parameters. The two-stage coding class has a penalty dependent term dependent on the class over which one is optimizing. There are, of course, other classes of MSPs since many functions of $n$ can define a penalty term, but these three classes currently have the virtue of interpretability and extensive study. Penalty terms that are powers of $n$ have also been studied, see [1, 3], so we have some idea how such models will perform. However, in practice, using an MSP outside a well understood class would require particularly good arguments.

## 2. THE BAYESIAN CONNECTION

The perceptiveness of Rissanen's approach masks its two-stage coding interpretation apart from the normalizing constant for the Jeffreys prior that appears in (9). Thinking of the prior as representing the code lengths for the first stage of a two-stage code, we see that Rissanen's approach is not incompatible with a Bayesian approach. Indeed, Barron and Xie [10] use a modification of the Jeffreys prior to obtain asymptotics for the same optimality criterion

[1] Visiting the Department of Statistical Science, University College London, UK.

of Shtarkov used by Rissanen. The Barron and Xie [10] strategies are only for the discrete memoryless case but their results are based on mixtures of likelihoods with respect to modifications of the Jeffreys prior. However, in some continuous memoryless cases one can get upper and lower bounds on (2): an upper bound follows from choosing $q$ in (2) to be the mixture of the parametric family with respect to the Jeffreys improper prior. (For the $\mathit{Normal}(\theta, 1)$, simple algebra suffices; for the $\mathit{Exponential}(\theta)$, use Stirling's formula.) For lower bounds, recall maximin is less than or equal to minimax. Thus one gets an integral over the sample space and again $q$ is chosen as the mixture. Fubini's theorem and multiplying and dividing by the parametric family in the logarithm gives two integrals whose asymptotics follow, in general, from Clarke and Barron [11, 12]. The upper and lower bounds are asymptotically identical and the result is the same as (9). The technique for the upper bound generalizes by using a Laplace integration argument, but the conditions seem restrictive. The technique for obtaining a lower bound stems from an examination of the optimality criterion based on redundancy that Rissanen notes can also lead to (9). All of these criteria directly or indirectly permit a two-stage coding interpretation.

Rissanen's approach—like many MSPs—has implications for hypothesis testing and it may be that the MDL can lead to techniques that perform better in some sense than the posterior probabilities many Bayesians use. However, posterior probabilities are used only because they are the Bayes optimal strategy under generalized 0–1 loss. If 0–1 loss is not the appropriate loss function, a Bayesian might not be led to use posterior probabilities. Indeed, if one had a physical motivation for using a two-stage coding optimality criterion, then a different loss function reflecting that reality—say the relative entropy—might be more appropriate, thereby leading to a different decision theory problem.

Moreover, one cannot in general reduce the testing of a point null (which is already problematic because of the lack of a common dominating measure) against a composite alternative to testing the null against the point alternative given by mixing the densities in the alternative with a fixed prior as Rissanen does in Section 3. In addition, power is a frequentist concept because it is an integral over the sample space, which a doctrinaire Bayesian would not use *a posteriori*. Of greater importance to Bayesians is the robustness of a hypothesis test to choice of prior or the other inputs to its formulation. Many Bayesian would use robustness arguments to address the subjectivity of the prior, or choose the prior to satisfy a physically meaningful optimality criterion. Finally, the testing in Section 3 is unusual because the regions are annuli with radii shrinking at the 'regular' $O(1/\sqrt{n})$ rate. It is unclear whether the conclusions follow from the choice of rate or from the technique. In short, it is difficult to interpret the results of this section.

## 3. LOOKING AHEAD

Finally, in the present context of model selection there are three important tasks which have yet to be addressed. One task is to select finitely many parametric families from the class of models one is willing to consider. One can imagine trying to do this by treating the models as messages to be coded and then using data compression techniques possibly stemming from the rate distortion function, analogous to vector quantization. However, this will be complicated by the requirement that one include—and choose amongst—potential explanatory variables. Second, given that one has such a collection of families one must choose a model selection technique intelligently and know how to use it: use the model chosen directly or use some version of model averaging—say only average about neighbourhoods of the best models? If two models reflecting incompatible physical assumptions are nearly equally good, how does one proceed? Should one average over the models chosen by different MSPs? Third, one must evaluate how good a model is—this may lead to the prequential approach of Dawid, see Skouras and Dawid [13]. Possibly, the examination of model uncertainty or mis-specification will necessitate topologizing the collection of models under consideration—in which case the present information-theoretic setting leads naturally to use of Shannon's mutual information.

## REFERENCES

[1] Yang, Y. and Barron, A. R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory*, **44**, 95–116.

[2] Barron, A. R. and Cover, T. (1990) Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **37**, 1034–1054.

[3] Bethel, J. and Shumway, R. (1988) *Asymptotic Properties of Information—Theoretic Methods of Model Selection*. U. C. Davis Div. of Stats. Tech. Rep. #112.

[4] Akaike, H. (1977) On entropy maximization principles. In Krishnaiah, P. K. (ed.), *Applications of Statistics*, pp. 27–41. North Holland, Amsterdam.

[5] Li, K.-C. (1987) Asymptotic optimality for $C_p$, $C_l$, cross validation and generalized cross validation: discrete index set. *Ann. Statist.*, **15**, 958–975.

[6] Shibata, R. (1981) An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

[7] Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.*, B, **41**, 190–195.

[8] Haughton, D. (1988) On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342–355.

[9] Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[10] Barron, A. R. and Xie, Q. (1996) Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Trans. Inform. Theory*. To appear.

[11] Clarke, B. and Barron, A. R. (1994) Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stat. Planning Inference*, **41**, 37–60.

[12] Clarke, B. and Barron, A. R. (1988) *Information Theoretic Asymptotics of Bayes Methods*. Department of Statistics Technical Report # 26, University of Illinois.

[13] Skouras, K. and Dawid, A. P. (1998) On efficient point prediction systems. *J. R. Statist.*, B, **60**, 765–780.