Rejoinder

C. S. WALLACE AND D. L. DOWE

Computer Science, Monash University, Clayton, Victoria 3168, Australia
Email: csw@cs.monash.edu.au

We would like to thank all the discussants for their contributions. Our paper attempts only a first step towards a satisfactory linkage between complexity theory and coding-based inference and these discussions will help to blaze the paths for future exploration. We also take this opportunity to remind the reader that, in our opinion, MDL and MML agree on many, many points. While we certainly disagree with Dr Rissanen and he with us on quite a few points, we certainly acknowledge that these disagreements would understandably appear both infrequent and minor from the perspective of someone who knew relatively little about MDL and MML.

1. RESPONSE TO PROFESSOR DAWID

It is perhaps appropriate to re-emphasize the distinctions among inductive inference, prediction and prediction with a loss function. Straight prediction yields a probability distribution over future data. Add a loss function and you get decision theory, giving the loss to be expected from betting on a particular future event. While models of the data-generating process may appear in this reasoning, there is no need to commit to any one model, and in general the predictive distribution is not a member of the family of models considered. Inductive inference (in the classical sense) aims to yield a general proposition about (model of) what is going on in the data-generating process, without explicit consideration of what future data may be generated. The complexity approach can be used for straight prediction, as in Solomonoff's work, and in no way restricts the loss function which may then be applied to determine how best to anticipate the future. In the inductive model-selection case, the complexity approach yields probabilities for various potential models, finite posterior probabilities in the case of minimum message length (MML), or something akin to a likelihood in the case of minimum description length (MDL). Again, there is no restriction on the application of a loss function to these results in order to choose the model which has least expected regret. Professor Dawid may have been misled by the common, but by no means universal, habit of complexity workers of emphasizing the discovery of the model or model class which is 'best' in message-length terms, but in fact the techniques yield comparisons of alternative models expressible as probability or likelihood ratios and any regret function may be applied to these ratios. Thus, the 'biggest weakness' he sees is an illusion.

In Professor Dawid's discussion, it is not clear whether the encoding of ξ is using a one-part or a two-part Bayesian message, but the asymptotic results would still appear to hold for MML even in a pathological case such as our awkward uniform example from Section 1.2 in our discussion paper in this issue.

Professor Dawid's conclusion that 'Bayes is a good thing' is, to us, a welcome and not overly surprising one. Indeed, information-theoretic coding approaches such as MDL, MML and Kolmogorov complexity admit of a Bayesian interpretation (cf. Professor Clarke's remark that 'Rissanen's approach is not inconsistent with a Bayesian approach').

Professor Dawid allocates some of his discussion to the choice of Bayesian priors, such as is implicit in the choice of his $\alpha(\cdot)$ and to the choice of loss function. Where there is prior knowledge, as in incorporating expert advice, we advocate its use. As we point out in Section 2.3 on the Jeffreys prior in our discussion paper in this issue, priors chosen on the basis of the mathematical form of the probabilistic relation between the unknown parameter and the possible outcomes of some observational protocol are not logically tenable, and, as Professor Dawid remarks, lead to an unwelcome dependence on such things as stopping rules.

Professor Dawid makes the interesting observation that such dependence on stopping rules disappears in the 'online' setting. However, this fact seems insufficient justification for accepting an untenable 'prior' such as Jeffreys.

The 'on-line' setting has a natural, indeed inevitable, role in the prediction of time-series data. However, it is not clear that it is a satisfactory setting for inductive inference when the data have no inherent sequence, since the results of an 'on-line' analysis depend on the order in which the data are considered.

Professor Dawid is perhaps unduly pessimistic about the future of coding-based approaches. The successes of MML/MDL inference over alternative methods are manifold as is attested to by many of the references in the contributed papers in this issue. It may be true that many of the successes could (with hindsight) have been obtained by conventional statistics. As the coding approach is not inconsistent with probability theory, this is not surprising. However, we may plausibly claim that many decades of conventional statistical theory did not produce any inductive inference method with the generality successfully to address all the problems to which the coding approach has been applied, despite the long standing of some of them. A conventional solution of these problems may well be possible, but might prove to be no more than the re-interpretation in older language of the insights developed in the coding methods.

2. RESPONSE TO DR RISSANEN

Dr Rissanen appears to have mistaken some technical aspects of MML. Some of his criticisms would appear to lack substance. MML wishes to obtain the 'best' two-part compression under the assumption that the second part of the message, encoding D given H, encodes D given the hypothesis H and nothing else, without paying attention to alternative hypotheses that might have been used but are not. Such a code is of course redundant if considered purely as an encoding of the data. It must be, because it also encodes something which is not deducible from the data, namely estimates of the unknown parameters. The reason that Dr Rissanen's MDL 'beats' this with the equation (6) of his discussion is that his optimization does not do such a two-part encoding and is, to us, misguided—see Section 1 on complete coding in our discussion paper, and see also Section 1.1 in our discussion paper to see that the amount by which MDL 'beats' MML is, in any event, rather small. The 'complete coding' now advocated in MDL approximately removes this redundancy, and with it removes any well-founded estimation of parameters. This is fine if we have no interest in these parameters, and wish only to infer the parametric model class, but gives no grounds for criticizing the MML form which does aim to yield a fully-specified model. MDL seems in recent years to have focused on the selection of a model class. We have given reason in Section 2.1 on partitioning models into 'model classes' in our discussion paper to suggest that the notion of a model class is not always well-defined. The MML school has advocated this principle of two-part coding since its inception in 1968 and the MDL school has differed on this point for some years now and may well continue to do so. It is a difference of objective, not a contradiction.

The remark about Kolmogorov's 'sufficient statistic' is not strictly applicable to the problem of inference from a given, finite data set, since the definition involves the behaviour of a quantity as the length of the data string increases indefinitely. In fact, the definition seems to make sense as a definition of a 'function' from all finite data strings generated by some source to the first-part strings which would appear in the two-part encodings of these data strings. If it is so interpreted, we may ask whether (for data coming from some unknown member θ of a known family Θ of sources) the MML estimator function satisfies the definition; for simple families such as the exponential family, it does. For the family of all sources with computable-probability distributions, the results of Barron and Cover [1] strongly suggest an affirmative answer also, although then the MML estimator 'function' is not computable.

Dr Rissanen's remark about the posterior, $P(y \mid x)$, should be contrasted with our equation for $M_T(S) - K_T(S)$ as the log of a posterior probability in Section 4.2 of our discussion paper in this issue. MML does two-part coding.

We contest the assertion that MDL is better-founded than MML. MML is able to select a model class, or a fully-specified hypothesis or whatever—depending on the problem specification and MML can do any of these with both statistical consistency and invariance under one-to-one re-parametrization. This can be done without any of Dr Rissanen's parameter-space restrictions which Professor Dawid finds 'unsatisfying'. Dr Rissanen appears to misunderstand the Bayesian posterior maximization (or maximum *a posteriori*, MAP) principle, which, unlike MML, is typically concerned with probability densities.

Dr Rissanen's way of 'suitably' choosing a range on his parameters after equation (4) of his discussion paper seems either arbitrary and unclear or subjectively Bayesian. Although Dr Rissanen's normalized maximum likelihood (NML) approach will give statistical invariance, as in our discussion paper, it seems seriously flawed even for some relatively simple statistical distributions. Professor Dawid suggests in his discussion some possible ways in which NML might be salvaged.

The $o(\log n)$ term in Dr Rissanen's discussion equation (8) admits of an $O(1)$ term of order one, within which, as in our contributed paper, there is more than ample room for the inclusion or even the concealment of many kinds of Bayesian prior. Indeed, as we note in our response to Professor Dawid and also in Sections 5 and 7 and the conclusion of our contributed paper, Kolmogorov complexity (as appealed to by Dr Rissanen in Section 2 on Model Selection of his contributed paper) has a Bayesian interpretation—it not only permits the use of Bayesian priors, it would also appear to insist that we use Bayesian priors. We agree with Dr Rissanen that his discussion equation (8), like our message length equation in Section 6.1.2 of our contributed paper, is an approximation which is only meant to be good for the exponential family and certain other families of functions.

We are at a loss to understand the claim that MML obtains results generally inferior to those of MDL and we also contest the remark that MDL 'works in a much wider set of problems than the MML principle'. While the similarity of MDL and MML ensures that the methods will necessarily give similar answers on a variety of problems, we note that MML has statistical invariance in addition to statistical consistency. Although MML is designed primarily for inference, its similarity to the minimum expected Kullback–Leibler distance estimator ensures that it will be good for prediction. We note that two examples, where MML demonstrably out-performs MDL and indeed where it is not at all clear that MDL works at all, are the awkward uniform example from Section 1.2 of our discussion paper and (when compared to normalized maximum likelihood) the negative binomial example from Section 2.3 of our discussion paper. With regard to Dr Rissanen's final claim that 'for non-parametric model classes the MML principle' supposedly 'produces inferior results' (to MDL) 'or fails completely', in the absence of an example exhibiting this 'failure', we find the claim unclear and unsubstantiated.

3. RESPONSE TO PROFESSOR SHEN

We assume that in Section 4 of Professor Shen's discussion, the earlier restriction that $|A|$ be finite is removed with the

introduction of a probability distribution over A. His first definition in that section of the complexity of the distribution P seems in accord with the view taken in our paper. The restriction in his second definition to finite domains would rule out some natural distributions, such as the negative binomial.

While appreciating the clear exposition of the relation between two-part complexity and inference given by Professor Shen, we remain cautious of the practical application of inequalities with unquantified $O(\log n)$ terms.

Section 5 raises important questions. Computational resources are not infinite and no unrestricted Turing machine exists. Just what impact resource limits have on the results of algorithmic complexity theory remains to be properly explored.

4. RESPONSE TO PROFESSOR CLARKE

With regard to the model selection principles (MSPs) considered, we note that the Akaike information criterion (AIC) is statistically inconsistent for mixture modelling, as well as for a variety of other problems discussed in the conclusion section from our contributed paper.

We certainly agree with Professor Clarke's remark that 'Rissanen's approach is not inconsistent with a Bayesian approach'.

In Professor Clarke's Section 3 on Looking Ahead, the choice between potential explanatory variables is not a problem for MML—although the search space could, of course, be rather large. We simply choose those which ultimately give rise to the shortest message length, the message of course including a specification of which of the available explanatory variables are used.

Where there are several alternative hypotheses under consideration, for inference, MML will take the best one. While MML can obtain posterior-probability ratios among competing hypotheses, the practical use of a weighted average of competing hypotheses about, say, molecular bonds, would become so cumbersome that it would rarely be fruitful except for the simplest deductions from the hypotheses. For prediction, MML adherents and other Bayesians will advocate model averaging. If there are two very different models with almost identical message lengths, then it will certainly be well worth averaging for prediction. (Indeed, methods other than MML would also be well-advised to consider averaging—especially in such cases.)

REFERENCES

[1] Barron, A. R. and Cover, T. (1991) Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, **37**, 1034–1054.