



Comment

NML, Bayes and true distributions: A comment on Karabatsos and Walker (2006)

Peter Grünwald^{a,*}, Daniel J. Navarro^b^a *Centrum voor Wiskunde en Informatica, P.O. Box 94079 NL-1090 GB, The Netherlands*^b *School of Psychology, University of Adelaide, Adelaide SA 5005, Australia*

ARTICLE INFO

Article history:

Received 23 November 2007

Received in revised form

14 October 2008

Available online 23 February 2009

Keywords:

Minimum description length

Normalized maximum likelihood

Bayesian inference

ABSTRACT

We review the normalized maximum likelihood (NML) criterion for selecting among competing models. NML is generally justified on information-theoretic grounds, via the principle of minimum description length (MDL), in a derivation that “does not assume the existence of a true, data-generating distribution”. Since this “agnostic” claim has been a source of some recent confusion in the psychological literature, we explain in detail what is meant by this statement. In doing so we discuss the work presented by [Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology*, 50, 517–520], who propose an alternative Bayesian decision-theoretic characterization of NML, which leads them to conclude that the claim of agnosticity is meaningless. In the KW derivation, one part of the NML criterion (the likelihood term) arises from placing a Dirichlet process prior over possible data-generating distributions, and the other part (the complexity term) is folded into a loss function. Whereas in the original derivations of NML, the complexity term arises naturally, in the KW derivation its mathematical form is taken for granted and not explained any further. We argue that for this reason, the KW characterization is incomplete; relatedly, we question the relevance of the characterization and we argue that their main conclusion about agnosticity does not follow.

© 2009 Published by Elsevier Inc.

1. Introduction

The normalized maximum likelihood (NML) criterion for the selection among a collection of models $\mathcal{M}_1, \dots, \mathcal{M}_D$ in light of observed data $\mathbf{x} = (x_1 \dots x_n)$ states that, where possible, we should prefer the model \mathcal{M} that maximizes the following probability,

$$p^*(\mathbf{x}|\mathcal{M}) = \frac{f(\mathbf{x}|\hat{\theta}(\mathbf{x}, \mathcal{M}))}{\int_{\mathcal{X}^n} f(\mathbf{y}|\hat{\theta}(\mathbf{y}, \mathcal{M}))d\mathbf{y}} \quad (1)$$

where $f(\mathbf{x}|\theta, \mathcal{M})$ denotes the probability of the data according to model \mathcal{M} with parameter values θ . In this expression, \mathcal{X}^n denotes the sample space of possible data sets of size n , and $\hat{\theta}(\mathbf{y}, \mathcal{M})$ is the maximum likelihood estimate obtained when model \mathcal{M} is fit to data \mathbf{y} .

The NML probability can be derived as the solution to a number of different optimality problems (Rissanen, 2001; Shtarkov, 1987). It plays a prominent role in the minimum description length (MDL) approach to statistical inference, originating from information theory. However, the NML distribution has also been given an interpretation from other statistical perspectives. Apart from the

information-theoretic derivation, there are three other standard derivations of the NML probability (see Grünwald (2007)): the *prequential* interpretation (briefly discussed in the Appendix), a *differential-geometric interpretation* (in which the denominator in (1) is interpreted as a volume (see, e.g. Balasubramanian (2005))) and a *Bayesian interpretation* (which links (1) to Bayes factor model selection based on a Jeffreys' prior). Importantly, the information-theoretic and prequential derivations of NML do not rely on the assumption of a “true”, data-generating distribution. In this sense, NML is an “agnostic” method, which suggests that it behaves robustly in situations in which all models under consideration are wrong, yet some are useful.

In a recent paper, Karabatsos and Walker (2006) (KW from now on) propose an alternative Bayesian decision theoretic interpretation for the NML criterion, from which they argue that it is meaningless to make claims about NML being an agnostic method. However, there are a number of difficulties with their proposal, which we discuss in this paper. The plan of this paper is as follows: we begin by providing a brief discussion of the information-theoretic view of NML (Section 2). Following this, in Section 3, we explain in detail the meaning and implication of the “agnostic” property of NML. We then turn to the KW characterization itself (Section 4), and our concerns with it (Sections 5 and 6). We make some concluding remarks in Section 7. For the benefit of readers who are not familiar with information theory, the paper ends with an Appendix in which one

* Corresponding author.

E-mail address: Peter.Grunwald@cwi.nl (P. Grünwald).

of the alternative interpretations of NML – the prequential one – is explained in some detail.

2. The information-theoretic view on NML

The MDL principle states that we should prefer those models that allow us to compress the data set \mathbf{x} to the greatest possible extent. That is, if the codelength $L_C(\mathbf{x})$ denotes the number of bits required to describe \mathbf{x} using some code C , then we should prefer those models that allows us to produce short codelengths. We are able to talk about data compression using probabilistic language thanks to the Kraft inequality, which tells us that for any probability mass function f defined on a sample space \mathcal{X}^n , there exists a uniquely decodable code C such that, for all $\mathbf{y} \in \mathcal{X}^n$, the codelength is given by $L_C(\mathbf{y}) = -\log f(\mathbf{y})$. Conversely, for any uniquely decodable code C , there exists a mass function f that satisfies this equality. This establishes a 1-to-1 correspondence between probability mass functions and uniquely decodable codes. Essentially the same correspondence holds, after appropriate discretization, if f is a density rather than a mass function.

The most well-known derivation of the NML distribution from the MDL perspective is Rissanen's (2001) work, which slightly extends an earlier derivation by Shtarkov (1987). Given a model \mathcal{M} that is parametrized by $\theta \in \Theta$, Shtarkov demonstrates that the NML probability $p^*(\mathbf{x}|\mathcal{M})$ in Eq. (1) corresponds to the “best” possible coding that can be achieved using \mathcal{M} . Shtarkov defines the best coding scheme that a model can achieve in a minimax sense, as the one that satisfies the following equality:

$$p^* = \arg_p \min_p \max_{\mathbf{y}} \left[(-\log p(\mathbf{y})) - \left(-\log f(\mathbf{y}|\hat{\theta}(\mathbf{y}, \mathcal{M})) \right) \right], \quad (2)$$

where the minimum is over all distributions p that can be defined on \mathcal{X}^n , and the maximum is over all possible datasets $\mathbf{y} \in \mathcal{X}^n$. The expression in square brackets is called the *regret*: when applied to the actually-observed data \mathbf{x} , it is the additional number of bits one needs to code the data \mathbf{x} using (the code based on) p , compared to the code in \mathcal{M} that, with hindsight, turns out to minimize the codelength (maximize the probability) of \mathbf{x} . The latter code is invariably the code based on the ML (maximum likelihood) estimator $f(\cdot|\hat{\theta}(\mathbf{x}, \mathcal{M}))$. Thus, we seek, among all distributions (codes) p on \mathcal{X}^n , the one such that the worst-case regret is minimized. Regarding the more general question of why it makes sense to solve a minimax problem of this kind, the Appendix contains a brief discussion; but the interested reader is referred to Grünwald (2007) for an extensive discussion. For the current purposes, it suffices to note that a key point in the specification of this minimax problem is that it does not matter what probability distribution generated the data \mathbf{x} , or whether such a “true” distribution even exists: the NML distribution satisfies certain optimality criteria that depend only on the data. We elaborate this point in detail in the following section. Then, in Sections 4–6, we discuss the KW derivation and our criticisms of it.

3. The role of true distributions

It is useful to think of hypothesis testing and model selection methods as algorithms. These algorithms usually take, as input, a finite or countably infinite list $\mathcal{M}_1, \mathcal{M}_2, \dots$ of models (families of probability distributions), as well as data $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$. They output a particular model \mathcal{M} from the list, or, more generally, they assign a weight or probability to each model on the list. We now look at the role of “true” distributions, first (Section 3.1) in the *design* of such algorithms, and then (Section 3.2) in the *analysis* of such algorithms. For the specific case of MDL, algorithms such as (but not restricted to) NML, Grünwald (2007, ch. 16 and 17) discusses these issues in far more detail.

3.1. True distributions in the design of algorithms

For some methods, such as traditional Neyman–Pearson hypothesis testing and AIC model selection, the corresponding algorithms have explicitly been designed to achieve a certain specified performance *under the assumption that one of the distributions p in one of the models under consideration is exactly true (i.e. the data are sampled from) p* . Other methods, such as cross-validation and NML-based model selection, do not rely on such an assumption in order to construct the algorithm. For instance, Shtarkov's derivation of NML as the solution to the minimax problem in Eq. (2) treats the observed data \mathbf{x} as fixed, without invoking any assumptions about what mechanism produced those data in the first place.

As an example of a procedure for which the design explicitly relies on some assumptions about the true generating mechanism, consider the following simple problem. Suppose we want to choose between a model $\mathcal{M}_1 = \{f(\cdot|\mu) \mid \mu \in \mathbb{R}\}$ and its submodel $\mathcal{M}_0 = \{f(\cdot|\mu) \mid \mu = 0\}$, where, for $\mathbf{x} \in \mathcal{X}^n$, $f(\mathbf{x}|\mu)$ is the standard normal density, extended to n outcomes by independence. In the Neyman–Pearson approach to this problem, we perform a hypothesis test with $\mu = 0$ as the null hypothesis, and $\mu \neq 0$ as the alternative. Viewed as an algorithm, such a test takes data $\mathbf{x} \in \mathcal{X}^n$ as input, and it outputs “reject \mathcal{M}_0 ,” or “accept \mathcal{M}_0 ,” possibly together with a p -value. For simplicity, we assume the significance level is fixed at 0.01. This means that the test (algorithm) has been designed such that the type-I error is at most 0.01: if the data are sampled from \mathcal{M}_0 , the probability of output “reject” is at most 0.01; moreover, among all algorithms with this property, we use the one for which the type-II error is minimized. Now, notice that the type-I error is defined in terms of the probability of obtaining a particular kind of data set if model \mathcal{M}_0 is true. Similarly, the type-II error describes the probability of obtaining a different kind of data set if (some element of) model \mathcal{M}_1 is true. The design of the algorithm thus crucially depends on the data being sampled either from \mathcal{M}_0 or \mathcal{M}_1 . As a consequence, an awkward problem arises if the data are not sampled from either of the two models. Under such circumstances, both the accept/reject decision and the corresponding p -value have no clear interpretation anymore, as they are probabilities of events according to some distributions that we already know are not the data-generating distributions. This situation is by no means uncommon: in practice, we often know in advance that all models under consideration are, to some extent, wrong. Instead of trying to identify the true model, in such a situation we may want to choose the model that, hopefully, is the “best” in the sense that it leads to the best predictions about future data coming from the same source. The Neyman–Pearson test has not been designed for such a situation and, as we have just seen, its outputs cannot easily be interpreted anymore. In particular, even though we put our significance level at 0.01, we certainly cannot claim anymore that, by following the procedure repeatedly in a variety of contexts, only once in about a 100 times will we encounter the situation that we reject \mathcal{M}_0 even though it leads to better predictions than \mathcal{M}_1 .

The example suggests that if none of our models are perfect – as is usually the case – then we should use statistical algorithms whose output is a function *only* of how well the actually observed sequence of data can be *predicted* based on the given models. To make this precise, we need to define what it means to “predict based on a given model”. This can be done in various ways. Let us consider two examples: leave-one-out cross-validation (LOOCV (see Browne (2000))), an approach to model selection that is popular in the machine learning community; and NML. In LOOCV, for all outcomes x_i , one predicts x_i on the basis of the maximum likelihood (ML) estimator $\hat{\theta}(\mathbf{x} \setminus x_i)$, i.e. based on all observed data except x_i itself. The quality of predicting x_i with density or mass function f_{θ} is measured in terms of the log

loss, defined as $\text{Loss}(x_i, f) := -\log f(x_i)$: the smaller the loss, the better the prediction. According to LOOCV, we should select the model \mathcal{M}_j which minimizes the sum of all prediction errors, $\sum_{i=1}^n \text{Loss}(x_i, f(\cdot | \hat{\theta}(\mathbf{x} \setminus x_i, \mathcal{M}_j)))$. The NML approach is based on the same loss function, but, as explained in the Appendix, rather than predicting by using the leave-one-out ML estimator, one sequentially predicts the full sequence $\mathbf{x} = (x_1, \dots, x_n)$ using the prediction strategy that is worst-case optimal relative to the element of \mathcal{M} that one should have used with hindsight, the worst-case being taken over all possible data sequences.

Summarizing, we may broadly distinguish between *truth-dependent* approaches such as Neyman–Pearson tests and AIC,¹ and *agnostic approaches* such as cross-validation and NML. Truth-dependent approaches are designed to give good results with high probability or in expectation according to some distribution p . In agnostic approaches, distributions are only used as predictors, and the merit of a model in light of the data \mathbf{x} is solely determined by how well such distributions predict \mathbf{x} . It is in this sense that introductory papers (e.g. Myung, Navarro, and Pitt (2006)) describe NML as being “free” from assumptions about true distribution: it is an agnostic method by design.

Having made this distinction between agnostic and truth-dependent procedures, it is worth considering the advantages built into the agnostic methods. Besides avoiding the previously-discussed problem of non-interpretatable outputs, agnostic methods also have another advantage: *when comparing a finite number of models with an agnostic approach, the better model must win, eventually*. To explain what this means (see Section 3.2 for more details) consider the case of just two models, \mathcal{M}_a and \mathcal{M}_b . Suppose one observes more and more data x_1, x_2, \dots , the sequence being such that the best predictor of the data in \mathcal{M}_a eventually keeps outperforming the best predictor of the data in \mathcal{M}_b . Given such a sequence, the agnostic approaches will eventually select \mathcal{M}_a . Specifically, for an agnostic approach it is guaranteed that, for all infinite sequences x_1, x_2, \dots such that

$$\begin{aligned} & \min_{f(\cdot|\theta) \in \mathcal{M}_a} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Loss}(x_i, f(\cdot | \theta, \mathcal{M}_a)) \\ & < \min_{f(\cdot|\theta) \in \mathcal{M}_b} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{Loss}(x_i, f(\cdot | \theta, \mathcal{M}_b)), \end{aligned} \quad (3)$$

one has the assurance that, for all large n larger than some n_0 , the model \mathcal{M}_a will be selected rather than \mathcal{M}_b . Here the number n_0 may depend on the particular sequence x_1, x_2, \dots : for some sequences, the better model will be identified earlier than for others.

For truth-dependent approaches, the guarantee that the best model will eventually be selected can only be given for a small subset of the sequences satisfying (3), namely those sequences x_1, x_2, \dots for which there exists a distribution f in $\mathcal{M}_a \cup \mathcal{M}_b$, so that x_1, x_2, \dots may be regarded as a “typical outcome” of f . In practice, however, we often have to deal with atypical outcomes: supposedly real-valued variables (e.g., normally distributed data) can very easily contain repeated values – cases where $x_i = x_j$ for some $i \neq j$ – due to round-off errors and other imperfections,

¹ To see that AIC is a truth-dependent approach, note that it tells us to select the model minimizing $\text{AIC}(\mathbf{x}, d) = -\log f(\mathbf{x} | \hat{\theta}(\mathbf{x}, \mathcal{M}_d)) + d$, where d is the model dimension. While the first term is “agnostic”, the second term (d) is truth-dependent, since it has been designed to make $\text{AIC}(\mathbf{x}, d)$ an unbiased estimator of the prediction loss that can be achieved with model \mathcal{M}_d . “Unbiased” means “giving the right answer in expectation”, the expectation being taken under a distribution p that is assumed to be in a (suitably defined) closure of the list of models $\mathcal{M}_1, \mathcal{M}_2, \dots$. We note that Bayesian inference cannot easily be put into one of the two categories: some variations may be called truth-dependent, others may not (Grünwald, 2007, ch. 17).

an occurrence that should have probability 0 (see, e.g. Grünwald (2007), ch. 17). More generally, real-world data sets tend to be riddled with data missing (not at random), data entry errors, and (particularly in the social sciences) a host of weak correlations (e.g. Meehl (1990)). The net result is that, in many cases, even very large empirical data sets will have some characteristics that make them rather atypical sequences. It is also for this reason that the predictive guarantees for the agnostic approaches are in practice somewhat reassuring.

The previous remarks notwithstanding, it is worth pointing out that there is, of course, a weak spot in the agnostic approaches: one can measure prediction error in many different ways, so why should one focus on the log loss? The model that predicts best in terms of log loss may not be the best in terms of some other loss functions such as 0/1-loss. Indeed, there are approaches which try to extend MDL and related approaches beyond the log loss (Grünwald, 2007, ch. 17); the methodology of *structural risk minimization* (Vapnik, 1998) may also be viewed in this manner. Nevertheless, there are certain properties of the log loss which make it particularly attractive, such as the fact that it is the only local proper scoring rule (Bernardo & Smith, 1994), that it has a clear interpretation in terms of data compression and sequential gambling (Grünwald, 2007), and, as we discuss below, that it has good convergence properties in the hypothetical case in which the true distribution does reside in one of the models after all.

3.2. True distributions in the analysis of algorithms

At this point, we turn to a discussion of the performance of different model selection algorithms. As with the previous discussion regarding the design of the methods, it is useful to analyze the methods under different assumptions about the nature of the data generating mechanism. Suppose that a method is applied to data $\mathbf{x} = x_1, \dots, x_n$, and the inferred model is then used to make predictions about future data $\mathbf{y} = x_{n+1}, \dots, x_{n+m}$. If the data generating machinery may change in arbitrary ways at time $n + 1$, then no method can be expected to work well. In such extreme scenarios, agnostic approaches will fail to make good predictions just as much as truth-dependent methods. In order for any method to work well, there has to be some kind of constraining mechanism which pertains to both \mathbf{x} and \mathbf{y} . It is therefore of some interest to compare the actual behavior of some well-known agnostic and truth-dependent model selection methods for a variety of such constraining mechanisms. Following Grünwald (2007), let us consider what are arguably the four most important cases:

1. *Mechanism satisfies Eq. (3)*. Suppose we are to choose between two possible models \mathcal{M}_0 and \mathcal{M}_1 , and that the constraining mechanism is such that Eq. (3) holds, either for $a = 0, b = 1$ or vice versa. This may be one of the weakest assumptions under which some form of inductive inference is possible at all. In this case, NML, the Bayes factor method, BIC, LOOCV and AIC will all select the best-predicting model \mathcal{M}_a for all large enough samples. In such cases, for large n , the truth-dependent component of $\text{AIC}(\mathbf{x}, d)$ becomes negligible compared to its agnostic component. If, however, we assume that \mathcal{M}_0 is nested into \mathcal{M}_1 , and Eq. (3) holds with equality, then NML, BIC and the Bayes factor method will select \mathcal{M}_0 for large n (a form of Occam’s razor), whereas for many sequences, AIC and LOOCV will not. Grünwald (2007) argues extensively why such a version of Occam’s razor is desirable. Note that all this holds quite irrespective of whether the “true” data generating mechanism is in any of the models, or is even a probability distribution; it may just as well be deterministic.

If we allow the list of models to contain an arbitrary but finite number of elements, then the same story still holds.

However, in practice, this list is often countably infinite, or (equivalently, as it turns out), it is allowed to grow with n . The prototypical example is linear regression with polynomials, where the outcomes are pairs (Z, X) , with, say, $Z \in [-1, 1]$ and $X \in \mathbb{R}$. Model \mathcal{M}_d prescribes that $X = \sum_{j=0}^{d-1} \alpha_j Z^j + U$, where $(\alpha_0, \dots, \alpha_{d-1})$ is a parameter vector and U is normally distributed noise with mean 0. We would like to learn the best polynomial model of the data, without assuming any a priori bound on the degree d . In such cases, there can be data sequences for which a particular degree d_0 leads, asymptotically, to the best predictions, yet, no matter how many data are observed, none of the methods will select degree d_0 , not even the agnostic ones.

2. *True distribution in one of the models.* At the other extreme, suppose we have the collection of models $\mathcal{M}_1, \mathcal{M}_2, \dots$, where the k -th model has k free parameters. Moreover, the data are sampled from a distribution $f(\cdot \mid \theta, \mathcal{M}_d)$ that falls inside the d -th model. In this situation, NML and other MDL-related methods, as well as BIC and the Bayes factor method, perform very well in the sense that, for all d such that \mathcal{M}_d is on the list, for almost all $f(\cdot \mid \theta, \mathcal{M}_d)$, with $f(\cdot \mid \theta, \mathcal{M}_d)$ -probability 1, they output “ \mathcal{M}_d ” for all large n . For an explanation of the “almost”, see Grünwald (2007). AIC and leave-one-out cross-validation do not share this property of statistical consistency, and may, with positive probability, output a model of larger dimension than the minimal d for which \mathcal{M}_d contains the true distribution. These results hold both if the list of models is finite and if it is countably infinite.
3. *True distribution in model closure.* A commonly studied situation in statistics is to assume that the list $\mathcal{M}_1, \mathcal{M}_2, \dots$ is countable, and that data are sampled from some distribution p , which is not in any of the models of the list, but which can be arbitrarily well-approximated by the list, in the sense that $\lim_{d \rightarrow \infty} \min_{f \in \mathcal{M}_d} D(p, f) = 0$. Here D is some suitably chosen distance measure for probability distributions. In our polynomial example, this would correspond to the true p stating that $X = g(Z) + U$, where g is a continuous function on $[-1, 1]$ that is, however, not itself a polynomial. In such cases, the best predictions can be obtained by choosing a small model at small sample sizes, and gradually choosing more complex models (higher-order polynomials) as the sample size increases. Qualitatively speaking, Bayes factor, BIC, AIC, NML and LOOCV all behave in this manner. But a more detailed view reveals important differences: if the models $\mathcal{M}_1, \mathcal{M}_2, \dots$ are sufficiently regular, and the distribution p is sufficiently smooth, then AIC and LOOCV will converge faster than NML, BIC and Bayes. More precisely, suppose we fix a method and for each n , we use it to infer a model and then predict future data based on that model. For all methods, the expected prediction loss will get smaller as n increases, and it will converge to the same asymptotic optimum. However, the convergence is slower (by a logarithmic factor) for Bayes, BIC and NML. On the other hand, if either (a) the models $\mathcal{M}_1, \mathcal{M}_2, \dots$ are not “regular”, or, (b), if the true p is not smooth, then AIC may fail dramatically, whereas Bayes factor, LOOCV and NML will still tend to converge. A common example of (a) is model selection for feature selection models, in which the number of considered models with d degrees of freedom is exponential in d (Yang, 1999). An example of (b) within the polynomial setting arises if the function g is discontinuous, or if it tends to $\pm\infty$ at the boundaries of its domain. This failure of AIC is due to its truth-dependent nature: it has simply not been designed to work well for true distributions that are as in situation (a) and (b).
4. *True distribution not in model closure.* Finally, consider the possibility that there exists a true distribution p that cannot be arbitrarily well-approximated by members of models

$\mathcal{M}_1, \mathcal{M}_2, \dots$, while nevertheless, some model \mathcal{M}_d contains a useful f that is “close” to p in that it tends to predict data reasonably well. This case is related to, but less general than, scenario 1 above, and essentially the same facts hold. To illustrate, suppose, for simplicity, that the data are i.i.d. according to both the ‘true’ p and all f in all of the $\mathcal{M}_1, \mathcal{M}_2, \dots$ under consideration, and suppose that one of the models is “best” in the sense that the following analogue of (3) holds: for some model \mathcal{M}_a on the list,

$$\min_{f(\cdot|\theta) \in \mathcal{M}_a} E_p[\text{LOSS}(X, f(\cdot \mid \theta, \mathcal{M}_a))] < \min_{b: b \neq a, \mathcal{M}_b \text{ on the list}} \min_{f(\cdot|\theta) \in \mathcal{M}_b} E_p[\text{LOSS}(X, f(\cdot \mid \theta, \mathcal{M}_b))]. \quad (4)$$

If the list is finite, say $\mathcal{M}_1, \dots, \mathcal{M}_D$, and (4) holds, then, with p -probability 1, all methods will select model \mathcal{M}_a for all large enough sample sizes n . This means that the p -probability that a suboptimal model $\mathcal{M}_b, b \neq a$ is selected based on data X_1, \dots, X_n goes to 0 with increasing n , where the exact rate at which it goes to 0 may depend on the precise relation between p and the various models on the list. In case the models are nested and (4) holds with equality, then, once again, for large n , NML, Bayes factor and BIC will tend to select the *smallest* model \mathcal{M}_a that achieves the minimum in (4), whereas, for some combinations of p and $\mathcal{M}_1, \dots, \mathcal{M}_D$, AIC and LOOCV will not. In case (4) holds but the list is countably infinite, then there exist scenarios in which none of the methods work fine for large samples, i.e. they keep selecting models that are further than some ϵ from the minimum (4), no matter how large n . Here ϵ is a positive constant and, being a constant, it does not tend to 0 with increasing n (Grünwald & Langford, 2007). Thus, neither NML (despite its agnosticity) nor the Bayes factor method are guaranteed to work in such a scenario. The only methods we are aware of that handle such a scenario well, are those developed in the structural risk minimization literature (Vapnik, 1998), but they tend to perform less than optimally in scenario 2 and 3 (Grünwald, 2007).

The upshot is that even agnostic methods may not always work well in all relevant settings. Nevertheless, we may still expect agnostic methods to be more robust than truth-dependent methods. Moreover, if a method that performs well in all settings 1–4 will ever be found, it is sure to be a method of the distribution-free kind. As an aside, Van Erven, Grünwald, and de Rooij (2008) present an agnostic approach that combines the best of NML and LOOCV, and is probably the first known method that probably performs well in all cases discussed under settings 2 and 3 above; yet it still fails with countably infinite lists in settings 1 and 4.

To summarize, in this section we have aimed to give a general overview of the role played by the concept of a “true distribution” for a variety of different model selection algorithms. We have done so, in part, because we think it provides a useful expansion of the necessarily-oversimplified treatment given in tutorial papers (e.g. Myung et al. (2006)). However, it also provides an appropriate foundation for our discussion of the claims made recently by KW. It is to this topic that we now turn.

4. A Bayesian decision-theoretic view on NML

In a recent paper, KW provide a Bayesian decision theoretic interpretation for the NML criterion, and use this interpretation to suggest that it is meaningless to refer to NML as an agnostic method. In order to characterize NML in terms of the more general Bayesian decision-theoretic framework, the derivation relies on three key premises:

- (1) Data arise from some unknown distribution (i.e., $\mathbf{x} \sim G$), and we have a prior over this distribution described by a Dirichlet process (DP (see Ferguson (1973))) with concentration parameter $c \rightarrow 0$ (i.e., $G \sim \text{DP}(G_0, 0)$).
- (2) We want to select a parameter θ that belongs to one of the models $\mathcal{M}_1, \dots, \mathcal{M}_D$, and in addition to the loss incurred due to the expected Kullback–Leibler discrepancy between $f(\cdot|\hat{\theta}, \mathcal{M})$ and the true distribution G , we suffer a “complexity penalty” $v(\mathcal{M}, n)$ that depends only on the model \mathcal{M} from which $\hat{\theta}$ is drawn and the sample size n .
- (3) The complexity penalty $v(\mathcal{M}, n)$ is defined by

$$v(\mathcal{M}, n) = \log \int_{\mathcal{X}^n} f(\mathbf{y}|\hat{\theta}(\mathbf{y}, \mathcal{M})) d\mathbf{y}. \quad (5)$$

KW show that under conditions (1) and (2), the optimal Bayesian choice for θ is the ML estimator $\hat{\theta}(\mathbf{x}, \mathcal{M}_d)$ within the model \mathcal{M}_d that minimizes, over all $d \in \{1, \dots, D\}$,

$$-\log f(\mathbf{x}|\hat{\theta}(\mathbf{x}, \mathcal{M}_d)) + v(\mathcal{M}_d, n). \quad (6)$$

Thus, they conclude, if the penalty term (5) is plugged into (6), then the optimal Bayesian choice is to select θ from the model \mathcal{M}_{d^*} , where d^* is given by

$$\begin{aligned} d^* &= \arg \min_d \left\{ -\log f(\mathbf{x}|\hat{\theta}(\mathbf{x}, \mathcal{M}_d)) + \log \int_{\mathcal{X}^n} f(\mathbf{y}|\hat{\theta}(\mathbf{y}, \mathcal{M}_d)) d\mathbf{y} \right\} \\ &= \arg \max_d p^*(\mathbf{x} | \mathcal{M}_d), \end{aligned} \quad (7)$$

where p^* is given by (1), and the second equality follows because the logarithm is a monotonically increasing function. Hence, when assumptions (1)–(3) are met, the Bayes optimal model coincides with the model preferred under the NML criterion. In the following sections we critically discuss this derivation and its supposed implications. In doing so, we distinguish between two major problems (Section 5) and three minor concerns (Section 6). We also briefly comment on another issue brought up by KW, namely the fact that for many models, the NML is undefined (Section 7).

5. Major problems

In this section, we raise two major sources of concern with the KW derivation, namely that it is incomplete in an essential sense (Section 5.1), and that the main conclusion drawn from the derivation does not follow (Section 5.2). However, we wish to emphasize that our concerns do not lie with the formal aspects to the derivation itself, which appears to be entirely correct.

5.1. Incompleteness of the characterization

In the context of discussing what conclusions can be drawn from their derivation, KW (p. 520) state that they have “discovered the NML criterion using Bayesian decision theory” (emphasis added). This statement highlights one of the main problems we have with their characterization, namely that it does not provide any Bayesian interpretation, characterization or explanation of the complexity term (5). Rather, they show that any model selection criterion of a “fit plus complexity” format is consistent with the Bayesian framework, using assumptions (1) and (2) above. The specific application to NML via assumption (3) is not explained anywhere in their paper – it is simply introduced on p. 519 with no justification given other than the statement that it is “[an] alternative penalty term ... for model simplicity”. They do not state why this particular penalty term would be of interest to the statistician, even though it is clearly an essential component to NML. After all, it is exactly this term that distinguishes the NML criterion from many other existing criteria such as AIC and BIC. In

our view, this is not really a “discovery” at all, and it makes it hard to see how their characterization is helpful or informative as to the nature of NML itself. Indeed, the KW derivation can also be used to “discover” BIC and (as KW in fact point out themselves) AIC – two criteria that behave very differently from NML in many situations (see Section 3). This is achieved simply by replacing $v(\mathcal{M}_d, n)$ as in (5) by $(k_d/2) \log n$ (which yields BIC) or k_d (producing AIC), where k_d is the dimensionality of model \mathcal{M}_d . There is no particular reason given for the use of one penalty function over any other one. This differs from all four previously existing interpretations of NML, each of which derives the penalty term from some more basic considerations.² In short, it seems to us that the KW derivation is incomplete in a very fundamental sense, because it does not give any reason why a statistical decision-maker should adopt a complexity term that has the specific mathematical form specified in assumption (5).

To illustrate the point, consider the following (highly exaggerated) example. To our knowledge, no-one has seriously proposed the use of a penalty function of the form

$$v(\mathcal{M}_d, n) = \begin{cases} 0 & \text{if } \mathcal{M}_d \text{ is Favorite Model X} \\ \infty & \text{otherwise} \end{cases} \quad (8)$$

but clearly, it would be straightforward to substitute this penalty function into the derivation provided by KW and thereby “discover” a model selection criterion that always prefers Favorite Model X. Taking KW at face value, we would be able to say that we have derived the criterion using Bayesian decision theory. However, it would be entirely unreasonable to specify $v(\mathcal{M}_d, n)$ in this fashion, and (we hope) no-one would accept the proposition that Bayesian methods actually justify this sort of behavior. Obviously, the problem is that we have provided no justification whatsoever for adopting this particular choice of $v(\mathcal{M}_d, n)$, and so any analyses we conduct on the basis of this choice would be of little interest to any statistician, Bayesian or otherwise. The point here is that the “Bayesian discovery” of NML made by KW is of exactly the same character as the “discovery” of the criterion that always prefers model X: namely, it demonstrates that NML is consistent with Bayesian theory, but provides no actual reason to use it in any practical situation. Their derivation is so broad as to encompass any criterion of a “fit plus penalty” format. This, in our view, cannot be called a “discovery” in any interesting sense.

The point of the previous example is to illustrate the importance of having some reason for choosing a particular penalty function. With that in mind, one way to think about our argument is to ask the following question: “if someone else had not already proposed the NML approach, would any Bayesian ever have contemplated the complexity term (5) in combination with this particular Dirichlet process prior?” It seems unlikely – indeed, KW state explicitly that it “is difficult to understand as a penalty term” (p. 520), with the implication that this is an inherent problem for NML. This is somewhat unfortunate, since the information-theoretic perspective provides a very natural interpretation of this term, as the minimax coding or prediction regret (Appendix A). Accordingly, we have a good *information-theoretic* reason to use NML. The problem here is that there is no corresponding *Bayesian* interpretation provided by KW. Without having been implicitly guided by the information-theoretic results provided by Rissanen (2001), Shtarkov (1987) and others, it seems highly unlikely that any Bayesian would be inclined to choose $v(\mathcal{M}, n)$ in the manner specified in (5), making KW’s (2006) derivation somewhat post hoc at best.

² Moreover, this is also the case for the original derivations of the AIC and the BIC. Akaike (1973) derived AIC by correcting for a bias in the model selection procedure implied by maximum likelihood methods, while Schwarz (1978) derived BIC by taking an asymptotic expansion of the logarithm of the Bayesian marginal probabilities.

