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*

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 and Walker (2006), 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.

Keywords: Minimum description length; normalized maximum likelihood; Bayesian inference

1 Introduction

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

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

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 (Shtarkov, 1987; Rissanen, 2001). 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 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 $L_C(\mathbf{y}) = -\log f(\mathbf{y})$. Vice versa, 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 Equation 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:

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

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 Section 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$ of models (families of probability distributions), as well as data $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$. 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 Equation 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 we want to choose between a model $\mathcal{M}_1 = \{f(\cdot \mid \mu) \mid \mu \in \mathbb{R}\}$ and its submodel $\mathcal{M}_0 = \{f(\cdot \mid \mu) \mid \mu = 0)$, where, for $\mathbf{x} \in \mathcal{X}^n$, $f(\mathbf{x} \mid \mu)$ 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 any more, 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 any more. 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 $\textsc{loss}(x_i, f) := -\log f(x_i)$: 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, $\sum_{i=1}^{n} \textsc{loss}(x_i, f(\cdot \mid \hat{\theta}(\mathbf{x} \setminus x_i, \mathcal{M}_j)))$. 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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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

[1] To see that AIC is a truth-dependent approach, note that it tells us to select the model minimizing $\mathrm{AIC}(\mathbf{x}, d) = -\log f(\mathbf{x} \mid \hat{\theta}(\mathbf{x}, \mathcal{M}_d)) + d$, 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 $\mathrm{AIC}(\mathbf{x}, d)$ 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$. 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).

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 et al. 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-interpretable 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 $x_1, x_2, \ldots$, 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 $x_1, x_2, \ldots$ such that

$$\min_{f(\cdot|\theta)\in\mathcal{M}_a} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \textsc{loss}(x_i, f(\cdot \mid \theta, \mathcal{M}_a)) < \min_{f(\cdot|\theta)\in\mathcal{M}_b} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \textsc{loss}(x_i, f(\cdot \mid \theta, \mathcal{M}_b)), \quad (3)$$

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 $x_1, x_2, \ldots$: 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 $x_1, x_2, \ldots$ for which there exists a distribution f in $\mathcal{M}_a \cup \mathcal{M}_b$, so that $x_1, x_2, \ldots$ 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, 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 $\mathbf{x} = x_1, \ldots, x_n$, and the inferred model is then used to make predictions about future data $\mathbf{y} = x_{n+1}, \ldots, x_{n+m}$. 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 Equation 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 Equation 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 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 Equation 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 $(\alpha_0, \ldots, \alpha_{d-1})$ 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$, 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$ 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 $\lim_{d\to\infty} \min_{f\in\mathcal{M}_d} D(p, f) = 0$. 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$ 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$ 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$, 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 $\mathcal{M}_1, \mathcal{M}_2, \ldots$ 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,

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

If the list is finite, say $\mathcal{M}_1, \ldots, \mathcal{M}_D$, 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 $X_1, \ldots, X_n$ 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 that 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 $\mathcal{M}_1, \ldots, \mathcal{M}_D$, 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 and 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 optimal 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 & De Rooij (2007) present an agnostic approach that combines the best of NML and LOOCV, and is probably the first known method that provably 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, tshe 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 $c \to 0$ (i.e., $G \sim \mathrm{DP}(G_0, 0)$).
(2) We want to select a parameter $\dot{\theta}$ that belongs to one of the models $\mathcal{M}_1, \ldots, \mathcal{M}_D$, and in addition to the loss incurred due to the expected Kullback-Leibler discrepancy between $f(\cdot|\dot{\theta}, \mathcal{M})$ and the true distribution G, we suffer a "complexity penalty" $v(\mathcal{M}, n)$ that depends only on the model $\mathcal{M}$ from which $\dot{\theta}$ is drawn and the sample size n.
(3) The complexity penalty $v(\mathcal{M}, n)$ is defined by

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

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

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

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

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

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,

[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.

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

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

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.

5.2 Unjustified Conclusions

In the previous section, we raised the concern that KW's derivation is incomplete, since it provides no basis for the choice of penalty function. In essence, we were arguing that the derivation – though technically correct – is not particularly helpful. In this section, we raise a different concern, namely the fact that KW appear to assert some kind of special privilege to their proof over other proofs, somehow invalidating the logic of previous justifications for the use of NML. The relevant quote from their paper is as follows: after completing their derivation, KW (p. 520) suggest that

> [T]he idea that NML makes no mention of a true distribution is a meaningless point. We have discovered the NML criterion using Bayesian decision theory and have, as a component of this procedure, explicitly introduced the notion of a true distribution function.

Importantly, this is the main conclusion of the paper. The *premise* here appears to be that "NML can be derived when we assume that a true distribution exists", from which they draw the *conclusion* that "previous derivations that did not need this assumption are meaningless".

We have four problems with this statement. Firstly and most importantly, it is hard to see how this conclusion can possibly follow from the premise. Nothing in KW's derivation falsifies the logic of the previous constructions provided by Shtarkov and Rissanen, so to the extent that those derivations were valid previously, they remain so now. Accordingly, there is still a perfectly good reason to use NML even if no data-generating distribution exists (see Section 3). While we agree that NML can be derived when a true distribution *is* assumed, it is hardly meaningless to observe that we can derive NML *without* having to make this assumption.

Our second problem is somewhat related to the first, in that one of the strengths of the original NML proof is that it makes only very weak assumptions (though, even so they are sometimes violated; see Section 7), implying that NML may be used in a broad range of situations. By contrast, the three conditions that apply to KW's derivation are fairly restrictive, and would only justify the use of NML in a few specific situations: for instance, KW require that our prior beliefs about the true data-generating distribution be captured by the statement $G \sim \mathrm{DP}(G_0, 0)$, whereas no such restrictions are required for Rissanen's or Shtarkov's proofs to hold. Thirdly, as argued previously in Section 5.1, the KW derivation is incomplete in a fashion that other derivations are not, so in our view it would be preferable to use one of the other proofs to justify the use of NML. Finally, as we will discuss in Section 6, there are some doubts as to how reasonable the underlying assumptions are, so unlike the other derivations of NML that hold quite generally, the KW approach does not necessarily apply in practical situations.

Similarly, though it is somewhat tangential to their derivation, KW also note that NML sometimes corresponds to the Bayes factor approach to model selection with Jeffreys' prior, and write (on p. 519)

> Bayes factors are recognized as being based on a 0-1 loss function which implicitly assumes that one of the models under consideration is the true model. This contradicts one of the key ideas for NML, namely that it is free from assumptions of a true model.

The first criticism of the previous statement still applies here: neither Rissanen's derivation, nor the prequential derivation given in the appendix, require the existence of a true distribution or a 0/1-loss function. The fact that there is *also* a Bayesian derivation that does assume these things and establishes a correspondence to Jeffreys' prior is irrelevant; it simply does not make the various other derivations meaningless or false.

6 Minor Issues

We proceed to discuss some minor concerns about the KW derivation, relating to the specification of the prior, the characterization of the decision problem, and inconsistencies with the assumptions used in previous work. Unlike the problems raised in the previous section, none of these issues should be taken to be strong criticisms of the paper, so much as minor caveats. We consider each of these in turn.

6.1 Specification of the prior

The KW paper relies on a Bayesian decision-maker who places a Dirichlet process (DP) prior (Ferguson, 1973) to describe his or her prior beliefs about an unknown probability distribution G. The DP prior is used to place a "nonparametric" prior over G, in which one seeks to avoid making restrictive assumptions about the family of distributions to which G might belong. From a Bayesian perspective, the nonparametric approach requires us to select a prior distribution that has broad support across the space of probability distributions. The DP prior serves this purpose, and specifies a distribution over random probability measures, parametrized by the base distribution G_0 (corresponding roughly to one's initial guess about G), and a concentration parameter c. However, although the DP has full (weak) support, it concentrates (with probability 1) on a set of discrete distributions (e.g., Sethuraman, 1994), which tends to limit its usefulness as a generic prior in some cases (e.g., Petrone & Raftery, 1997). In some contexts, however, the restriction to discrete distributions is actually quite useful, and for this very reason the DP has become a popular choice for specifying priors over countable mixtures (e.g., Escobar & West 1995).

It is important to note that KW use the DP in the general sense, using the limiting DP prior with $c \to 0$ to describe the prior belief about a *generic* unknown distribution G. The result is that, as they note, they rely on a prior that concentrates with probability 1 on point-mass distributions. This reliance plays an important role in their subsequent derivation: under this limiting prior, the Bayesian predictive distribution for future data converges to the empirical distribution of $\mathbf{x}$ (e.g., Ghosh & Ramamoorthi, 2003, Theorem 3.2.7). This in turn implies that the Bayesian maximum utility parameter estimate under Kullback-Leibler loss is equivalent to the frequentist MLE $\hat{\theta}(\mathbf{x}, \mathcal{M})$ (as discussed by KW). Obviously, this does not hold for other values of c, since in general the predictive distribution under a DP prior is a weighted mixture of G_0 and the empirical distribution. In short, although technically correct, the correspondence that they establish holds only for this rather odd special case; a case that KW appear to have chosen primarily to ensure that their Bayesian parameter estimation procedure mimics a frequentist one. [3]

6.2 Specification of the decision problem

A second issue relates to the manner in which KW specify the decision problem. The Bayesian decision procedure described by KW equates the utility of a model with the utility of its best parameter value. This manner of setting up the problem is highly biased towards complex models, since in its simplest form it reduces to picking the model that can provide the best fit in a maximum likelihood sense. In order to redress this, they then introduce a complexity penalty into the utility function, as suggested by Kadane and Dickey (1980). We note that such an approach, while certainly correct, is by no means standard Bayesian practice. In a standard Bayesian textbook, Berger (1985, p. 284) argues that one of the advantages of the Bayesian approach is that it *automatically* takes model complexity into account, without the need for any explicit penalties (Mackay (2003) refers to this as the "Bayesian Occam's razor"). Indeed, this does happen under standard parametric priors and standard utility functions (possibly, but not necessarily of the 0/1-type; see, e.g. Bernardo & Smith 1994, ch. 6). However, by associating model utility with maximum likelihood parameter utility, KW are unable to take advantage of one of the most useful features of Bayesian inference, and are forced to reintroduce it via the unexplained penalty function $v(\mathcal{M}, n)$.

[3] Moreover, although it appears in the *model-selection* procedure described by Equation 1, when using MDL one would generally *not* use the MLE as one's optimal parameter choice within the selected model. This point is particularly important and we will return to it in Section 6.3.

6.3 *Incompatibility with assumptions of the original derivation*

A third point to make is that KW's characterization actually discards a number of the existing parallels between MDL and Bayesian methods. As discussed, the NML criterion originally arose as a specific instance of the broader MDL approach to inductive inference, which is in fact closely related to Bayesian inference (see Grünwald 2007, p. 531-550). Just as in the Bayesian approach, MDL inference invariably starts by putting a distribution on observables. In the NML version of MDL discussed here, one actually puts a uniform prior $\pi(d) = 1/D$ on the model set $\{\mathcal{M}_1, \ldots, \mathcal{M}_D\}$ (and indeed when one compares countably infinitely many models, the prior on the model index d becomes *essential* in the MDL approach; see Grünwald, 2007, p. 406 & p. 423). One then associates each model $\mathcal{M}_d$ with a distribution on $\mathcal{X}^n$, in this case the NML distribution $p^*(\mathbf{x} \mid \mathcal{M}_d)$ given by (1). Thus, under the more typical Bayesian characterization of NML, the criterion may be interpreted as advocating a "maximum posterior model" $\mathcal{M}_d$ under a uniform prior distribution on d. Accordingly, the NML criterion (as with other versions of MDL) already has a Bayesian flavor, with the NML distribution $p^*(\mathbf{x} \mid \mathcal{M}_d)$ playing a role similar to the Bayesian marginal distribution $p(\mathbf{x} \mid \mathcal{M}_d) = \int_\Theta p(\mathbf{x} \mid \theta, \mathcal{M}_d)\ dw(\theta)$, for some prior distribution w. In fact, although we do not do so here, it is not too difficult to construct cases in which the Bayesian marginal probability corresponds to the NML probability more or less exactly (this can be made to hold for any sample size n, with the usual convergence to a Jeffreys' prior as $n \to \infty$). These relationships, however, are completely lost in the KW derivation, because KW decouple the two terms that comprise the NML criterion. In their approach, the numerator $f(\mathbf{x}|\hat{\theta}(\mathbf{x}, \mathcal{M}))$ arises as a consequence of the DP$(G_0, 0)$ prior, while the denominator $\int_{\mathcal{X}^n} f(\mathbf{y}|\hat{\theta}(\mathbf{y}, \mathcal{M}))d\mathbf{y}$ is a consequence of the loss function. The fact that the denominator is in fact the integral of the maximized likelihood in the numerator (hence giving rise to the name *normalized* maximum likelihood, and making $p^*(\cdot \mid \mathcal{M}_d)$ a distribution over possible data sets) actually becomes irrelevant in KW's approach.

In much the same manner, it should be noted that in MDL approaches to density estimation relative to a parametric model $\mathcal{M}_d$, one generally does *not* use a maximum likelihood estimator (this is explained further in the final paragraph of the appendix). Instead, MDL's information-theoretic derivations lead one to adopt either a truncated ML estimator (in "two-part MDL") or a *predictive distribution* (in "predictive MDL") corresponding to a Bayesian predictive distribution relative to $\mathcal{M}_d$ and some smooth prior on the model $\mathcal{M}_d$ which varies from case to case. For example, with the Bernoulli model, the ML estimator after observing n biased coin tosses with h heads and $n-h$ tails, would be $p(\text{heads}) = h/n$, whereas the predictive MDL estimator would be a smoothed version thereof, $p(\text{heads}) = (h + \frac{1}{2})/(n + 1)$ (Grünwald, 2007, section 15.4). It is then surprising to see that KW's derivation is based on a special nonparametric prior under which the Bayesian predictive distribution coincides with the maximum likelihood estimate. Whereas most Bayesians, when working with a fixed parametric model, would

prefer using a Bayesian predictive distribution based on a smooth prior defined relative to model $\mathcal{M}_d$, and MDL prescribes the use of the same or similar predictive distributions, KW rederive NML using a *different* predictive distribution.

Summarizing, KW have discarded two facts which already make MDL inference closely related to mainstream Bayesian inference: the fact that the NML distribution is a distribution, and the fact that MDL estimates/predictive distributions often coincide with Bayesian predictive distributions, but not with ML estimators. Of course, discarding these facts does not introduce any errors into their derivation, but it does mean that they are missing some of the very key components of the original work.

7 Concluding Remarks

Our main goal in this paper has been to discuss some of the problems associated with the KW derivation of NML probabilities, and to elaborate on the claim that NML does not depend on the assumption of a true distribution. However, we would like to end the paper by noting that there is one serious issue in which we agree with KW's position: the NML approach has some technical difficulties which (at least in the simple form presented here), make it useless for many practical model selection tasks. The main problem is that the integral in the denominator diverges for some of the simplest and most often used parametric models, including the normal location and scale families. This issue has in fact been known since 1996 and is the subject of considerable discussion in the literature (see Grünwald 2007, ch. 11). Although it is not central to their derivation, the issue is briefly raised by KW (on p. 520), so it is worth explicitly stating that we agree that this is a genuine, and quite serious, issue with the NML approach.

More generally, we suspect that it may be the case that some researchers (including us) have at times overemphasized the importance of NML within the MDL framework, perhaps giving the impression that the two are equivalent. Given this possibility, it is important to note that the central idea in MDL is to base statistical inference on universal coding (see Grünwald 2007, for an extensive discussion): as it happens, the NML method is only one of at least five good methods for constructing universal codes, so the MDL framework is much broader than NML (only two of the 19 chapters in Grünwald's (2007) book deal primarily with NML, for instance). That said, because it has certain optimality properties which the other methods lack, NML has tended to be the preferred method in recent years. However, in those cases where NML cannot be applied, the other methods usually still can. Importantly, one of these five types of universal codes is based on Bayesian marginal likelihoods, and so it should be no surprise that there is generally a close correspondence between Bayesian and MDL methods. Nevertheless, this correspondence is of quite a different type than the KW derivation suggests.

8 Acknowledgements

We sincerely thank the anonymous reviewer who suggested that the meaning of "assuming a true distribution" should be discussed in more detail. This led to some essential additions to the text. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and by an Australian Research Fellowship (ARC grant DP-0773794). This publication only reflects the authors' views.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.) *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado

Balasubramanian, V. (2005). MDL, Bayesian inference and the geometry of the space of probability distributions. In P. Grünwald, J. I. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis,* revised and expanded 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.

Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. Chichester, UK: Wiley.

Browne (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108-132.

Dawid, A.P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *Journal of the Royal Statistical Society, Series A 147*(2), 278-292.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association, 90*, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1*, 209-230.

Ghosh, J. & Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. New York: Springer

Grünwald, P. (2005). Minimum description length tutorial. In P. Grünwald, J. I. Myung & M. A. Pitt (Eds). *Advances in Minimum Description Length: Theory & Applications*. Cambridge, MA: MIT Press. Note that Chapter 17, the most relevant chapter for this article, is available freely on the web.

Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press

Grünwald, P. & J. Langford (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning 66*(2-3), pages 119-149. Cambridge, MA: MIT Press

Van Erven, T., Grünwald, P. & de Rooij, S. (2008) Catching up faster in Bayesian model

selection and model averaging. *Advances in Neural Information Processing Systems, 20.*

Karabatsos, G. & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology, 50*, 517-520.

Kadane, J. & Dickey, J. (1980). Bayesian decision theory and the simplification of models. In J. Kmenta, & J. Ramsey (Eds.), *Evaluation of Econometric Models.* New York: Academic Press.

Mackay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge, UK: Cambridge University Press.

Meehl (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195-244 (Monograph supplement 1-V66).

Myung, J. A., Navarro, D. J. & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167-179.

Petrone, S. & Raftery, A.E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics and Probability Letters, 36*, 69-83.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*, 1712-1717.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*, 639-650.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission 23*(3), 3–17.

Vapnik, V. (1998). Statistical Learning Theory. New York: John Wiley

Yang, Y. (1999). Model Selection for Nonparametric Regression. *Statistica Sinica 9*, 475–499.

A The Prequential Interpretation of NML

Suppose we want to predict a sequence of outcomes $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where each x_i is an element of some space $\mathcal{X}$. The x_i are given to us one at a time. At each point in time i, we want to predict the next outcome x_i, and, as we have already observed $x_1, \ldots, x_{i-1}$, we can use these previous outcomes to guide our prediction. We assume that our predictions are probabilistic, i.e. they take the form of a probability distribution on $\mathcal{X}$, identified with its density or mass function f. We measure the loss of predicting with f when the actual outcome is x_i by $\textsc{loss}(x_i, f) := -\log f(x_i)$. This loss function arises naturally in data compression and gambling (Grünwald 2007), but, being the only so-called "local proper scoring rule" (Bernardo & Smith, 1994), it is also frequently used in Bayesian statistics, where it is known as the "logarithmic score".

A *sequential prediction strategy* S is a function that maps from the union of sample spaces $\cup_{n\geq 0}\ \mathcal{X}^n$ to the set $\mathcal{P}$ of distributions on $\mathcal{X}$ where the distributions are again identified with their densities or mass functions. That is,

$$S : \cup_{n\geq 0}\mathcal{X}^n \to \mathcal{P}.$$

In other words, a strategy S maps each possible sequence of arbitrary length (i.e., each element of $\cup_{n\geq 0}\ \mathcal{X}^n$) to a probabilistic prediction (i.e., an element of $\mathcal{P}$) for the next outcome. Thus $S(x_1, \ldots, x_{i-1}) = q$ means that somebody who uses strategy S will, upon observing sequence $x_1, \ldots, x_{i-1}$, predict the next observation using the distribution q. If we adopt the standard convention that X_i denotes the ith random variable and x_i denotes its observed outcome, then we would say that strategy S predicts that $X_i \mid x_1, \ldots, x_{i-1} \sim q$. When the actual outcome x_i is then observed, we would suffer the loss $\textsc{loss}(x_i, q) = -\log q(x_i)$. We define the loss of strategy S as the sum of its individual losses:

$$\textsc{loss}(x_1, \ldots, x_n, S) := \sum_{i=1}^{n} \textsc{loss}(x_i, S(x_1, \ldots, x_{i-1})).$$

In a seminal paper, Dawid (1984) called such strategies *prequential forecasting systems*; for this reason we also call the following interpretation of NML "prequential".

Let $\mathcal{M}$ be a statistical model, i.e. a family of distributions on $\mathcal{X}^n$. Each distribution $f(\cdot \mid \theta)$ in $\mathcal{M}$ can be used as a sequential prediction strategy S_θ in a straightforward fashion. To do so, we observe that $x_1, \ldots, x_i \in \mathcal{X}^i$ for all i, and so we can define

$$S_\theta(x_1, \ldots, x_i) := f(X_{i+1} = \cdot \mid x_1, \ldots, x_i, \theta).$$

That is, the $(i+1)$-st outcome is predicted using the conditional distribution for this outcome, given all past outcomes $x_1, \ldots, x_i$. If the model assumes that data are *i.i.d.*, then the parameter set θ produces the simple prediction strategy $S_\theta(x_1, \ldots, x_i) = f(\cdot \mid \theta)$, in which the predictions for each variable X_{i+1} are the same, irrespective of the previously observed outcomes. For simplicity, we will henceforth assume that this simplification holds for the model $\mathcal{M}$ under consideration.

Among all strategies S_θ corresponding to some $f(\cdot \mid \theta) \in \mathcal{M}$, the best predictor for any given full sequence $\mathbf{x} = (x_1, \ldots, x_n)$ is given by $S_{\hat{\theta}(\mathbf{x})}$, where $\hat{\theta}(\mathbf{x})$ is the maximum likelihood distribution for $\mathbf{x}$. To see this, note that for each S_θ, the loss incurred on $\mathbf{x}$ is

$$\sum_{i=1}^{n} -\log f(x_i \mid \theta) = -\log \prod_{i=1}^{n} f(x_i \mid \theta) = -\log f(\mathbf{x} \mid \theta),$$

so that the higher $f(\mathbf{x} \mid \theta)$, the smaller the loss. The loss is minimized for $\hat{\theta}(\mathbf{x})$, which is thus optimal among all $f(\cdot \mid \theta) \in \mathcal{M}$ *with hindsight*. In reality, we do not have hindsight:

we do not know $\hat{\theta}(\mathbf{x})$ until we have seen all x_i, so we cannot expect to predict, for all $\mathbf{x} \in \mathcal{X}^n$, as well as $\hat{\theta}(\mathbf{x})$. But we can design a prediction strategy which, for each $\mathbf{x}$, is *almost* as good as $\hat{\theta}(\mathbf{x})$, in the sense that the additional loss it incurs overs $\hat{\theta}(\mathbf{x})$ is as small as possible in the worst case over all $\mathbf{x}$. Thus, we look for a prediction strategy S such that

$$\max_{\mathbf{x} \in \mathcal{X}^n} \left[\textsc{loss}(\mathbf{x}, S) - \textsc{loss}(\mathbf{x}, S_{\hat{\theta}(\mathbf{x})}) \right] \tag{A.1}$$

is as small as possible. The expression between square brackets is called the *prediction regret* of strategy S relative to model $\mathcal{M}$. It is straightforward to show that, whenever $\mathcal{M}$ is such that a minimax optimal strategy minimizing (A.1) exists, then, for all i, all $x_1, \ldots, x_i \in \mathcal{X}^i$, its predictions $S(x_1, \ldots, x_i)$ coincide with $p^*(X_{i+1} = \cdot \mid x_1, \ldots, x_i)$ where p^* is the NML distribution (1). In other words, the NML distribution can be thought of as a sequential prediction strategy that achieves the minimax optimal regret under logarithmic score. We emphasize that, even if the data are *i.i.d.* according to each $f(\cdot \mid \theta)$, they are certainly not *i.i.d.* according to p^*: $p^*(X_{i+1} \mid x_1, \ldots, x_i)$ will strongly depend on $x_1, \ldots, x_i$, and will essentially behave like a smoothed version of the ML estimator $f(\cdot \mid \hat{\theta}(x_1, \ldots, x_i))$.

Thus, if we use NML to select between a finite number of models $\mathcal{M}_1, \ldots, \mathcal{M}_D$, we are effectively, for each $\mathcal{M}_d$, sequentially predicting $x_1, \ldots, x_n$ using the strategy that is optimal *relative* to $\mathcal{M}_d$, and in the end we select the model whose predictions yield the smallest total loss. Thus, we select the model that allows for the best possible sequential prediction of *unseen* data. As will be clear from the discussion in Section 3.1, this scheme is quite reminiscent of leave-one-out cross-validation with a logarithmic score. The precise relationship is discussed by Grünwald (2007, ch. 17).

Finally, we note that, just as there is an MDL approach to model selection, there also exist MDL methods for prediction and density estimation. One standard way to define such MDL predictions and estimates based on a sample $x_1, \ldots, x_i$ is in fact based on the "prequential" setup above. The distribution $f(\cdot \mid \theta) \in \mathcal{M}$ that is imagined to have generated the data is estimated as $p^*(X_{i+1} = \cdot \mid x_1, \ldots x_i)$, i.e. the conditional distribution of x_{i+1} according to the NML distribution p^*, defined relative to some $n \gg i$. As we have said before, in general $p^*(X_{i+1} \mid x_1, \ldots x_i)$ is *not* the maximum likelihood distribution $f(\cdot \mid \hat{\theta}(x_1, \ldots, x_i))$. It is a complicated distribution that can usually be very well approximated by the predictive distribution based on Jeffreys' prior (for large enough i, this predictive distribution is well-defined even if Jeffreys' prior is improper). The goal in MDL is to design an estimator that, when used for sequentially predicting outcomes, predicts nearly as well as the ML estimator for the final sample $x_1, \ldots, x_n$. This goal, however, is *not* achieved by predicting the individual x_{i+1} based on the ML estimator for $x_1, \ldots, x_i$. Therefore, MDL parameter estimation is, in general, quite different from ML parameter estimation.