

ACCEPTED VERSION

Peter Grünwald and Daniel J. Navarro

NML, Bayes and true distributions: A comment on Karabatsos and Walker (2006)
Journal of Mathematical Psychology, 2009; 53(1):43-51

Copyright ©2009 Published by Elsevier Inc.

[10.1016/j.jmp.2008.11.005](https://doi.org/10.1016/j.jmp.2008.11.005)

NOTICE: this is the author's version of a work that was accepted for publication in Journal of Mathematical Psychology. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Journal of Mathematical Psychology, 2009; 53(1):43-51, [10.1016/j.jmp.2008.11.005](https://doi.org/10.1016/j.jmp.2008.11.005)

Source:

<http://www.elsevier.com/journal-authors/policies/open-access-policies/article-posting-policy>

Elsevier's Policy: An author may use the preprint for personal use, internal institutional use and for permitted scholarly posting.

...

Elsevier's AAM Policy: Authors retain the right to use the accepted author manuscript for personal use, internal institutional use and for permitted scholarly posting provided that these are not for purposes of commercial use or systematic distribution.

Elsevier believes that individual authors should be able to distribute their AAMs for their personal voluntary needs and interests, e.g. posting to their websites or their institution's repository, e-mailing to colleagues. However, our policies differ regarding the systematic aggregation or distribution of AAMs to ensure the sustainability of the journals to which AAMs are submitted. Therefore, deposit in, or posting to, subject-oriented or centralized repositories (such as PubMed Central), or institutional repositories with systematic posting mandates is permitted only under specific agreements between Elsevier and the repository, agency or institution, and only consistent with the publisher's policies concerning such repositories. Voluntary posting of AAMs in the arXiv subject repository is permitted.

...

Permitted scholarly posting: Voluntary posting by an author on open websites operated by the author or the author's institution for scholarly purposes, as determined by the author, or (in connection with preprints) on preprint servers.

April 30, 2015

<http://hdl.handle.net/2440/60483>

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

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

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

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

9 The NML probability can be derived as the solution to a number of different optimality
10 problems (Shtarkov, 1987; Rissanen, 2001). It plays a prominent role in the minimum
11 description length (MDL) approach to statistical inference, originating from information
12 theory. However, the NML distribution has also been given an interpretation from other
13 statistical perspectives. Apart from the information-theoretic derivation, there are three
14 other standard derivations of the NML probability (see Grünwald 2007): the *prequential*
15 interpretation (briefly discussed in the appendix), a *differential-geometric interpretation*
16 (in which the denominator in (1) is interpreted as a volume; see, e.g., Balasubramanian
17 2005) and a *Bayesian interpretation* (which links (1) to Bayes factor model selection based
18 on a Jeffreys' prior). Importantly, the information-theoretic and prequential derivations
19 of NML do not rely on the assumption of a “true”, data-generating distribution. In this
20 sense, NML is an “agnostic” method, which suggests that it behaves robustly in situations
21 in which all models under consideration are wrong, yet some are useful.

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

33 **2 The Information-Theoretic View on NML**

34 The MDL principle states that we should prefer those models that allow us to compress
 35 the data set \mathbf{x} to the greatest possible extent. That is, if the codelength $L_C(\mathbf{x})$ denotes
 36 the number of bits required to describe \mathbf{x} using some code C , then we should prefer
 37 those models that allows us to produce short codelengths. We are able to talk about data
 38 compression using probabilistic language thanks to the Kraft inequality, which tells us
 39 that for any probability mass function f defined on a sample space \mathcal{X}^n , there exists a
 40 uniquely decodable code C such that, for all $\mathbf{y} \in \mathcal{X}^n$, the codelength is given by $L_C(\mathbf{y}) =$
 41 $-\log f(\mathbf{y})$. Vice versa, for any uniquely decodable code C , there exists a mass function f
 42 that satisfies this equality. This establishes a 1-to-1 correspondence between probability
 43 mass functions and uniquely decodable codes. Essentially the same correspondence holds,
 44 after appropriate discretization, if f is a density rather than a mass function.

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

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

51 where the minimum is over all distributions p that can be defined on \mathcal{X}^n , and the maximum
 52 is over all possible datasets $\mathbf{y} \in \mathcal{X}^n$. The expression in square brackets is called the *regret*:
 53 when applied to the actually-observed data \mathbf{x} , it is the additional number of bits one
 54 needs to code the data \mathbf{x} using (the code based on) p , compared to the code in \mathcal{M} that,
 55 with hindsight, turns out to minimize the codelength (maximize the probability) of \mathbf{x} .
 56 The latter code is invariably the code based on the ML (maximum likelihood) estimator
 57 $f(\cdot|\hat{\theta}(\mathbf{x}, \mathcal{M}))$. Thus, we seek, among all distributions (codes) p on \mathcal{X}^n , the one such that
 58 the worst-case regret is minimized. Regarding the more general question of why it makes
 59 sense to solve a minimax problem of this kind, the appendix contains a brief discussion;
 60 but the interested reader is referred to Grünwald (2007) for an extensive discussion. For
 61 the current purposes, it suffices to note that a key point in the specification of this minimax
 62 problem is that it does not matter what probability distribution generated the data \mathbf{x} ,
 63 or whether such a “true” distribution even exists: the NML distribution satisfies certain
 64 optimality criteria that depend only on the data. We elaborate this point in detail in the
 65 following section. Then, in Section 4–6, we discuss the KW derivation and our criticisms
 66 of it.

67 3 The Role of True Distributions

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

76 3.1 True Distributions in the Design of Algorithms

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

86 As an example of a procedure for which the design explicitly relies on some assumptions
87 about the true generating mechanism, consider the following simple problem. Suppose we
88 we want to choose between a model $\mathcal{M}_1 = \{f(\cdot | \mu) | \mu \in \mathbb{R}\}$ and its submodel $\mathcal{M}_0 =$
89 $\{f(\cdot | \mu) | \mu = 0\}$, where, for $\mathbf{x} \in \mathcal{X}^n$, $f(\mathbf{x} | \mu)$ is the standard normal density, extended
90 to n outcomes by independence. In the Neyman-Pearson approach to this problem, we
91 perform a hypothesis test with $\mu = 0$ as the null hypothesis, and $\mu \neq 0$ as the alternative.
92 Viewed as an algorithm, such a test takes data $\mathbf{x} \in \mathcal{X}^n$ as input, and it outputs “reject
93 \mathcal{M}_0 ,” or “accept \mathcal{M}_0 ,” possibly together with a p -value. For simplicity, we assume the
94 significance level is fixed at 0.01. This means that the test (algorithm) has been designed
95 such that the type-I error is at most 0.01: *if the data are sampled from \mathcal{M}_0 , the probability*
96 *of output “reject” is at most 0.01; moreover, among all algorithms with this property, we*
97 *use the one for which the type-II error is minimized.* Now, notice that the type-I error
98 is defined in terms of the probability of obtaining a particular kind of data set *if model*
99 *\mathcal{M}_0 is true.* Similarly, the type-II error describes the probability of obtaining a different
100 kind of data set *if (some element of) model \mathcal{M}_1 is true.* The design of the algorithm thus
101 crucially depends on the data being sampled either from \mathcal{M}_0 or \mathcal{M}_1 . As a consequence,
102 an awkward problem arises if the data are not sampled from either of the two models.
103 Under such circumstances, both the accept/reject decision and the corresponding p -value

104 have no clear interpretation any more, as they are probabilities of events according to
105 some distributions that we already know are not the data-generating distributions. This
106 situation is by no means uncommon: in practice, we often know in advance that all
107 models under consideration are, to some extent, wrong. Instead of trying to identify the
108 true model, in such a situation we may want to choose the model that, hopefully, is the
109 “best” in the sense that it leads to the best predictions about future data coming from
110 the same source. The Neyman-Pearson test has not been designed for such a situation,
111 and, as we have just seen, its outputs cannot easily be interpreted any more. In particular,
112 even though we put our significance level at 0.01, we certainly cannot claim anymore that,
113 by following the procedure repeatedly in a variety of contexts, only once in about a 100
114 times will we encounter the situation that we reject \mathcal{M}_0 even though it leads to better
115 predictions than \mathcal{M}_1 .

116 The example suggests that if none of our models are perfect – as is usually the case –
117 then we should use statistical algorithms whose output is a function *only* of how well
118 the actually observed sequence of data can be *predicted* based on the given models. To
119 make this precise, we need to define what it means to “predict based on a given model.”
120 This can be done in various ways. Let us consider two examples: leave-one-out cross-
121 validation (LOOCV; see Browne 2000), an approach to model selection that is popular in
122 the machine learning community; and NML. In LOOCV, for all outcomes x_i , one predicts
123 x_i on the basis of the maximum likelihood (ML) estimator $\hat{\theta}(\mathbf{x} \setminus x_i)$, i.e. based on all
124 observed data except x_i itself. The quality of predicting x_i with density or mass function
125 f_θ is measured in terms of the log loss, defined as $\text{LOSS}(x_i, f) := -\log f(x_i)$: the smaller
126 the loss, the better the prediction. According to LOOCV, we should select the model \mathcal{M}_j
127 which minimizes the sum of all prediction errors, $\sum_{i=1}^n \text{LOSS}(x_i, f(\cdot \mid \hat{\theta}(\mathbf{x} \setminus x_i, \mathcal{M}_j)))$. The
128 NML approach is based on the same loss function, but, as explained in the appendix,
129 rather than predicting by using the leave-one-out ML estimator, one sequentially predicts
130 the full sequence $\mathbf{x} = (x_1, \dots, x_n)$ using the prediction strategy that is worst-case optimal
131 relative to the element of \mathcal{M} that one should have used with hindsight, the worst-case
132 being taken over all possible data sequences.

133 Summarizing, we may broadly distinguish between *truth-dependent* approaches such as
134 Neyman-Pearson tests and AIC,¹ and *agnostic approaches* such as cross-validation and
135 NML. Truth-dependent approaches are designed to give good results with high probability
136 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 $\text{AIC}(\mathbf{x}, d) = -\log f(\mathbf{x} \mid \hat{\theta}(\mathbf{x}, \mathcal{M}_d)) + d$, 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 $\text{AIC}(\mathbf{x}, d)$ 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).

137 are only used as predictors, and the merit of a model in light of the data \mathbf{x} is solely
 138 determined by how well such distributions predict \mathbf{x} . It is in this sense that introductory
 139 papers (e.g., Myung et al. 2006) describe NML as being “free” from assumptions about
 140 true distribution: it is an agnostic method by design.

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

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

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

154 For truth-dependent approaches, the guarantee that the best model will eventually be
 155 selected can only be given for a small subset of the sequences satisfying (3), namely those
 156 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
 157 may be regarded as a “typical outcome” of f . In practice, however, we often have to deal
 158 with atypical outcomes: supposedly real-valued variables (e.g., normally distributed data)
 159 can very easily contain repeated values – cases where $x_i = x_j$ for some $i \neq j$ – due
 160 to round-off errors and other imperfections, an occurrence that should have probability
 161 0 (see, e.g., Grünwald, 2007, ch. 17). More generally, real-world data sets tend to be
 162 riddled with data missing not at random, data entry errors, and (particularly in the social
 163 sciences) a host of weak correlations (e.g., Meehl 1990). The net result is that, in many
 164 cases, even very large empirical data sets will have some characteristics that make them
 165 rather atypical sequences. It is also for this reason that the predictive guarantees for the
 166 agnostic approaches are in practice somewhat reassuring.

167 The previous remarks notwithstanding, it is worth pointing out that there is, of course, a
 168 weak spot in the agnostic approaches: one can measure prediction error in many different
 169 ways, so why should one focus on the log loss? The model that predicts best in terms of
 170 log loss may not be the best in terms of some other loss functions such as 0/1-loss. Indeed,
 171 there are approaches which try to extend MDL and related approaches beyond the log loss
 172 (Grünwald 2007, ch. 17); the methodology of *structural risk minimization* (Vapnik, 1998)
 173 may also be viewed in this manner. Nevertheless, there are certain properties of the log

174 loss which make it particularly attractive, such as the fact that it is the only local proper
175 scoring rule (Bernardo & Smith, 1994), that it has a clear interpretation in terms of data
176 compression and sequential gambling (Grünwald, 2007), and, as we discuss below, that it
177 has good convergence properties in the hypothetical case in which the true distribution
178 does reside in one of the models after all.

179 3.2 True Distributions in the Analysis of Algorithms

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

193 **1. Mechanism satisfies Equation 3.** Suppose we are to choose between two possible
194 models \mathcal{M}_0 and \mathcal{M}_1 , and that the constraining mechanism is such that Equation 3
195 holds, either for $a = 0, b = 1$ or vice versa. This may be one of the weakest assumptions
196 under which some form of inductive inference is possible at all. In this case, NML,
197 the Bayes factor method, BIC, LOOCV and AIC will all select the best-predicting
198 model \mathcal{M}_a for all large enough samples. In such cases, for large n , the truth-dependent
199 component of $\text{AIC}(\mathbf{x}, d)$ becomes negligible compared to its agnostic component. If,
200 however, we assume that \mathcal{M}_0 is nested into \mathcal{M}_1 , and Equation 3 holds with equality,
201 then NML, BIC and the Bayes factor method will select \mathcal{M}_0 for large n (a form of
202 Occam’s razor), whereas for many sequences, AIC and LOOCV will not. Grünwald
203 (2007) argues extensively why such a version of Occam’s razor is desirable. Note that
204 all this holds quite irrespective of whether the “true” data generating mechanism is
205 in any of the models, or is even a probability distribution; it may just as well be
206 deterministic.

207 If we allow the list of models to contain an arbitrary but finite number of elements,
208 then the same story still holds. However, in practice, this list is often countably infinite,
209 or (equivalently, as it turns out), it is allowed to grow with n . The prototypical exam-
210 ple is linear regression with polynomials, where the outcomes are pairs (Z, X) , with,
211 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
212 $(\alpha_0, \dots, \alpha_{d-1})$ is a parameter vector and U is normally distributed noise with mean 0.

213 We would like to learn the best polynomial model of the data, without assuming any
 214 a priori bound on the degree d . In such cases, there can be data sequences for which a
 215 particular degree d_0 leads, asymptotically, to the best predictions, yet, no matter how
 216 many data are observed, none of the methods will select degree d_0 , not even the agnostic
 ones.

217

218 **2. True distribution in one of the models.** At the other extreme, suppose we have
 219 the collection of models $\mathcal{M}_1, \mathcal{M}_2, \dots$, where the k -th model has k free parameters.
 220 Moreover, the data are sampled from a distribution $f(\cdot \mid \theta, \mathcal{M}_d)$ that falls inside the
 221 d -th model. In this situation, NML and other MDL-related methods, as well as BIC and
 222 the Bayes factor method, perform very well in the sense that, for all d such that \mathcal{M}_d
 223 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
 224 “ \mathcal{M}_d ” for all large n . For an explanation of the “almost”, see Grünwald (2007). AIC and
 225 leave-one-out cross-validation do not share this property of statistical consistency, and
 226 may, with positive probability, output a model of larger dimension than the minimal
 227 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.

228

229 **3. True distribution in model closure.** A commonly studied situation in statistics is
 230 to assume that the list $\mathcal{M}_1, \mathcal{M}_2, \dots$ is countable, and that data are sampled from some
 231 distribution p , which is not in any of the models of the list, but which can be arbitrarily
 232 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
 233 some suitably chosen distance measure for probability distributions. In our polynomial
 234 example, this would correspond to the true p stating that $X = g(Z) + U$, where g is a
 235 continuous function on $[-1, 1]$ that is, however, not itself a polynomial. In such cases,
 236 the best predictions can be obtained by choosing a small model at small sample sizes,
 237 and gradually choosing more complex models (higher-order polynomials) as the sample
 238 size increases. Qualitatively speaking, Bayes factor, BIC, AIC, NML and LOOCV all
 239 behave in this manner. But a more detailed view reveals important differences: if the
 240 models $\mathcal{M}_1, \mathcal{M}_2, \dots$ are sufficiently regular, and the distribution p is sufficiently smooth,
 241 then AIC and LOOCV will converge faster than NML, BIC and Bayes. More precisely,
 242 suppose we fix a method and for each n , we use it to infer a model and then predict
 243 future data based on that model. For all methods, the expected prediction loss will get
 244 smaller as n increases, and it will converge to the same asymptotic optimum. However,
 245 the convergence is slower (by a logarithmic factor) for Bayes, BIC and NML. On the
 246 other hand, if either (a) the models $\mathcal{M}_1, \mathcal{M}_2, \dots$ are not “regular”, or, (b), if the true
 247 p is not smooth, then AIC may fail dramatically, whereas Bayes factor, LOOCV and
 248 NML will still tend to converge. A common example of (a) is model selection for feature
 249 selection models, in which the number of considered models with d degrees of freedom
 250 is exponential in d (Yang 1999). An example of (b) within the polynomial setting arises
 251 if the function g is discontinuous, or if it tends to $\pm\infty$ at the boundaries of its domain.
 252 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).

253
 254 **4. True distribution not in model closure.** Finally, consider the possibility that there
 255 exists a true distribution p that cannot be arbitrarily well-approximated by members
 256 of models $\mathcal{M}_1, \mathcal{M}_2, \dots$, while nevertheless, some model \mathcal{M}_a contains a useful f that is
 257 “close” to p in that it tends to predict data reasonably well. This case is related to but
 258 less general than scenario 1 above, and essentially the same facts hold. To illustrate,
 259 suppose for simplicity that the data are i.i.d. according to both the ‘true’ p and all f
 260 in all of the $\mathcal{M}_1, \mathcal{M}_2, \dots$ under consideration, and suppose that one of the models is
 261 “best” in the sense that the following analogue of (3) holds: for some model \mathcal{M}_a on the
 262 list,

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

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

280 The upshot is that even agnostic methods may not always work well in all relevant set-
 281 tings. Nevertheless, we may still expect agnostic methods to be more robust than truth-
 282 dependent methods. Moreover, *if* a method that performs well in all settings 1–4 will
 283 ever be found, it is sure to be a method of the distribution-free kind. As an aside, Van
 284 Erven, Grünwald & De Rooij (2007) present an agnostic approach that combines the best
 285 of NML and LOOCV, and is probably the first known method that provably performs
 286 well in all cases discussed under settings 2 and 3 above; yet it still fails with countably
 287 infinite lists in settings 1 and 4.

288 To summarize, in this section we have aimed to give a general overview of the role played
 289 by the concept of a “true distribution” for a variety of different model selection algorithms.

290 We have done so in part because we think it provides a useful expansion of the necessarily-
 291 oversimplified treatment given in tutorial papers (e.g., Myung et al. 2006). However, it
 292 also provides an appropriate foundation for our discussion of the claims made recently by
 293 KW. It is to this topic that we now turn.

294 4 A Bayesian Decision-Theoretic View on NML

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

- 299 (1) Data arise from some unknown distribution (i.e., $\mathbf{x} \sim G$), and we have a prior over
 300 this distribution described by a Dirichlet process (DP; see Ferguson, 1973) with
 301 concentration parameter $c \rightarrow 0$ (i.e., $G \sim \text{DP}(G_0, 0)$).
- 302 (2) We want to select a parameter $\hat{\theta}$ that belongs to one of the models $\mathcal{M}_1, \dots, \mathcal{M}_D$, and
 303 in addition to the loss incurred due to the expected Kullback-Leibler discrepancy
 304 between $f(\cdot|\hat{\theta}, \mathcal{M})$ and the true distribution G , we suffer a “complexity penalty”
 305 $v(\mathcal{M}, n)$ that depends only on the model \mathcal{M} from which $\hat{\theta}$ is drawn and the sample
 306 size n .
- 307 (3) The complexity penalty $v(\mathcal{M}, n)$ is defined by

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

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

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

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

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

312 where p^* is given by (1), and the second equality follows because the logarithm is a mono-
 313 tonically increasing function. Hence, when assumptions (1)-(3) are met, the Bayes optimal
 314 model coincides with the model preferred under the NML criterion. In the following sec-
 315 tions we critically discuss this derivation and its supposed implications. In doing so, we

316 distinguish between two major problems (Section 5) and three minor concerns (Section 6).
317 We also briefly comment on another issue brought up by KW, namely the fact that for
318 many models, the NML is undefined (Section 7).

319 5 Major Problems

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

325 5.1 Incompleteness of the Characterization

326 In the context of discussing what conclusions can be drawn from their derivation, KW
327 (p. 520) state that they have “*discovered* the NML criterion using Bayesian decision the-
328 ory.” (emphasis added). This statement highlights one of the main problems we have
329 with their characterization, namely that it does not provide any Bayesian interpretation,
330 characterization or explanation of the complexity term (5). Rather, they show that any
331 model selection criterion of a “fit plus complexity” format is consistent with the Bayesian
332 framework, using assumptions (1) and (2) above. The specific application to NML via
333 assumption (3) is not explained anywhere in their paper – it is simply introduced on p.
334 519 with no justification given other than the statement that it is “[an] alternative penalty
335 term . . . for model simplicity”. They do not state *why* this particular penalty term would
336 be of interest to the statistician, even though it is clearly an essential component to NML.
337 After all, it is exactly this term that distinguishes the NML criterion from many other ex-
338 isting criteria such as AIC and BIC. In our view, this is not really a “discovery” at all, and
339 it makes it hard to see how their characterization is helpful or informative as to the nature
340 of NML itself. Indeed, the KW derivation can also be used to “discover” BIC and (as KW
341 in fact point out themselves) AIC — two criteria that behave very differently from NML
342 in many situations (see Section 3). This is achieved simply by replacing $v(\mathcal{M}_d, n)$ as in (5)
343 by $(k_d/2) \log n$ (which yields BIC) or k_d (producing AIC), where k_d is the dimensionality
344 of model \mathcal{M}_d . There is no particular reason given for the use of one penalty function
345 over any other one. This differs from all four previously existing interpretations of NML,
346 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.

347 it seems to us that the KW derivation is incomplete in a very fundamental sense, because
 348 it does not give any reason why a statistical decision-maker should adopt a complexity
 349 term that has the specific mathematical form specified in assumption (5).

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

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

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

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

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

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

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

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

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

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

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

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

428 6 Minor Issues

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

434 6.1 *Specification of the prior*

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

450 It is important to note that KW use the DP in the general sense, using the limiting DP
451 prior with $c \rightarrow 0$ to describe the prior belief about a *generic* unknown distribution G .
452 The result is that, as they note, they rely on a prior that concentrates with probability
453 1 on point-mass distributions. This reliance plays an important role in their subsequent
454 derivation: under this limiting prior, the Bayesian predictive distribution for future data
455 converges to the empirical distribution of \mathbf{x} (e.g., Ghosh & Ramamoorthi, 2003, Theorem
456 3.2.7). This in turn implies that the Bayesian maximum utility parameter estimate under
457 Kullback-Leibler loss is equivalent to the frequentist MLE $\hat{\theta}(\mathbf{x}, \mathcal{M})$ (as discussed by KW).
458 Obviously, this does not hold for other values of c , since in general the predictive distri-
459 bution under a DP prior is a weighted mixture of G_0 and the empirical distribution. In
460 short, although technically correct, the correspondence that they establish holds only for
461 this rather odd special case; a case that KW appear to have chosen primarily to ensure
462 that their Bayesian parameter estimation procedure mimics a frequentist one.³

463 6.2 *Specification of the decision problem*

464 A second issue relates to the manner in which KW specify the decision problem. The
465 Bayesian decision procedure described by KW equates the utility of a model with the
466 utility of its best parameter value. This manner of setting up the problem is highly biased
467 towards complex models, since in its simplest form it reduces to picking the model that
468 can provide the best fit in a maximum likelihood sense. In order to redress this, they
469 then introduce a complexity penalty into the utility function, as suggested by Kadane
470 and Dickey (1980). We note that such an approach, while certainly correct, is by no
471 means standard Bayesian practice. In a standard Bayesian textbook, Berger (1985, p.
472 284) argues that one of the advantages of the Bayesian approach is that it *automatically*
473 takes model complexity into account, without the need for any explicit penalties (Mackay
474 (2003) refers to this as the “Bayesian Occam’s razor”). Indeed, this does happen under
475 standard parametric priors and standard utility functions (possibly, but not necessarily
476 of the 0/1-type; see, e.g. Bernardo & Smith 1994, ch. 6). However, by associating model
477 utility with maximum likelihood parameter utility, KW are unable to take advantage of
478 one of the most useful features of Bayesian inference, and are forced to reintroduce it via
479 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.

481 A third point to make is that KW’s characterization actually discards a number of the
 482 existing parallels between MDL and Bayesian methods. As discussed, the NML criterion
 483 originally arose as a specific instance of the broader MDL approach to inductive inference,
 484 which is in fact closely related to Bayesian inference (see Grünwald 2007, p. 531-550). Just
 485 as in the Bayesian approach, MDL inference invariably starts by putting a distribution on
 486 observables. In the NML version of MDL discussed here, one actually puts a uniform prior
 487 $\pi(d) = 1/D$ on the model set $\{\mathcal{M}_1, \dots, \mathcal{M}_D\}$ (and indeed when one compares countably
 488 infinitely many models, the prior on the model index d becomes *essential* in the MDL
 489 approach; see Grünwald, 2007, p. 406 & p. 423). One then associates each model \mathcal{M}_d with
 490 a distribution on \mathcal{X}^n , in this case the NML distribution $p^*(\mathbf{x} \mid \mathcal{M}_d)$ given by (1). Thus,
 491 under the more typical Bayesian characterization of NML, the criterion may be interpreted
 492 as advocating a “maximum posterior model” \mathcal{M}_d under a uniform prior distribution on d .
 493 Accordingly, the NML criterion (as with other versions of MDL) already has a Bayesian
 494 flavor, with the NML distribution $p^*(\mathbf{x} \mid \mathcal{M}_d)$ playing a role similar to the Bayesian
 495 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
 496 w . In fact, although we do not do so here, it is not too difficult to construct cases in
 497 which the Bayesian marginal probability corresponds to the NML probability more or
 498 less exactly (this can be made to hold for any sample size n , with the usual convergence
 499 to a Jeffreys’ prior as $n \rightarrow \infty$). These relationships, however, are completely lost in the
 500 KW derivation, because KW decouple the two terms that comprise the NML criterion.
 501 In their approach, the numerator $f(\mathbf{x} \mid \hat{\theta}(\mathbf{x}, \mathcal{M}))$ arises as a consequence of the $DP(G_0, 0)$
 502 prior, while the denominator $\int_{\mathcal{X}^n} f(\mathbf{y} \mid \hat{\theta}(\mathbf{y}, \mathcal{M})) d\mathbf{y}$ is a consequence of the loss function.
 503 The fact that the denominator is in fact the integral of the maximized likelihood in the
 504 numerator (hence giving rise to the name *normalized* maximum likelihood, and making
 505 $p^*(\cdot \mid \mathcal{M}_d)$ a distribution over possible data sets) actually becomes irrelevant in KW’s
 506 approach.

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

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

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

529 7 Concluding Remarks

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

542 More generally, we suspect that it may be the case that some researchers (including
543 us) have at times overemphasized the importance of NML within the MDL framework,
544 perhaps giving the impression that the two are equivalent. Given this possibility, it is
545 important to note that the central idea in MDL is to base statistical inference on uni-
546 versal coding (see Grünwald 2007, for an extensive discussion): as it happens, the NML
547 method is only one of at least five good methods for constructing universal codes, so the
548 MDL framework is much broader than NML (only two of the 19 chapters in Grünwald's
549 (2007) book deal primarily with NML, for instance). That said, because it has certain
550 optimality properties which the other methods lack, NML has tended to be the preferred
551 method in recent years. However, in those cases where NML cannot be applied, the other
552 methods usually still can. Importantly, one of these five types of universal codes is based
553 on Bayesian marginal likelihoods, and so it should be no surprise that there is generally a
554 close correspondence between Bayesian and MDL methods. Nevertheless, this correspon-
555 dence is of quite a different type than the KW derivation suggests.

556 8 Acknowledgements

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

563 References

- 564 Akaike, H. (1973). Information theory and an extension of the maximum likelihood prin-
565 ciple. In B. Petrov & F. Csaki (Eds.) *Second International Symposium on Information*
566 *Theory* (pp. 267-281). Budapest: Akademiai Kiado
- 567 Balasubramanian, V. (2005). MDL, Bayesian inference and the geometry of the space of
568 probability distributions. In P. Grünwald, J. I. Myung, and M. A. Pitt (Eds.), *Advances*
569 *in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT Press.
- 570 Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, revised and ex-
571 panded 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.
- 572 Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. Chichester, UK: Wiley.
- 573 Browne (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108-
574 132.
- 575 Dawid, A.P. (1984). Present position and potential developments: some personal views,
576 statistical theory, the prequential approach. *Journal of the Royal Statistical Society,*
577 *Series A* 147(2), 278-292.
- 578 Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using
579 mixtures. *Journal of the American Statistical Association*, 90, 577-588.
- 580 Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of*
581 *Statistics*, 1, 209-230.
- 582 Ghosh, J. & Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. New York: Springer
- 583 Grünwald, P. (2005). Minimum description length tutorial. In P. Grünwald, J. I. Myung
584 & M. A. Pitt (Eds). *Advances in Minimum Description Length: Theory & Applications*.
585 Cambridge, MA: MIT Press. Note that Chapter 17, the most relevant chapter for this
586 article, is available freely on the web.
- 587 Grünwald, P. (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT
588 Press
- 589 Grünwald, P. & J. Langford (2007). Suboptimal behavior of Bayes and MDL in classi-
590 fication under misspecification. *Machine Learning* 66(2-3), pages 119-149. Cambridge,
591 MA: MIT Press
- 592 Van Erven, T., Grünwald, P. & de Rooij, S. (2008) Catching up faster in Bayesian model

593 selection and model averaging. *Advances in Neural Information Processing Systems*,
594 20.

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

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

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

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

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

606 Petrone, S. & Raftery, A.E. (1997). A note on the Dirichlet process prior in Bayesian
607 nonparametric inference with partial exchangeability. *Statistics and Probability Letters*,
608 36, 69-83.

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

611 Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6,
612 461-464.

613 Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4,
614 639-650.

615 Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of In-*
616 *formation Transmission* 23(3), 3-17.

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

618 Yang, Y. (1999). Model Selection for Nonparametric Regression. *Statistica Sinica* 9, 475-
619 499.

620 A The Prequential Interpretation of NML

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

631 A *sequential prediction strategy* S is a function that maps from the union of sample spaces
 632 $\cup_{n \geq 0} \mathcal{X}^n$ to the set \mathcal{P} of distributions on \mathcal{X} where the distributions are again identified
 633 with their densities or mass functions. That is,

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

634 In other words, a strategy S maps each possible sequence of arbitrary length (i.e., each
 635 element of $\cup_{n \geq 0} \mathcal{X}^n$) to a probabilistic prediction (i.e., an element of \mathcal{P}) for the next
 636 outcome. Thus $S(x_1, \dots, x_{i-1}) = q$ means that somebody who uses strategy S will, upon
 637 observing sequence x_1, \dots, x_{i-1} , predict the next observation using the distribution q . If we
 638 adopt the standard convention that X_i denotes the i th random variable and x_i denotes its
 639 observed outcome, then we would say that strategy S predicts that $X_i \mid x_1, \dots, x_{i-1} \sim q$.
 640 When the actual outcome x_i is then observed, we would suffer the loss $\text{LOSS}(x_i, q) =$
 641 $-\log q(x_i)$. We define the loss of strategy S as the sum of its individual losses:

$$\text{LOSS}(x_1, \dots, x_n, S) := \sum_{i=1}^n \text{LOSS}(x_i, S(x_1, \dots, x_{i-1})).$$

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

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

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

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

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

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

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

658 we do not know $\hat{\theta}(\mathbf{x})$ until we have seen all x_i , so we cannot expect to predict, for all
659 $\mathbf{x} \in \mathcal{X}^n$, as well as $\hat{\theta}(\mathbf{x})$. But we can design a prediction strategy which, for each \mathbf{x} , is
660 *almost* as good as $\hat{\theta}(\mathbf{x})$, in the sense that the additional loss it incurs over $\hat{\theta}(\mathbf{x})$ is as
661 small as possible in the worst case over all \mathbf{x} . Thus, we look for a prediction strategy S
662 such that

$$\max_{\mathbf{x} \in \mathcal{X}^n} [\text{LOSS}(\mathbf{x}, S) - \text{LOSS}(\mathbf{x}, S_{\hat{\theta}(\mathbf{x})})] \quad (\text{A.1})$$

663 is as small as possible. The expression between square brackets is called the *prediction*
664 *regret* of strategy S relative to model \mathcal{M} . It is straightforward to show that, whenever
665 \mathcal{M} is such that a minimax optimal strategy minimizing (A.1) exists, then, for all i , all
666 $x_1, \dots, x_i \in \mathcal{X}^i$, its predictions $S(x_1, \dots, x_i)$ coincide with $p^*(X_{i+1} = \cdot \mid x_1, \dots, x_i)$ where
667 p^* is the NML distribution (1). In other words, the NML distribution can be thought
668 of as a sequential prediction strategy that achieves the minimax optimal regret under
669 logarithmic score. We emphasize that, even if the data are *i.i.d.* according to each $f(\cdot \mid \theta)$,
670 they are certainly not *i.i.d.* according to p^* : $p^*(X_{i+1} \mid x_1, \dots, x_i)$ will strongly depend
671 on x_1, \dots, x_i , and will essentially behave like a smoothed version of the ML estimator
672 $f(\cdot \mid \hat{\theta}(x_1, \dots, x_i))$.

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

680 Finally, we note that, just as there is an MDL approach to model selection, there also
681 exist MDL methods for prediction and density estimation. One standard way to define
682 such MDL predictions and estimates based on a sample x_1, \dots, x_i is in fact based on
683 the “prequential” setup above. The distribution $f(\cdot \mid \theta) \in \mathcal{M}$ that is imagined to have
684 generated the data is estimated as $p^*(X_{i+1} = \cdot \mid x_1, \dots, x_i)$, i.e. the conditional distribution
685 of x_{i+1} according to the NML distribution p^* , defined relative to some $n \gg i$. As we have
686 said before, in general $p^*(X_{i+1} \mid x_1, \dots, x_i)$ is *not* the maximum likelihood distribution $f(\cdot \mid$
687 $\hat{\theta}(x_1, \dots, x_i))$. It is a complicated distribution that can usually be very well approximated
688 by the predictive distribution based on Jeffreys’ prior (for large enough i , this predictive
689 distribution is well-defined even if Jeffreys’ prior is improper). The goal in MDL is to design
690 an estimator that, when used for sequentially predicting outcomes, predicts nearly as well
691 as the ML estimator for the final sample x_1, \dots, x_n . This goal, however, is *not* achieved by
692 predicting the individual x_{i+1} based on the ML estimator for x_1, \dots, x_i . Therefore, MDL
693 parameter estimation is, in general, quite different from ML parameter estimation.