



On the normalized maximum likelihood and Bayesian decision theory[☆]

George Karabatsos^{a,*}, Stephen G. Walker^b

^aUniversity of Illinois-Chicago, 1040 W. Harrison Street (MC 147), Chicago, IL 60607, USA

^bUniversity of Kent, Institute of Mathematics, Statistics and Actuarial Science, Canterbury CT2 7NZ, UK

Received 13 October 2005; received in revised form 10 July 2006

Abstract

Under the principle of minimum description length, the optimal predictive model maximizes the normalized maximum likelihood (NML). While the Bayesian approach to model selection aims to identify the model that best describes the (unknown) true distribution that generated a set of data, the NML approach to model selection makes no reference to a true distribution, and this is seen as a significant advantage of the latter approach. In contrast, this article shows that, for a specific choice of utility function, the NML approach is equivalent to a Bayesian model selection under the Bayesian bootstrap and with a specific penalty function for model complexity. This new characterization uncovers some statistical issues about the NML approach.

© 2006 Published by Elsevier Inc.

Keywords: Bayesian non-parametrics; Normalized maximum likelihood; Model selection

Under the principle of minimum description length (MDL, Rissanen, 2001), given a set of models $\mathcal{M} = \{M_d\}_1^D$ and available data $\mathbf{x} = \{x_i \in \mathcal{X} \subseteq \mathfrak{R}^q\}_{i=1}^n$, the optimal decision is to select the model $\hat{M} \in \mathcal{M}$ that maximizes the normalized maximum likelihood (NML), according to:

$$\begin{aligned} \hat{M} &= \arg \sup_{M_d \in \mathcal{M}} \left\{ p_d^*(\mathbf{x}) = \frac{f_d(\mathbf{x}; \hat{\theta}_{x,d})}{\int f_d(\mathbf{y}; \hat{\theta}_{y,d}) d\mathbf{y}} \right\} \\ &= \arg \sup_{M_d \in \mathcal{M}} \left\{ \log p_d^*(\mathbf{x}) = \sum_{i=1}^n \log f_d(x_i; \hat{\theta}_{x,d}) \right. \\ &\quad \left. - \log \int f_d(\mathbf{y}; \hat{\theta}_{y,d}) d\mathbf{y} \right\}, \end{aligned}$$

where for every model $M_d \in \mathcal{M}$, the NML $p_d^*(\mathbf{x})$ depends on the predictive distribution $f_d(\mathbf{x}; \hat{\theta}_{x,d})$ conditioned on the maximum likelihood estimate $\hat{\theta}_{x,d}$, which is normalized by the integral $\int f_d(\mathbf{y}; \hat{\theta}_{y,d}) d\mathbf{y}$ of such predictive distributions

[☆]This research is supported by National Science Foundation Grant SES-0242030 in the program of Methodology, Measurement, and Statistics. The second author is financed by an EPSRC Advanced Research Fellowship.

*Corresponding author. Fax: +1 312 996 5651.

E-mail addresses: georgek@uic.edu (G. Karabatsos), S.G.Walker@kent.ac.uk (S.G. Walker).

over all possible data sets $\{\mathbf{y}\}$. This integral is a model penalty that increases with the “flexibility” of $f_d(\cdot; \theta_d)$, and thus the NML approach rewards model simplicity. In other words, the NML identifies the optimal model $\hat{M} \in \mathcal{M}$ as the one that “allows the greatest compression of the data” (Myung, Navarro, & Pitt, 2006). It has been argued that one significant advantage of the NML approach is that it does not assume the existence of a true distribution function (see Grünwald, 2005a, Section 1.5), and Myung et al. (2006) go so far as to state: “We are . . . ill-advised to base our inferential procedures on an unjustifiable faith in an unknown truth.”

In this paper, we adopt a Bayesian approach to model selection which will be based on a formal decision theoretic set-up. Therefore, it is worth outlining the Bayesian decision theoretic approach to model selection, recently discussed by Gutierrez-Peña and Walker (2005), Karabatsos and Walker (2005), and Karabatsos (2006). We will show how the NML approach to model selection arises as a special case of Bayesian model selection, for a specific choice of utility function and a specific choice of prior distribution.

It is known that Bayes theorem is a mathematical consequence of the axioms of quantitative coherence, and provides the only coherent approach to statistical infer-

ence, in which the aim is to find the optimal decision that maximizes expected utility (Bernardo & Smith, 1994). If the decision-maker (statistician) defines all the “possible states of the world” as the set of all sampling distribution functions $\Omega_{\mathcal{X}} = \{F\}_{\mathcal{X}}$ (defined on some sample space \mathcal{X}), then according to Bayes Theorem, s/he must specify a prior distribution Π over $\Omega_{\mathcal{X}}$. A set of data \mathbf{x} updates this prior distribution Π to the posterior distribution Π_n , such that for every $F \in \Omega_{\mathcal{X}}$, the posterior probability distribution $\Pi_n(dF)$ represents the degree of belief that F is the true sampling distribution. The decision-maker specifies a utility function $u(a; F)$ over $\mathcal{D} \times \Omega_{\mathcal{X}}$, that measures the desirability of decision $a \in \mathcal{D}$ given each possible true “state of the world” $F \in \Omega_{\mathcal{X}}$.

The first essential component is the prior distribution $\Pi(dF)$, defined on a space of distribution functions. For the purposes of the paper, our choice of prior distribution will be the so-called Dirichlet process prior (Ferguson, 1973). Here we discuss in more detail the Dirichlet process prior. The first point to make is that this prior has full (weak) support, that is, positive mass is put on all weak neighbourhoods of all distribution functions which are absolutely continuous with respect to the mean distribution of the Dirichlet process prior. To expand on this, if d_w is a metric which is equivalent to convergence in distribution; that is, for distribution functions P_n and P ,

$$d_w(P, P_n) \rightarrow 0 \iff \int \phi(x) dP_n(x) \rightarrow \int \phi(x) dP(x)$$

for all continuous and bounded ϕ , then the prior is said to have full (weak) support whenever

$$\Pi\{P : d_w(P, Q) < \varepsilon\} > 0$$

for all $\varepsilon > 0$ and for all Q absolutely continuous with respect to the mean distribution function of the Dirichlet process prior. Hence, if this is correct terminology for a Bayesian, the prior is not “wrong”. That is, the prior gives support to all possible distributions that may generate a given set of data \mathbf{x} .

The prior model generates random distribution functions. Essentially, a random path (stochastic process) is generated which behaves as a distribution function. That is, it starts at zero and moves to one in a non-decreasing way. It is possible to sample a Dirichlet process via the strategy of taking $\{\theta_i\}_{i=1}^{\infty}$ to be independent and identically distributed from some fixed distribution G and $\{v_i\}_{i=1}^{\infty}$ to be independent and identically distributed from beta(1, c) for some $c > 0$. Then

$$F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j},$$

where $w_1 = v_1$ and for $j > 1$,

$$w_j = v_j \prod_{l=1}^{j-1} (1 - v_l).$$

It is straightforward to show that the sum of the w_j 's is one.

It is that $E(F) = G$ and for suitable sets B ,

$$\text{Var}\{F(B)\} = \frac{G(B)\{1 - G(B)\}}{c + 1}. \quad 59$$

Using the Dirichlet process itself for modelling independent and identically distributed observations, say $\{x_1, \dots, x_n\}$, can be done and the posterior is also a Dirichlet process with updated parameters $c \rightarrow c + n$ and

$$G \rightarrow F_n = \frac{cG + nG_n}{c + n}, \quad 61$$

where G_n is the empirical distribution function of $\{x_1, \dots, x_n\}$. Hence, the Bayes estimate, or predictive distribution F_n , is a nice mixture of the prior choice and the empirical distribution. The parameter c has been viewed as a “prior sample size”. According to this view, a “non-informative” Dirichlet process prior is defined by taking the limit $c \rightarrow 0$. The posterior under this limit is a Dirichlet process with expectation the probability measure which assigns probability $1/n$ to each observation $x_i \in \mathbf{x}$. In other words, the posterior mean is the empirical distribution function G_n , which provides the basis for the Bayesian bootstrap (Rubin, 1981). Under this scenario, the predictive distribution function is the empirical distribution function.

The other key component for the coherent approach to decision making is the utility function defined on the joint space of actions and distribution functions. We will denote this by $u(\theta; F)$, with $\theta \in \Theta$. This assumes that the decision space is a parameter space Θ , which indexes a parametric family of densities $f(x; \theta)$. This makes sense when the decision is to select a parameter from a parametric family of densities for estimation purposes. The idea is that $u(\theta; F)$ rewards those θ which make $f(x; \theta)$ close to the true distribution function F . This is the only way to understand the construction of $u(\theta; F)$ and so it is clear that one must be thinking about the notion of a true distribution function. For this reason one wants a prior which is supported by all distribution functions and the Dirichlet process prior does precisely this.

Following the theory, the optimal choice, based on the principle of the maximization of expected utility, is then given by

$$\arg \sup_{\theta \in \Theta} \int u(\theta; F) \Pi_n(dF), \quad 101$$

where Π_n is the posterior distribution obtained with the prior Π and the data $\mathbf{x} = \{x_1, \dots, x_n\}$. We will denote the maximizer (optimal value) by $\hat{\theta}$. This is standard Bayesian theory, see for example Bernardo and Smith (1994, Chapter 2).

We now extend the ideas above to the situation which covers D possible parametric models, with corresponding parametric families $f_d(x; \theta_d)$ and $\theta_d \in \Theta_d$. We will label the model by M_d , for $d = 1, \dots, D$. Extending the ideas in an obvious way, one can choose the optimal model by selecting the model M_d which maximizes

$$\bar{u}(M_d|\mathbf{x}) = \int u(\hat{\theta}_d; F) \Pi_n(dF).$$

This is simply an obvious extension of the parameter estimation choice. One way of looking at this is to extend the parameter space to

$$\Theta = \bigcup_{d=1}^D \Theta_d$$

and choose the model which provides the $\theta \in \Theta$ which maximizes the posterior expected utility.

We now discuss the choice of utility function $u(\theta; F)$. As has been mentioned earlier, those $\theta \in \Theta$ which make $f(x; \theta)$ close to F receive a high utility. It is natural to measure closeness through a distance between probability distribution functions and a popular choice is the Kullback–Leibler distance, or divergence as it is more usually known. We also seek a penalty term which penalizes model complexity. Based on the Kullback–Leibler divergence, having removed terms that do not influence the maximization of expected utility, we can define the utility function as provided by the logarithmic scoring rule, having the form

$$u(\theta; F) = A \int \log f(x; \theta) dF(x) + B,$$

where $A > 0$ and B do not depend on θ or F (see Bernardo & Smith, 1994, Chapter 2; Bernardo, 1979). Hence, we can write

$$\bar{u}(M_d|\mathbf{x}) = \int \left\{ \int \log f_d(x; \hat{\theta}_d) dF(x) \right\} \times \Pi_n(dF) - v(d, n),$$

where $v(d, n)$ is an as yet unspecified function of the model complexity associated with M_d , the sample size n , and $\hat{\theta}_d$ maximizes over Θ_d the expression

$$\int \left\{ \int \log f_d(x; \theta_d) dF(x) \right\} \Pi_n(dF).$$

Here we have put $A = 1$ and $B = -v(d, n)$. Now

$$\begin{aligned} & \int \left\{ \int \log f_d(x; \theta_d) dF(x) \right\} \Pi_n(dF) \\ &= \int \log f_d(x; \theta_d) dF_n(x), \end{aligned}$$

where F_n is the predictive distribution function given by

$$F_n = \int F \Pi_n(dF).$$

In the case of the “non-informative” Dirichlet process prior Π over $\Omega_{\mathcal{X}}$, F_n becomes the empirical distribution function and so

$$\int \log f_d(x; \theta_d) dF_n(x) = \frac{1}{n} \sum_{i=1}^n \log f_d(x_i; \theta_d).$$

Thus in the Bayesian approach, under a “non-informative” Dirichlet process prior, the optimal decision $\hat{\theta}_d \in \Theta_d$ corresponds to the maximum likelihood estimator.

We now consider the penalty term for model dimension/complexity. The most usual penalty term is $v(d, n) = m(d)$, where $m(d)$ denotes the dimension of model M_d . This gives the Akaike criterion (Akaike, 1973). An alternative penalty term is based on the reward

$$R_d = -\frac{1}{n} \log \int f_d(\mathbf{y}; \hat{\theta}_{\mathbf{y},d}) d\mathbf{y}$$

for model simplicity.¹ That is

$$v(d, n) = \frac{1}{n} \log \int f_d(\mathbf{y}; \hat{\theta}_{\mathbf{y},d}) d\mathbf{y}. \quad (1)$$

Under this penalty term, and assuming the non-informative Dirichlet process prior Π , the optimal choice \hat{M} of model that maximizes the posterior expected utility, satisfies

$$\hat{M} = \arg \sup_{d \in \{1, \dots, D\}} \bar{u}(M_d|\mathbf{x}),$$

where

$$\begin{aligned} \bar{u}(M_d|\mathbf{x}) &= \frac{1}{n} \sum_{i=1}^n \left\{ \log f_d(x_i; \hat{\theta}_{x,d}) - \frac{1}{n} \log \int f_d(\mathbf{y}; \hat{\theta}_{\mathbf{y},d}) d\mathbf{y} \right\} \\ &= \frac{1}{n} \log \left\{ \frac{f_d(\mathbf{x}; \hat{\theta}_{\mathbf{x},d})}{\int f_d(\mathbf{y}; \hat{\theta}_{\mathbf{y},d}) d\mathbf{y}} \right\}. \end{aligned}$$

Thus, under the Bayesian bootstrap, which effectively uses the empirical distribution function as the predictive distribution, the Kullback–Leibler divergence as a measure between distribution functions and the $v(d, n)$ penalty term for dimension given in (1), NML and Bayesian decision theory coincide. It has been mentioned that under exponential family assumptions (or when one of the models under consideration is true), that NML corresponds to the Bayes factor approach to model selection with Jeffrey’s prior. In fact, this is only an approximate representation, which apparently becomes more accurate as the sample size increases. Additionally, it carries through when the Jeffrey’s prior is proper; however, this prior is invariably improper. Finally, on this point, 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.

We now comment on the two components for achieving NML under a Bayesian decision theoretic approach. The first is to do with the prior and the second the penalty term.

The “non-informative” Dirichlet process prior is intrinsic to the NML approach, yet this prior is actually a highly informative prior that assigns probability 1 to the strict subset $\mathcal{G} \subseteq \Omega_{\mathcal{X}}$ of degenerate sampling distributions (Ghosh & Ramamoorthi, 2003, Theorem 3.2.6). In fact,

¹See Kadane and Dickey (1980) for a general discussion about assigning rewards for model simplicity, within a posterior expected utility framework.

1 this prior generates random distribution functions consist- 45
 2 ing of a single atom, the location of which is coming from
 3 G . However, this said, the predictive is the empirical 47
 4 distribution function, which can be regarded as “objec-
 5 tive”. However, this finding contrasts with the view that 49
 6 NML corresponds to objective statistical inference under a
 7 Jeffreys prior (Grünwald, 2005a). Our contention is that it
 8 arises under the non-parametric “non-informative” Dirich-
 9 let process prior. Additionally, the idea that NML makes
 10 no mention of a true distribution is a meaningless point.
 11 We have discovered the NML criterion using Bayesian
 12 decision theory and have, as a component of this
 13 procedure, explicitly introduced the notion of a true
 14 distribution function in the construction of the utility
 15 function $u(\theta; F)$.

16 We now comment on the penalty term for model
 17 complexity. The first point to make is that, in our view,
 18 the $v(d, n)$ in (1) is difficult to understand as a penalty term.
 19 It does increase with model complexity, but to what extent
 20 is incalculable. Another point is that this choice of $v(d, n)$ is
 21 not invariant over the (usually subjective) choice of sample
 22 space. However, our specific charge is that $v(d, n)$ is too
 23 large, even infinite, for reasonable models, including the
 24 normal model.

25 **Example.** Here we consider a simple normal model of
 26 dimension 1, and so we have

$$27 f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\{-(x - \theta)^2/2\}.$$

28 Given a sample of size two, say x_1 and x_2 , it is well known
 29 that the maximum likelihood estimator for the model is
 30 given by $\hat{\theta} = (x_1 + x_2)/2$. Consequently,

$$31 v(1, 2) = \frac{1}{2} \log \left\{ \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\{-(y_1 - \bar{y})^2/2 \right.$$

$$32 \left. -(y_2 - \bar{y})^2/2\} dy_1 dy_2 \right\},$$

33 where $\bar{y} = (y_1 + y_2)/2$. Hence,

$$34 v(1, 2) = \frac{1}{2} \log \left\{ \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp\{-(y_1 - y_2)^2/4\} dy_1 dy_2 \right\}$$

43

and it is easy to see that this leads to $v(1, 2) = +\infty$, and as a
 consequence the normal model receives a posterior
 expected utility of $\bar{u}(M_d|x) = -\infty$. A simple model, yet a
 disaster for NML. This is true especially in light of the fact
 that the NML approach intends to encourage the selection
 of simple models. There have been ad hoc methods
 proposed to solve this particular problem (e.g., restrict
 the parameter space), however, they are not fully satisfac-
 tory (e.g., Grünwald, 2005b, p. 70).

References

- 35 Akaike, H. (1973). Information theory and the an extension of the
 36 maximum likelihood principle. In B. Petrov, & F. Csaki (Eds.), *Second
 37 international symposium on information theory*. Budapest: Academiai
 38 Kiado.
- 39 Bernardo, J. (1979). Expected information as expected utility. *Annals of
 40 Statistics*, 7, 686–690.
- 41 Bernardo, J., & Smith, A. (1994). *Bayesian theory*. Chichester, UK: Wiley.
- 42 Ferguson, T. (1973). A Bayesian analysis of some nonparametric
 43 problems. *Annals of Statistics*, 1, 209–230.
- 44 Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. New
 45 York: Springer.
- 46 Grünwald, P. (2005a). Introducing MDL. In P. Grünwald, J. Myung, &
 47 M. Pitt (Eds.), *Advances in minimum description length: Theory and
 48 applications*. Cambridge, MA: MIT Press.
- 49 Grünwald, P. (2005b). Tutorial on MDL. In P. Grünwald, J. Myung, &
 50 M. Pitt (Eds.), *Advances in minimum description length: Theory and
 51 applications*. Cambridge, MA: MIT Press.
- 52 Gutierrez-Peña, E., & Walker, S. (2005). Statistical decision problems and
 53 Bayesian nonparametric methods. *International Statistical Review*, 73,
 54 309–330.
- 55 Kadane, J., & Dickey, J. (1980). Bayesian decision theory and the
 56 simplification of models. In J. Kmenta, & J. Ramsey (Eds.), *Evaluation
 57 of econometric models*. New York: Academic Press.
- 58 Karabatsos, G. (2006). Bayesian nonparametric model selection and
 59 model testing. *Journal of Mathematical Psychology*, 50, 123–148.
- 60 Karabatsos, G., & Walker, S. (2005). Solving incoherence in model
 61 selection with Bayesian nonparametrics. Under review.
- 62 Myung, J., Navarro, D., & Pitt, M. (2006). Model selection by normalized
 63 maximum likelihood. *Journal of Mathematical Psychology*, 50,
 64 167–179.
- 65 Rissanen, J. (2001). Strong optimality of the normalized ML models as
 66 universal codes and information in data. *IEEE Transactions on
 67 Information Theory*, 47, 1712–1717.
- 68 Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9,
 69 130–134.