Biometrika (2020), **107**, 2, *pp.* 489–496 doi: 10.1093/biomet/asz077
Printed in Great Britain Advance Access publication 24 January 2020

On the marginal likelihood and cross-validation

By E. FONG and C. C. HOLMES

Department of Statistics, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, U.K.
edwin.fong@stats.ox.ac.uk cholmes@stats.ox.ac.uk

Summary

In Bayesian statistics, the marginal likelihood, also known as the evidence, is used to evaluate model fit as it quantifies the joint probability of the data under the prior. In contrast, non-Bayesian models are typically compared using cross-validation on held-out data, either through k-fold partitioning or leave-p-out subsampling. We show that the marginal likelihood is formally equivalent to exhaustive leave-p-out cross-validation averaged over all values of p and all held-out test sets when using the log posterior predictive probability as the scoring rule. Moreover, the log posterior predictive score is the only coherent scoring rule under data exchangeability. This offers new insight into the marginal likelihood and cross-validation, and highlights the potential sensitivity of the marginal likelihood to the choice of the prior. We suggest an alternative approach using cumulative cross-validation following a preparatory training phase. Our work has connections to prequential analysis and intrinsic Bayes factors, but is motivated in a different way.

Some key words: Cross-validation; Marginal likelihood; Prequential scoring.

1. Introduction

Probabilistic model evaluation and selection is an important task in statistics and machine learning, particularly when multiple models are under initial consideration. In the non-Bayesian literature, models are typically compared using out-of-sample performance criteria such as cross-validation (Geisser & Eddy, 1979; Shao, 1993; Vehtari & Lampinen, 2002) and predictive information (Watanabe, 2010). Computing the leave-p-out cross-validation score requires n-choose-p test set evaluations for n data points, which in most cases is computationally unviable, so approximations such as k-fold cross-validation are often used instead (Geisser, 1975). A survey is provided in Arlot & Celisse (2010), and a Bayesian perspective on cross-validation can be found in Vehtari & Ojanen (2012) or Gelman et al. (2014).

In Bayesian statistics, the marginal likelihood or model evidence is the natural measure of model fit. For a model $\mathcal{M}$ with likelihood function or sampling distribution $\{f_\theta(y) : \theta \in \Theta\}$ parameterized by θ, a prior $\pi(\theta)$ and observations $y_{1:n} \in \mathcal{Y}^n$, the marginal likelihood or the prior predictive is defined as

$$p_{\mathcal{M}}(y_{1:n}) = \int f_\theta(y_{1:n})\, \mathrm{d}\pi(\theta). \tag{1}$$

The marginal likelihood can be used to calculate the posterior probability of the model given the data, $p(\mathcal{M} \mid y_{1:n}) \propto p_{\mathcal{M}}(y_{1:n})\, p(\mathcal{M})$, as it is the probability of the data being generated under the prior when the model is correctly specified (Robert, 2007, Ch. 7). The ratio of marginal likelihoods between models is known as a Bayes Factor, and quantifies the prior-to-posterior odds on observing the data. The marginal likelihood can be difficult to compute if the likelihood is peaked with respect to the prior, although Monte Carlo solutions exist; see Robert & Wraith (2009) for a survey. Under vague priors, the marginal likelihood may also be highly sensitive to the prior dispersion even if the posterior is not; a well-known example is Lindley's paradox (Lindley, 1957; O'Hagan & Forster, 2004; Robert, 2014). As a result, its approximations such as the Bayesian information criterion (Schwarz, 1978) or the deviance information criterion (Spiegelhalter et al., 2002) are widely used; see also Gelman et al. (2014).

For the present work, it is useful to note that from the property of probability distributions, the log marginal likelihood can be written as the sum of log conditionals,

$$\log p_{\mathcal{M}}(y_{1:n}) = \sum_{i=1}^{n} \log p_{\mathcal{M}}(y_i \mid y_{1:i-1}), \tag{2}$$

where $p_{\mathcal{M}}(y_i \mid y_{1:i-1}) = \int f_\theta(y_i)\, \mathrm{d}\pi(\theta \mid y_{1:i-1})$ is the posterior predictive probability for $i > 1$, $p_{\mathcal{M}}(y_1 \mid y_{1:0}) = \int f_\theta(y_1)\, \mathrm{d}\pi(\theta)$, and this representation is true for any permutation of the data indices.

While Bayesian inference formally assumes that the model space captures the truth, in the model-misspecified or so-called M-open scenario (Bernardo & Smith, 2009, Ch. 6) the log marginal likelihood can be simply interpreted as a predictive sequential, or prequential (Dawid, 1984), scoring rule of the form $S(y_{1:n}) = \sum_i s(y_i \mid y_{1:i-1})$ with score function $s(y_i \mid y_{1:i-1}) = \log p_{\mathcal{M}}(y_i \mid y_{1:i-1})$. This interpretation of the log marginal likelihood as a predictive score (Kass & Raftery, 1995; Gneiting & Raftery, 2007; Bernardo & Smith, 2009, Ch. 6) has resulted in alternative scoring functions for Bayesian model selection (Dawid & Musio, 2014, 2015; Watson & Holmes, 2016; Shao et al., 2019), and provides insight into the relationship between the marginal likelihood and posterior predictive methods (Vehtari & Ojanen, 2012). Key et al. (1999) considered cross-validation from an M-open perspective and introduced a mixture utility for model selection that trades off fidelity to data against predictive power.

2. Uniqueness of the marginal likelihood under coherent scoring

To begin, we prove that under an assumption of data exchangeability, the log posterior predictive score is the only prequential scoring rule that guarantees coherent model evaluation. The coherence property under exchangeability, where the indices of the data points carry no information, refers to the principle that identical models on seeing the same data should be scored equally irrespective of data ordering.

In demonstrating the uniqueness of the log posterior predictive, it is useful to introduce the notion of a general Bayesian model (Bissiri et al., 2016), which is a framework for Bayesian updating without the requirement of a true model. Define a parameter of interest by

$$\theta_0 = \arg\min_\theta \int l(\theta, y)\, \mathrm{d}F_0(y), \tag{3}$$

where $F_0(y)$ is the unknown true sampling distribution giving rise to the data, and $l : \Theta \times \mathcal{Y} \to \mathbb{R}$ is a loss function linking an observation y to the parameter θ. Bissiri et al. (2016) argue that after observing $y_{1:n}$, a coherent update of beliefs about θ_0 from a prior $\pi_G(\theta)$ to the posterior $\pi_G(\theta \mid y_{1:n})$ exists and must take the form

$$\pi_G(\theta \mid y_{1:n}) \propto \exp\{-w\, l(\theta, y_{1:n})\}\pi_G(\theta), \tag{4}$$

where $l(\theta, y_{1:n}) = \sum_i l(\theta, y_i)$ is an additive loss function and $w > 0$ is a loss scale parameter; see Holmes & Walker (2017), and see Lyddon et al. (2019) on the selection of w. For $w = 1$ and $l(\theta, y) = -\log f_\theta(y)$, we obtain traditional Bayesian updating without assuming that the model $f_\theta(y)$ is true for some value of θ. From (3), M-open Bayesian inference is simply targeting the value of θ that minimizes the Kullback–Leibler divergence between $\mathrm{d}F_0(y)$ and $f_\theta(y)$. The form (4) is uniquely implied by the assumptions in Theorem 1 of Bissiri et al. (2016), and we now focus on the coherence property of the updating rule. An update function $\psi\{l(\theta, y), \pi_G(\theta)\} = \pi_G(\theta \mid y)$ is coherent if for some inputs $y_{1:2}$ it satisfies

$$\psi\big[l(\theta, y_2), \psi\{l(\theta, y_1), \pi_G(\theta)\}\big] = \psi\{l(\theta, y_1) + l(\theta, y_2), \pi_G(\theta)\}.$$

This coherence condition is natural under an assumption of exchangeability, as we expect posterior inferences about θ_0 to be unchanged whether we observe $y_{1:2}$ in any order or all at once, as it is in traditional Bayesian updating.

We now extend this coherence condition to general Bayesian model choice, where the goal is to evaluate the fit of the observed data under the general Bayesian model class $\mathcal{M}_G = \{l(\theta, y) : \theta \in \Theta\}$ with a prior $\pi_G(\theta)$. We treat w as a parameter outside of the model specification, as there are principled methods to select it from the model, prior and data. We define the log posterior predictive score as

$$s_G(\tilde{y} \mid y_{1:n}) = \log \int g\{l(\theta, \tilde{y})\} \, \mathrm{d}\pi_G(\theta \mid y_{1:n}),$$

where $g : \mathbb{R} \to (0, \infty)$ is a continuous monotonically decreasing scoring function that transforms $l(\theta, y)$ into a predictive score for a test point $\tilde{y}$. We define the cumulative prequential log score as

$$S_G(y_{1:n}) = \sum_{i=1}^{n} s_G(y_i \mid y_{1:i-1}),$$

where $s_G(y_1 \mid y_{1:0}) = \log \int g\{l(\theta, y_1)\} \, \mathrm{d}\pi_G(\theta)$. The cumulative prequential log score sums the log posterior predictive score of each consecutive data point in a prequential manner, with a large score indicating that the model is predicting well. It might seem that there are many choices for g, but we will see that all but one violate coherency, as defined below.

Definition 1. *The model scoring function $g(l)$ is coherent if it satisfies*

$$\sum_{i=1}^{n} s_G(y_i \mid y_{1:i-1}) = \log \int g\{l(\theta, y_{1:n})\} \, \mathrm{d}\pi_G(\theta)$$

for all Θ, $\pi(\theta)$ and $n > 0$, such that $S_G(y_{1:n})$ is invariant with respect to the ordering or partitioning of the observations.

We now present our main result on the uniqueness of the choice of g.

Proposition 1. *If the model scoring function $g : \mathbb{R} \to (0, \infty)$ is continuous, monotonically decreasing and coherent, then the unique choice of scoring rule $g(l)$ is*

$$g(l) = \exp(-w\, l),$$

where w is the loss scale in the general Bayesian posterior.

The proof is given in the Supplementary Material. This result holds irrespective of whether the model is true or not. More important for us is the corollary below.

Corollary 1. *The marginal likelihood is the unique coherent marginal score for Bayesian inference.*

Proof. Let $w = 1$ and $l(\theta, y) = -\log f_\theta(y)$; then $g\{l(\theta, y)\} = f_\theta(y)$. □

The marginal likelihood arises naturally as the unique prequential scoring rule under coherent belief updating in the Bayesian framework. The coherence of the marginal likelihood implies invariance with respect to the permutation of the observations $y_{1:n}$ under exchangeability, including independent and identically distributed data, a property that is not shared by other prequential scoring rules such as those proposed by Dawid & Musio (2014), Grünwald & van Ommen (2017) and Shao et al. (2019).

3. The marginal likelihood and cross-validation

3.1. *Equivalence of the marginal likelihood and cumulative cross-validation*

The leave-p-out cross-validation score is defined as

$$S_{\mathrm{CV}}(y_{1:n};p) = \frac{1}{\binom{n}{p}} \sum_{t=1}^{\binom{n}{p}} \frac{1}{p} \sum_{j=1}^{p} s\big(\tilde{y}_j^{(t)} \mid y_{1:n-p}^{(t)}\big), \tag{5}$$

where $\tilde{y}_{1:p}^{(t)}$ denotes the tth of n-choose-p possible held-out test sets, with $y_{1:n-p}^{(t)}$ being the corresponding training set, such that $y_{1:n} = \{\tilde{y}^{(t)}, y^{(t)}\}$; S_{CV} records the average predictive score per datum. Although leave-one-out cross-validation is a popular method, it was shown in Shao (1993) that it is asymptotically inconsistent for a linear model selection problem, and requires $p/n \to 1$ as $n \to \infty$ for consistency. We do not go into further detail here, but instead refer the reader to Arlot & Celisse (2010). Selecting a larger p has the interpretation of penalizing complexity (Vehtari & Ojanen, 2012), as complex models will tend to overfit to a small training set. However, the number of test set evaluations grows rapidly with p, and hence k-fold cross-validation is often adopted for computational convenience.

From a Bayesian perspective it is natural to consider the log posterior predictive probability, $s(\tilde{y} \mid y) = \log \int f_\theta(\tilde{y})\, \mathrm{d}\pi(\theta \mid y)$, as the scoring function, particularly as we have now shown that it is the only coherent scoring mechanism, which leads us to the following result.

Proposition 2. *The Bayesian marginal likelihood is equivalent to the cumulative leave-p-out cross-validation score using the log posterior predictive probability as the scoring rule, such that*

$$\log p_{\mathcal{M}}(y_{1:n}) = \sum_{p=1}^{n} S_{\mathrm{CV}}(y_{1:n};p) \tag{6}$$

with $s(\tilde{y}_j \mid y_{1:n-p}) = \log p_{\mathcal{M}}(\tilde{y}_j \mid y_{1:n-p}) = \log \int f_\theta(\tilde{y}_j)\, \mathrm{d}\pi(\theta \mid y_{1:n-p})$.

Proof. This follows from the invariance of the marginal likelihood under arbitrary permutation of the sequence $y_{1:n}$ in (2). We provide a proof in the Appendix, and an alternative proof by induction is given in the Supplementary Material. □

The Bayesian marginal likelihood is simply n times the average leave-p-out cross-validation score, $n \times (1/n) \sum_{p=1}^{n} S_{\mathrm{CV}}(y_{1:n};p)$, where the scaling by n is due to (5) being a per-datum score. Bayesian models are evaluated through out-of-sample predictions on all $2^n - 1$ possible held-out test sets, whereas cross-validation with fixed p only captures a snapshot of model performance. Evaluating the predictive performance on $2^n - 1$ test sets would appear intractable for most applications, but we can see from (6) and (1) that it is computable as a single integral.

3.2. *Sensitivity to the prior and preparatory training*

The representation of the marginal likelihood as a cumulative cross-validation score (6) provides insight into the sensitivity to the prior. The last term on the right-hand side of (6), $S_{\mathrm{CV}}(y_{1:n};n) = (1/n) \sum_{i=1}^{n} \log \int f_\theta(y_i)\, \mathrm{d}\pi(\theta)$, involves no training data and scores the model entirely on how well the analyst is able to specify the prior. In many situations, the analyst may not want this term to contribute to model evaluation. Moreover, there is conflict between the desire to specify vague priors to safeguard their influence and the fact that diffuse priors can lead to an arbitrarily large and negative model score for real-valued parameters from (6). It may seem inappropriate to penalize a model based on the subjective ability to specify the prior, or to compare models using a score that includes contributions from predictions made using only a handful of training points even with informative priors. For example, we see that 10% of terms contributing to the marginal likelihood come from out-of-sample predictions, using on average less than 5% of the available training data. This is related to the start-up problem in prequential analysis (Dawid, 1992).

A natural and obvious solution is to begin evaluating the model performance after a preparatory phase, for example using 10% or 50% of the data for preparatory training prior to testing. This leads to a Bayesian cumulative leave-P-out cross-validation score, defined as

$$S_{\text{CCV}}(y_{1:n}; P) = \sum_{p=1}^{P} S_{\text{CV}}(y_{1:n}; p), \tag{7}$$

with a matched preparatory cross-validation score $S_{\text{PCV}}(y_{1:n}; P) = \sum_{p=P+1}^{n} S_{\text{CV}}(y_{1:n}; p)$ for $1 \leqslant P < n$. We suggest, as reasonable default choices, setting P to leave out $0.9n$, $0.5n$ or $\max(0.9n, n - 10d)$, where d is the total number of model parameters, but clearly this is situation-specific. One may be interested in reporting both S_{CCV} and S_{PCV}, as the latter can be regarded as an evaluation of the prior, but we recommend that only S_{CCV} be used for model evaluation from the arguments above. Although full coherency is now lost, we still have coherency conditioned on a preparatory training set, where permutation of the data within the training and test sets does not affect the score; therefore we can write (7) as

$$S_{\text{CCV}}(y_{1:n}; P) = \frac{1}{\binom{n}{P}} \sum_{t=1}^{\binom{n}{P}} \log p_{\mathcal{M}}\big(\tilde{y}_{1:P}^{(t)} \mid y_{1:n-P}^{(t)}\big). \tag{8}$$

This equivalence is derived in the Supplementary Material in a similar fashion to Proposition 2. This has precisely the form of the log geometric intrinsic Bayes factor of Berger & Pericchi (1996), but is motivated by a different route. The intrinsic Bayes factor was developed in an objective Bayesian setting (Berger & Pericchi, 2001), where improper priors cause indeterminacies in the evaluation of the marginal likelihood. The intrinsic Bayes factor remedies this with a partition of the data into $y_{1:l}$ and $y_{l+1:n}$, where $y_{1:l}$ is the minimum training sample used to convert an improper prior $\pi(\theta)$ into a proper prior $\pi(\theta \mid y_{1:l})$. In contrast, we set $n-P$ to provide preparatory training and $\pi(\theta)$ can be subjective. Moreover, in modern applications one often has $d \gg n$, in which case intrinsic Bayes factors cannot be applied in their original form.

We can approximate (8) via Monte Carlo where the training datasets $y_{1:n-P}^{(t)}$ are drawn uniformly at random, and for non-conjugate models the inner term must also be estimated, for example by

$$\hat{S}_{\text{CCV}}(y_{1:n}; P) = \frac{1}{T} \sum_{t=1}^{T} \log \left\{ \frac{1}{B} \sum_{b=1}^{B} f_{\theta_b^{(t)}}(\tilde{y}_{1:P}^{(t)}) \right\}, \tag{9}$$

where samples $\theta_b^{(t)} \sim \pi(\theta \mid y_{1:n-P}^{(t)})$ are obtained via T Markov chain Monte Carlo samplers. If we assume that the number B of samples per chain is sufficiently large, then the variance of the estimate $\hat{S}_{\text{CCV}}$ is approximately of the form τ^2/T. Fitting T models may be costly, but we can run the chains in parallel. To avoid the need for T Markov chain Monte Carlo chains in (9), we can instead take advantage of the fact that the partial posteriors for different training sets will be similar, and utilize importance sampling (Bhattacharya & Haslett, 2007; Vehtari et al., 2017) or sequential Monte Carlo (Bornn et al., 2010) to estimate the posterior predictives for computational savings. Further details on efficient computation of (9) are provided in the Supplementary Material.

4. Illustration for the normal linear model

We illustrate the use of Bayesian cumulative cross-validation in a polynomial regression example, where the rth polynomial model is defined as

$$f_\theta(y \mid x, r) = N\{y; \theta^{\mathrm{T}}\phi_r(x), \sigma^2\}, \quad \phi_r(x) = \begin{bmatrix} 1 & x & \ldots & x^{r-1} & x^r \end{bmatrix}^{\mathrm{T}}.$$

Table 1. *Log marginal likelihoods and cumulative cross-validation scores for the normal linear model*

s^2	Model r	$\log p_r(y_{1:n})$	$\hat{S}_{\mathrm{CCV}}(y_{1:n}; P) \times n/P$ $P = 0.9n$	$P = 0.5n$	$P = 0.1n$
10^{-1}	0	−158.82	−153.80	−153.21	−153.06
	1	−155.57	−150.39	−149.55	−149.27
	2	−156.12	−150.94	−149.81	−149.38
10^{0}	0	−158.82	−153.80	−153.21	−153.06
	1	−156.26	−150.77	−149.66	−149.34
	2	−157.80	−151.90	−150.04	−149.50
10^{4}	0	−158.82	−153.80	−153.21	−153.06
	1	−160.81	−150.91	−149.68	−149.35
	2	−166.93	−152.30	−150.08	−149.53
	Maximum standard error		0.002	0.008	0.023

We observe the data $\{y_{1:n}, x_{1:n}\}$, and we place a fixed vague prior on the intercept term, $\theta_0 \sim N(\theta_0; 0, 100^2)$, and place priors $\theta_d \sim N(\theta_d; 0, s^2)$ $(d \in \{1, \ldots, r\})$ on the remaining coefficients. In our example, $n = 100$ and the true model is $r = 1$ and $\theta = [1 \quad 0.5]^{\mathrm{T}}$ with known $\sigma^2 = 1$. For our prior, we vary the value of $s^2 \in \{10^{-1}, 10^0, 10^4\}$ to investigate the impact of the prior tails. For each prior setting, we calculate $\log p_{\mathcal{M}}(y_{1:n})$ and $S_{\mathrm{CCV}}(y_{1:n}; P)$ for models $r \in \{0, 1, 2\}$. In this example, $\log p_{\mathcal{M}}(y_{1:n})$ is tractable, whereas S_{CCV} requires a Monte Carlo average over tractable log posterior predictives. We report the mean over 10 runs of estimating S_{CCV} with $T = 10^6$ random training/test splits. We calculate the Monte Carlo standard error over the 10 runs and report the maximum for each setting of P.

The results are shown in Table 1, where $\hat{S}_{\mathrm{CCV}}$ is normalized to be on the same scale as $\log p_r(y_{1:n})$. Under the strong prior $s^2 = 10^{-1}$ and the moderate prior $s^2 = 10^0$, the marginal likelihood correctly identifies the true model, but as s^2 is increased to 10^4 it heavily overpenalizes the more complex models and prefers $r = 0$. In fact, the magnitude of the marginal likelihood and the discrepancy just described can be made arbitrarily large by simply increasing s^2, which should be guarded against when a modeller has weak prior beliefs. This issue is not observed with $\hat{S}_{\mathrm{CCV}}$ for the values of P we consider. The vague prior does not impede the ability of $\hat{S}_{\mathrm{CCV}}$ to correctly identify the true model $r = 1$, and the scores are stable within each column of P.

In the Supplementary Material, we present graphical tools for exploring the cumulative cross-validation and the effect of the choice of P on S_{CCV}. We also provide an additional example using probit regression on the Pima Indian dataset.

5. Discussion

We have shown that for coherence, the unique scoring rule for Bayesian model evaluation in either the M-open or the M-closed scenario is provided by the log posterior predictive probability, and that the marginal likelihood is equivalent to a cumulative cross-validation score over all training/test data partitions. The coherence stems from the fact that the scoring rule and the Bayesian update both use the same information, namely the likelihood function, which is appropriate as the alternative would be to learn and score under different criteria. If one is interested in an alternative loss function to the negative loglikelihood, we advocate a general Bayesian update (Bissiri et al., 2016; Lyddon et al., 2019) that targets the parameters minimizing the expected loss, with models evaluated using the corresponding coherent cumulative cross-validation score.

Acknowledgement

The authors thank Lucian Chan, George Nicholson, the editor, an associate editor and two referees for their helpful comments. Fong was funded by the Alan Turing Institute. Holmes was supported by the Alan Turing Institute, Health Data Research UK, the Li Ka Shing Foundation, the Medical Research Council, and the U.K. Engineering and Physical Sciences Research Council.

Supplementary material

Supplementary material available at *Biometrika* online includes the proof of Proposition 1, further derivations, a discussion of computational methods and graphical tools, and an additional probit regression example.

Appendix

Proof of Proposition 2

Consider the $n! \times n$ matrix Z with elements $(Z)_{ti} = \log p_{\mathcal{M}}(y_i^{(t)} \mid y_{1:i-1}^{(t)})$ such that the tth row of Z records the prequential sequence of log posterior predictives under the tth of $n!$ permutations of $y_{1:n}$. By the property of conditional probabilities, we have that the row sums of Z are equal, $\sum_i (Z)_{ti} = \sum_i (Z)_{t'i}$ for all t and t', and hence

$$\log p_{\mathcal{M}}(y_{1:n}) = \frac{1}{n!}\sum_{t=1}^{n!}\sum_{i=1}^{n}(Z)_{ti} = \sum_{i=1}^{n}\frac{1}{n!}\sum_{t=1}^{n!}(Z)_{ti}.$$

Within each column of Z, the values $(Z)_{ti}$ are invariant with respect to the permutation of $y_{1:i-1}$ in the preceding $i-1$ columns under exchangeability. Thus there are n-choose-$(i-1)$ distinct training sets and $n-i+1$ choices for y_i given the training set. For each column $i \in \{1,\ldots,n\}$, we can then write

$$\begin{aligned}\frac{1}{n!}\sum_{t=1}^{n!}(Z)_{ti} &= \frac{1}{\binom{n}{i-1}}\sum_{t=1}^{\binom{n}{i-1}}\frac{1}{n-i+1}\sum_{j=1}^{n-i+1} s\big(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)}\big)\\ &= S_{\mathrm{CV}}(y_{1:n}; n-i+1),\end{aligned}$$

where $s(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)}) = \log p_{\mathcal{M}}(\tilde{y}_j^{(t)} \mid y_{1:i-1}^{(t)})$. The result follows upon taking $p = n-i+1$. In the Supplementary Material we provide an alternative proof by induction.

References

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79.

Berger, J. O. & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Assoc.* **91**, 109–22.

Berger, J. O. & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model selection*, P. Lahiri, ed., vol. 38 of *Lecture Notes–Monograph Series*. Beachwood, Ohio: Institute of Mathematical Statistics, pp. 135–207.

Bernardo, J. M. & Smith, A. F. M. (2009). *Bayesian Theory*. Chichester: Wiley.

Bhattacharya, S. & Haslett, J. (2007). Importance re-sampling MCMC for cross-validation in inverse problems. *Bayesian Anal.* **2**, 385–407.

Bissiri, P. G., Holmes, C. C. & Walker, S. G. (2016). A general framework for updating belief distributions. *J. R. Statist. Soc.* B **78**, 1103–30.

Bornn, L., Doucet, A. & Gottardo, R. (2010). An efficient computational approach for prior sensitivity analysis and cross-validation. *Can. J. Statist.* **38**, 47–64.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. R. Statist. Soc.* A **147**, 278–92.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Oxford: Oxford University Press, pp. 109–25.

Dawid, A. P. & Musio, M. (2014). Theory and applications of proper scoring rules. *METRON* **72**, 169–83.

Dawid, A. P. & Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Anal.* **10**, 479–99.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Am. Statist. Assoc.* **70**, 320–8.

Geisser, S. & Eddy, W. (1979). A predictive approach to model selection. *J. Am. Statist. Assoc.* **74**, 153–60.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statist. Comp.* **24**, 997–1016.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.* **102**, 359–78.

Grünwald, P. & van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12**, 1069–103.

Holmes, C. C. & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104**, 497–503.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Am. Statist. Assoc.* **90**, 773–95.

Key, J. T., Pericchi, L. R. & Smith, A. F. M. (1999). Bayesian model choice: What and why? (with Discussion). In *Bayesian Statistics 6 (Proceedings of the Sixth Valencia International Meeting)*. Oxford: Oxford University Press, pp. 343–70.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–92.

Lyddon, S. P., Holmes, C. C. & Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106**, 465–78.

O'Hagan, A. & Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. London: Arnold.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer, 2nd ed.

Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Phil. Sci.* **81**, 216–32.

Robert, C. P. & Wraith, D. (2009). Computational methods for Bayesian model choice. *AIP Conf. Proc.* **1193**, 251–62.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Statist. Assoc.* **88**, 486–94.

Shao, S., Jacob, P. E., Ding, J. & Tarokh, V. (2019). Bayesian model comparison with the Hyvärinen score: Computation and consistency. *J. Am. Statist. Assoc.* **114**, 1826–37.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc.* B **64**, 583–639.

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comp.* **27**, 1413–32.

Vehtari, A. & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comp.* **14**, 2339–468.

Vehtari, A. & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statist. Surveys.* **6**, 142–228.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–94.

Watson, J. & Holmes, C. C. (2016). Approximate models and robust decisions. *Statist. Sci.* **31**, 465–89.

[*Received on* 22 *May* 2019. *Editorial decision on* 22 *August* 2019]