entropy

MDPI

Article

Posterior Averaging Information Criterion

Shouhao Zhou

Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Pennsylvania State University, Hershey, PA 17033, USA; shouhao.zhou@psu.edu

Abstract: We propose a new model selection method, named the posterior averaging information criterion, for Bayesian model assessment to minimize the risk of predicting independent future observations. The theoretical foundation is built on the Kullback–Leibler divergence to quantify the similarity between the proposed candidate model and the underlying true model. From a Bayesian perspective, our method evaluates the candidate models over the entire posterior distribution in terms of predicting a future independent observation. Without assuming that the true distribution is contained in the candidate models, the new criterion is developed by correcting the asymptotic bias of the posterior mean of the in-sample log-likelihood against out-of-sample log-likelihood, and can be generally applied even for Bayesian models with degenerate non-informative priors. Simulations in both normal and binomial settings demonstrate superior small sample performance.

Keywords: Bayesian modeling; expected out-of-sample likelihood; Kullback–Leibler divergence; misspecified model; predictive model selection

Citation: Zhou, S. Posterior Averaging Information Criterion. *Entropy* **2023**, *25*, 468. https://doi.org/10.3390/e25030468

Academic Editor: Udo Von Toussaint

Received: 15 November 2022
Revised: 22 February 2023
Accepted: 27 February 2023
Published: 7 March 2023

1. Introduction

Model selection plays a key role in statistical modeling and machine learning. Information theoretic criteria, such as Akaike information criterion (AIC) [1] minimum description length [2] and Schwarz information criterion [3], have been frequently and widely exploited with profound impact on many research fields.

Among these popular methods, a substantial group of model selection criteria was proposed based on the Kullback–Leibler (K-L) information divergence [4]. In the context of model selection, it provides an objective measure to quantify the overall closeness of a probability distribution (the candidate model) and the underlying true model. On both theoretical and applied fronts, K-L based information criteria have drawn a huge amount of attention, and a rich body of literature now exists for both frequentist and Bayesian modeling.

Here we will focus on predictive model selection. To choose a proper criterion for a statistical data analysis project, it is essential to distinguish the ultimate goal of modeling. Geisser & Eddy [5] challenged researchers with two fundamental questions that should be asked in advance of any procedure conducted for model selection:

- Which of the models best explains a given set of data?
- Which of the models yields the best predictions for future observations from the same process that generated the given set of data?

The first question, which concerns the accuracy of the model in describing the current data, has been an empirical problem for many years. It represents the explanatory perspective. The second question, which represents the predictive perspective, concerns the accuracy of the model in predicting future data, having drawn substantial attention in recent decades. If an infinitely large quantity of data is available, the predictive perspective and the explanatory perspective may converge. However, with a limited number of observations we encounter in practice, predictive model selection methods should achieve an optimal balance between goodness of fit and parsimony, for example, as we have seen in AIC.

Compared with frequentist methods, Bayesian approaches to statistical inference have unique concerns regarding the interpretation of parameters and models. However, many earlier Bayesian K-L information criteria, such as Deviance Information Criterion (DIC) [6], follow essentially the frequentist philosophy insofar as they select a model using plug-in estimators of the parameters. Subsequently, the parameter uncertainty is largely ignored. Such a paradigm has changed since the Bayesian predictive information criterion (BPIC) [7], as model selection criteria were developed over the entire posterior distribution. Nevertheless, BPIC has its own limitations, particularly with asymmetric posterior distributions. More importantly, BPIC is undefined under improper prior distributions, which limits its use in practice. More details can be found in Section 3 with a review of alternative methods.

The rest of this article is organized as follows. To explain the motivation of the proposed Bayesian criterion, in Section 2 we review the K-L divergence, its application and development in frequentist statistics, and the adaption to Bayesian modeling based on plug-in parameter estimation. In Section 3, major attention is given to the K-L based predictive criterion for models evaluated by averaging over the posterior distributions of parameters. To select models with better predictive performance, a generally applicable method, named the posterior averaging information criterion (PAIC), is proposed for comparing different Bayesian statistical models under mild regularity conditions. The new criterion is developed by correcting the asymptotic bias of using the posterior mean of the log-likelihood as an estimator of its expected log-likelihood, and we prove that the asymptotic property holds even though the candidate models are misspecified. In Section 4, we present some numerical studies in both normal and binomial cases to investigate its performance with small sample sizes. We also provide a real data variable selection example in Section 5 to exhibit possible differences between the explanatory and predictive approaches. We conclude with a few summary remarks and discussions in Section 6.

2. Kullback–Leibler Divergence and Model Selection

Kullback & Leibler [4] derived an information measure to assess the dissimilarity between any two models. If we assume that $f(y)$ and $g(y)$, respectively, represent the probability density distributions of the 'true model' and the 'approximate model' on the same measurable space, the K-L divergence is defined by

$$KL(f||g) = \int f(y) \cdot \log \frac{f(y)}{g(y)} dy = E_y[\log f(y)] - E_y[\log g(y)], \tag{1}$$

which is always non-negative, reaching the minimum value of 0 when f is the same as g almost surely. It is interpreted as the 'information' loss when g is used to approximate f. Namely, the smaller the value of $KL(f||g)$, the closer we consider the model g to be to the true distribution.

Only the second term of $KL(f||g)$ in (1) is relevant in practice to compare different possible models g without full knowledge of the true distribution. This is because the first term, $E_y[\log f(y)]$, is a constant that depends on only the unknown true distribution f, and can be neglected in model comparison for given data.

Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ be n independent observations of the data following probability density function $f(y)$. $\tilde{y}$ is a future independent observation following the same density function $f(y)$, representing an unknown but potentially observable quantity [8]. Without exactly knowing $f(y)$, we denote a model m with density $g_m(y|\theta^m)$ among a list of potential operating models $m = 1, 2, \cdots, M$. For notational purposes, we ignore the model index m when there is no ambiguity. The true model f is referred to as the *unknown* data generating mechanism, not necessarily to be encompassed in any approximate model family of g_m.

As the sample size $n \to \infty$, the average of the log-likelihood

$$\frac{1}{n} L(\theta|\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \log g(y_i|\theta)$$

tends to $E_{\tilde{y}}[\log g(\tilde{y}|\theta)]$ by the law of large numbers, which suggests how we can estimate the second term of $KL(f||g)$.

The model selection based on the K-L divergence is straightforward when all the operating models are fixed probability distributions, i.e., $g(y|\theta) = g(y)$. The model with the largest empirical log-likelihood $\sum_i \log g(y_i)$ is favored, when the observed data $\mathbf{y}$ are used as the test sample. However, when the distribution family $g(\tilde{y}|\theta)$ contains some unknown parameters θ, the direct comparison becomes no longer feasible. A typical strategy is to conduct the model fitting first, and then compare the operating models specified at the fitted parameters. In this case, the same data are indeed used twice—in both model fitting (as the training sample) and evaluation (as the test sample). Therefore, the in-sample log-likelihood is not optimal for the predictive modeling. For a desirable out-of-sample predictive performance, a common idea is to identify a bias correction term to rectify the over-estimation bias of the in-sample estimator, which is also the focus of this work.

In the frequentist setting, the general model selection procedure chooses candidate models specified by some point estimate $\hat{\theta}$ based on a certain statistical principle such as maximum likelihood. A considerable amount of theoretical research has addressed this problem by correcting for the bias of $\frac{1}{n}\sum_i \log g(y_i|\hat{\theta})$ in estimation of $E_{\tilde{y}}[\log g(\tilde{y}|\hat{\theta})]$ [1,9–11]. A nice review can be found in Burnham & Anderson [12].

Since the introduction of the AIC [1], researchers have commonly applied frequentist model selection methods into Bayesian modeling. However, the differences in the underlying philosophies between Bayesian and frequentist statistical inference caution against such direct applications. There also have been a few attempts to specialize the K-L divergence for Bayesian model selection (see, for example, [5,13,14]) in the last century. These methods are limited either in the scope of methodology or computational feasibility, especially when the parameters of the Bayesian models are in high-dimensional hierarchical structures.

The seminal work of Spiegelhalter et al. [6,15] proposed DIC,

$$DIC = D(\bar{\theta}) + 2p_D$$

as a Bayesian adaption of AIC and implemented it using Gibbs sampling (BUGS) [16], where $D(\theta)$ is the deviance function, $\bar{\theta}$ is the posterior mean and p_D is the effective number of parameters. Although its establishment lacks a theoretical foundation [17,18], $-dic/2n$, as a model selection measure, heuristically estimates $E_{\tilde{y}}[\log g(\tilde{y}|\bar{\theta})]$, the expected out-of-sample log-likelihood specified at the posterior mean, after assuming that the proposed model encompasses the true model. Alternative methods can be found either using a similar approach for mixed-effects models [19–21] or using numerical approximation [22] to estimate cross-validative predictive loss [23].

3. Posterior Averaging Information Criterion

The preceding methods in general can be viewed as Bayesian adaptation of the information criteria originally designed for frequentist statistics, when each model is assessed in terms of the similarity between the true distribution f and the model density function specified by the plug-in parameters. This may not be ideal since, in contrast to frequentist modeling, "Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by *a probability distribution on the parameters* of the model and on unobserved quantities such as predictions for new observations" [8]. Rather than considering a model specified by a point estimate, it is more reasonable to assess the goodness of a Bayesian model in terms of the posterior distribution.

3.1. Rationale and the Proposed Method

Ando [7] proposed an estimator for the posterior averaged discrepancy function,

$$\eta = E_{\tilde{y}}[E_{\theta|\mathbf{y}} \log g(\tilde{y}|\theta)].$$

Under certain regularity conditions, it was shown that an asymptotic unbiased estimator of η is

$$\begin{aligned}\hat{\eta}^{BPIC} &= \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - \frac{1}{n}[\, E_{\theta|\mathbf{y}}\log\{\pi(\theta)L(\theta|\mathbf{y})\} - \log\{\pi(\hat{\theta})L(\hat{\theta}|\mathbf{y})\} \\ &\quad + tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} + \frac{K}{2}] \\ &\triangleq \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - BC_1.\end{aligned} \tag{2}$$

Here, $\pi(\theta)$ is the prior distribution, $\hat{\theta}$ is the posterior mode, K is the cardinality of θ, and matrices J_n and I_n are some empirical estimators for the Bayesian asymptotic Hessian matrix,

$$J(\theta) = -E_{\tilde{y}}\left(\frac{\partial^2 \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta\partial\theta'}\right)$$

and Bayesian asymptotic Fisher information matrix,

$$I(\theta) = E_{\tilde{y}}\left(\frac{\partial \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}\frac{\partial \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta'}\right),$$

where $\log \pi_0(\theta) = \lim_{n\to\infty} n^{-1}\log \pi(\theta)$.

The Bayesian predictive information criterion (BPIC) was introduced as $-2n\cdot\hat{\eta}^{BPIC}$. It is applicable when the true model f is not necessarily in the specified family of probability distributions. In model comparison, the candidate model with a minimum BPIC value is favored. However, it has the following limitations in practice.

1. Equation (2) was from the original presentation for BPIC in Equation (5) of Ando [7]. After some math canceling out the term $\frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y})$ in both estimator and bias correction term, $\hat{\eta}^{BPIC}$ can be simplified as

$$\begin{aligned}\hat{\eta}^{BPIC} &= \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) - \frac{1}{n}[\, E_{\theta|\mathbf{y}}\log \pi(\theta) - \log \pi(\hat{\theta}) + tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} + \frac{K}{2}] \\ &\triangleq \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) - BC_2,\end{aligned} \tag{3}$$

which shows that it was actually the plug-in estimator $\frac{1}{n}\log L(\hat{\theta}|\mathbf{y})$, rather than natural estimator $\frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y})$, was used in estimation of η for bias correction. Compared with the natural estimator, the estimation efficiency of η using plug-in estimator is suboptimal when the posterior distribution is asymmetric.

2. The BPIC cannot be calculated when the prior distribution $\pi(\theta)$ is degenerate, a common situation in Bayesian analysis when an objective non-informative prior is selected. For example, if we use non-informative prior $\pi(\mu) \propto 1$ for the mean parameter μ of the normal distribution in the following Section 4.1, the values of $\log \pi(\hat{\theta})$ and $E_{\theta|\mathbf{y}}\log \pi(\theta)$ in Equation (3) are undefined.

In order to avoid those drawbacks, we propose a new model selection criterion in terms of the posterior mean of the empirical log-likelihood $\hat{\eta} = \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) = \frac{1}{n}\sum_i E_{\theta|\mathbf{y}}[\log g(y_i|\theta)]$, a natural estimator of estimand η. Without losing any of the attractive properties of BPIC, the new criterion expands the model scope to all regular Bayesian models. As we will show in the simulation study, empirically it also improves the unbiased property for small samples, and enhances the robustness of the estimation.

Because all the data $\mathbf{y}$ are used for both model fitting and model selection, $\hat{\eta}$ always overestimates η. To correct the estimation bias from the overuse of the data, we have the following theorem.

Theorem 1. *Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ be n independent observations drawn from the probability cumulative distribution $F(\tilde{y})$ with density function $f(\tilde{y})$. Consider $\mathcal{G} = \{g(\tilde{y}|\theta); \theta \in \Theta \subseteq \mathbb{R}^p\}$ as a family of candidate statistical models that do not necessarily contain the true distribution f, where $\theta = (\theta_1, \ldots, \theta_p)'$ is the p-dimensional vector of unknown parameters, with prior distribution $\pi(\theta)$. Under the following three regularity conditions:*

C1: Both the log density function $\log g(\tilde{y}|\theta)$ and the log unnormalized posterior density $\log\{L(\theta|\boldsymbol{y})\pi(\theta)\}$ are twice continuously differentiable in the compact parameter space Θ;
C2: The expected posterior mode $\theta_0 = \arg\max_\theta E_{\tilde{y}}[\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}]$ is unique in Θ;
C3: The Hessian matrix of $E_{\tilde{y}}[\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}]$ is non-singular at θ_0,

the bias of $\hat{\eta}$ for η can be approximated asymptotically without bias by

$$\hat{\eta} - \eta = \widehat{b_\theta} \cong \frac{1}{n} tr\{J_n^{-1}(\hat{\theta}) I_n(\hat{\theta})\}, \tag{4}$$

where $\hat{\theta}$ is the posterior mode that maximizes the posterior distribution $\propto \pi(\theta)\prod_{i=1}^n g(y_i|\theta)$ and

$$\begin{aligned} J_n(\theta) &= -\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial^2 \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta\partial\theta'}\right) \\ I_n(\theta) &= \frac{1}{n-1}\sum_{i=1}^n \left(\frac{\partial \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}\frac{\partial \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta'}\right). \end{aligned}$$

Proof. Recall that the quantity of interest is $E_{\tilde{y}}E_{\theta|\mathbf{y}}\log g(\tilde{y}|\theta)$. To estimate it, we first check $E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\} = E_{\tilde{y}}E_{\theta|\mathbf{y}}\{\log g(\tilde{y}|\theta) + \log\pi_0(\theta)\}$ and expand it about θ_0,

$$\begin{aligned} E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\} &= E_{\tilde{y}}\log\{g(\tilde{y}|\theta_0)\pi_0(\theta_0)\} + E_{\theta|\mathbf{y}}(\theta-\theta_0)'\frac{\partial E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad + \frac{1}{2}E_{\theta|\mathbf{y}}[(\theta-\theta_0)'\frac{\partial^2 E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\theta_0}(\theta-\theta_0)] + o_p(n^{-1}) \\ &= E_{\tilde{y}}\log\{g(\tilde{y}|\theta_0)\pi_0(\theta_0)\} + E_{\theta|\mathbf{y}}(\theta-\theta_0)'\frac{\partial E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad - \frac{1}{2}E_{\theta|\mathbf{y}}[(\theta-\theta_0)'J(\theta_0)(\theta-\theta_0)] + o_p(n^{-1}) \\ &\triangleq I_1 + I_2 + I_3 + o_p(n^{-1}) \end{aligned} \tag{5}$$

The first term I_1 can be linked to the empirical log likelihood function as follows:

$$\begin{aligned} E_{\tilde{y}}\log\{g(\tilde{y}|\theta_0)\pi_0(\theta_0)\} &= E_{\tilde{y}}\log g(\tilde{y}|\theta_0) + \log\pi_0(\theta_0) \\ &= E_y\frac{1}{n}\log L(\theta_0|\mathbf{y}) + \log\pi_0(\theta_0) \\ &= E_y\frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} - \frac{1}{n}\log\pi(\theta_0) + \log\pi_0(\theta_0) \\ &= E_y E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\} - \frac{1}{2n}tr\{J_n^{-1}(\theta_0)I(\theta_0)\} \\ &\quad + \frac{1}{2n}tr\{J_n^{-1}(\hat{\theta})J_n(\theta_0)\} - \frac{1}{n}\log\pi(\theta_0) + \log\pi_0(\theta_0) + o_p(n^{-1}) \end{aligned}$$

where the last equation holds due to Lemma A5 (together with other Lemmas, provided in the Appendix A).

The second term I_2 vanishes since

$$\frac{\partial E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}|_{\theta=\theta_0} = 0$$

as θ_0 is the expected posterior mode.

Using Lemma A4, the third term I_3 can be rewritten as

$$\begin{aligned} I_3 &= -\frac{1}{2}E_{\theta|\mathbf{y}}(\theta-\theta_0)'J(\theta_0)(\theta-\theta_0) \\ &= -\frac{1}{2}tr\{E_{\theta|\mathbf{y}}[(\theta-\theta_0)(\theta-\theta_0)']J(\theta_0)\} \\ &= -\frac{1}{2n}(tr\{J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)J(\theta_0)\}+tr\{J_n^{-1}(\hat{\theta})J(\theta_0)\})+o_p(n^{-1}) \end{aligned}$$

By substituting each term in Equation (5) and neglecting the residual term, we obtain

$$\begin{aligned} E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\} \simeq\ & E_yE_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\}-\frac{1}{2n}tr\{J_n^{-1}(\theta_0)I(\theta_0)\} \\ & +\frac{1}{2n}tr\{J_n^{-1}(\hat{\theta})J_n(\theta_0)\}-\frac{1}{n}\log\pi(\theta_0)+\log\pi_0(\theta_0) \\ & -\frac{1}{2n}(tr\{J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)J(\theta_0)\}+tr\{J_n^{-1}(\hat{\theta})J(\theta_0)\}) \end{aligned}$$

Recall that we have defined $\log\pi_0(\theta)=\lim_{n\to\infty}n^{-1}\log\pi(\theta)$, so that asymptotically we have

$$\begin{aligned} \log\pi_0(\theta_0)-\tfrac{1}{n}\log\pi(\theta_0)&\simeq 0, \\ E_{\theta|\mathbf{y}}\log\{\pi_0(\theta)\}-E_{\theta|\mathbf{y}}\tfrac{1}{n}\log\{\pi(\theta)\}&\simeq 0. \end{aligned}$$

Therefore, $E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\}$ can be estimated by

$$\begin{aligned} E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\} &= E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}-E_{\theta|\mathbf{y}}\log\{\pi_0(\theta)\} \\ &\simeq E_yE_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\}-\frac{1}{2n}tr\{J_n^{-1}(\theta_0)I(\theta_0)\}+\frac{1}{2n}tr\{J_n^{-1}(\hat{\theta})J_n(\theta_0)\} \\ &\quad -\frac{1}{2n}(tr\{J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)J(\theta_0)\}+tr\{J_n^{-1}(\hat{\theta})J(\theta_0)\}) \\ &\quad -\frac{1}{n}\log\pi(\theta_0)+\log\pi_0(\theta_0)-E_{\theta|\mathbf{y}}\log\{\pi_0(\theta)\} \\ &\simeq E_yE_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\}-\frac{1}{2n}tr\{J_n^{-1}(\theta_0)I(\theta_0)\}+\frac{1}{2n}tr\{J_n^{-1}(\hat{\theta})J_n(\theta_0)\} \\ &\quad -\frac{1}{2n}(tr\{J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)J(\theta_0)\}+tr\{J_n^{-1}(\hat{\theta})J(\theta_0)\}) \end{aligned}$$

Replacing θ_0 by $\hat{\theta}$, $J(\theta_0)$ by $J_n(\hat{\theta})$ and $I(\theta_0)$ by $I_n(\hat{\theta})$, we obtain $E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\}-\frac{1}{n}tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\}$ as an asymptotically unbiased estimate for $E_{\tilde{y}}E_{\theta|\mathbf{y}}\log\{g(\tilde{y}|\theta)\}$. □

With the above result, we propose a new predictive criterion for Bayesian modeling, named the Posterior Averaging Information Criterion (PAIC),

$$PAIC=-2\sum_i E_{\theta|\mathbf{y}}[\log g(y_i|\theta)]+2tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\}. \tag{6}$$

The candidate models with small criterion values (6) are preferred for the purpose of model selection.

Remark 1. *PAIC selects the candidate models with optimal performance to predict a future outcome.*

The optimality is defined in a sense to maximize the out-of-sample log density η, which is equivalent to minimize the posterior predictive K-L divergence.

Remark 2. *PAIC is derived without assuming that the approximating distributions contain the truth.*

In another word, PAIC is generally applicable even if all candidate models are misspecified. In such settings, rather than select the true model, the goal is to identify the best candidate model(s) with small PAICs among all models under consideration. Similar to other K-L based information criteria, we consider a model is better if its PAIC is smaller with a difference larger than 2.

Remark 3. *The averaging over the posterior distribution in empirical likelihood helps differentiate the candidate models.*

The posterior distribution function, rather than a point estimator, represents the current best knowledge from a Bayesian perspective. In some cases, two candidate models may have identical posterior mean but different posterior distributions. (A simple example could be in the setting of Section 4.1, when model A has $\tau_0 = 1000$ and model B has $\tau_0 = 1$ in the prior distribution.) Apparently, Bayesian model assessment with respect to the posterior distribution is more effective in model selection. When the posterior distribution of the parameters is asymmetric, the estimation of information criterion averaged over the posterior is also more robust than plugging in a point estimator.

Remark 4. *PAIC can be applied to Bayesian models with flexible prior structures.*

For example, in cases when the prior distributions are consistent and sample size dependent [24,25], the information in the prior distribution does not degenerate asymptotically, but is accommodated spontaneously in empirical log-likelihood and bias-correction for predictive assessment. Unlike the BPIC, PAIC relaxes the restriction in common prior distribution specification. It is well-defined and can cope with degenerate non-informative prior distributions for parameters. The bias correction term $tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\}$ is closely related to the concept of measuring a Bayesian model's complexity [26]. Particularly, when the candidate model is true and has no hierarchical structures, and the prior distribution is non-informative with a dimension of p, we have exactly $tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} = p$, which is similar to the bias correction in AIC [1].

3.2. Relevant Methods for the Posterior Averaged K-L Discrepancy

Rather than deriving the bias correction analytically, resampling approaches, such as cross-validation and bootstrap, can also be used to measure the posterior averaged K-L discrepancy. Plummer [22] introduced the expected deviance penalized loss with 'expected deviance' defined as

$$L^e(y_i, z) = -2E_{\theta|z} \log g(y_i|\theta),$$

which is a special case of the predictive discrepancy measure [27]. The standard cross-validation method can also be applied in this circumstance to estimate η, simply by considering the K-L discrepancy as the utility function of [28] and further investigated by [29]. The estimation of the bootstrap error correction $\eta^{(b)} - \hat{\eta}^{(b)}$ with bootstrap analogues

$$\eta^{(b)} = E_{\tilde{y^*}}[E_{\theta|y^*} \log g(\tilde{y}|\theta)]$$

and

$$\hat{\eta}^{(b)} = E_{\tilde{y^*}}[n^{-1}E_{\theta|y^*} \log L(\theta|y^*)]$$

for $\eta - \hat{\eta}$ was discussed by Ando [7] as a Bayesian adaptation of frequentist model selection [10]. Although numeric algorithms such as importance sampling can be used for intensive computation, one caveat is that it may cause inaccurate estimation in practice if some observation y_i was influential [28]. To address that problem, Vehtari [30] proposed Pareto smoothed importance sampling, a new algorithm for regularizing importance weights, and developed a numerical tool [31] to facilitate computation. Watanabe [32] established a singular learning theory and proposed a new criterion named Watanabe–Akaike [29], or widely applicable information criterion (WAIC) [33,34], while

WAIC_1 was proposed for the plug-in discrepancy and WAIC_2 for the posterior averaged discrepancy. However, compared with BPIC and PAIC, we found that WAIC_2 tends to have a larger bias and variation for regular Bayesian models, as shown in simulation studies in the next section.

4. Simulation Study

In this section, we present some numerical results to illustrate the performance of the proposed method under small sample sizes. Assuming K-L divergence is a good measure for model selection, our goal is simply to assess how it can be estimated with the smallest bias. In the simulation experiments, we estimate the true expected bias η either analytically in a Gaussian setting (Section 4.1) or numerically by averaging $E_{\theta|\mathbf{y}}[\log g(\tilde{y}|\theta)]$ over a large number of extra independent draws of $\tilde{y}$ when there is asymmetric posterior distribution and no closed form for the integration (Section 4.2). To have BPIC well-defined for comparison, only the proper prior distributions are considered.

4.1. A Case with Closed-Form Expression for Bias Estimators

Suppose observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are a vector of iid samples generated from $N(\mu_T, \sigma_T^2)$, with unknown true mean μ_T and variance $\sigma_T^2 = 1$. Assume the data are analyzed by the approximating model $g(y_i|\mu) = N(\mu, \sigma_A^2)$ with prior $\pi(\mu) = N(\mu_0, \tau_0^2)$, where σ_A^2 is fixed, but not necessarily equal to the true variance σ_T^2. When $\sigma_A^2 \neq \sigma_T^2$, the model is misspecified.

The posterior distribution of μ is normally distributed with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, where

$$\begin{aligned} \hat{\mu} &= (\mu_0/\tau_0^2 + \sum_{i=1}^{n} y_i/\sigma_A^2)/(1/\tau_0^2 + n/\sigma_A^2) \\ \hat{\sigma}^2 &= 1/(1/\tau_0^2 + n/\sigma_A^2). \end{aligned}$$

Therefore, the K-L discrepancy function and its estimator are

$$\begin{aligned} \eta &= E_{\tilde{y}}[E_{\mu|\mathbf{y}}[\log g(\tilde{y}|\mu)]] = -\frac{1}{2}\log(2\pi\sigma_A^2) - \frac{\sigma_T^2 + (\mu_T - \hat{\mu})^2 + \hat{\sigma}^2}{2\sigma_A^2} \\ \hat{\eta} &= \frac{1}{n}\sum_{i=1}^{n} E_{\mu|\mathbf{y}}[\log g(y_i|\mu)]] = -\frac{1}{2}\log(2\pi\sigma_A^2) - \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu})^2 + \hat{\sigma}^2}{2\sigma_A^2}. \end{aligned}$$

We assess the bias estimator defined in Theorem 1, $\hat{b}_\mu^{PAIC}$ and four other bias estimators: $\hat{b}_\mu^{BPIC}$ [7], $\hat{b}_\mu^{WAIC_2}$ [33], $\hat{b}_\mu^{p_{opt}}$ [22], and $\hat{b}_\mu^{CV}$ [35].

$$\begin{aligned} \hat{b}_\mu^{PAIC} &= \frac{1}{n-1}\hat{\sigma}^2\sum_{i=1}^{n}((\mu_0 - \hat{\mu})/(n\tau_0^2) + (y_i - \hat{\mu})/\sigma_A^2)^2 \\ \hat{b}_\mu^{BPIC} &= \frac{1}{n}\hat{\sigma}^2\sum_{i=1}^{n}((\mu_0 - \hat{\mu})/(n\tau_0^2) + (y_i - \hat{\mu})/\sigma_A^2)^2 \\ \hat{b}_\mu^{WAIC_2} &= \frac{\hat{\sigma}^2}{\sigma_A^4}(n\hat{\sigma}^2/2 + \sum_{i=1}^{n}(y_i - \hat{\mu})^2) \\ \hat{b}_\mu^{p_{opt}} &= \frac{1}{2n}p_{opt} = 1/(1/\tau_0^2 + (n-1)/\sigma_A^2)/\sigma_A^2 \\ \hat{b}_\mu^{CV} &= \hat{\eta} - (\sum_{i=1}^{n}(y_i - (\mu_0/\tau_0^2 + \sum_{j\neq i} y_j/\sigma_A^2)/(1/\tau_0^2 + (n-1)/\sigma_A^2))^2/n + \hat{\sigma}^2)/\sigma_A^2/2. \end{aligned}$$

We compare them with the true bias

$$b_\mu = E_{\mathbf{y}}(\hat{\eta} - \eta) = E_{\mathbf{y}}\{\frac{\sigma_T^2}{2\sigma_A^2} + \frac{(\mu_T - \hat{\mu})^2}{2\sigma_A^2} - \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu})^2}{2\sigma_A^2}\} = \sigma_T^2\hat{\sigma}^2/\sigma_A^4.$$

The results are in accordance with the theory (Figure 1). All of the estimates are close to the true bias-correction values when the model is well-specified with $\sigma_A^2 = \sigma_T^2 = 1$, especially when the sample size becomes moderately large (Figure 1, panels (a), (b), and (c)). The estimated values based on the PAIC are consistently closer to the true values than either those based on Ando's method, which underestimates the bias, or the WAIC_2, cross-validation or expected deviance penalized loss, which overestimate the bias, especially when the sample size is small. When the models are misspecified, it is not surprising that in all of the plots given in panels (d)–(i) of Figure 1, only the expected deviance penalized loss misses the target even asymptotically since its assumption is violated, whereas all the other approaches converge to b_μ. In summary, PAIC achieves the best overall performance.

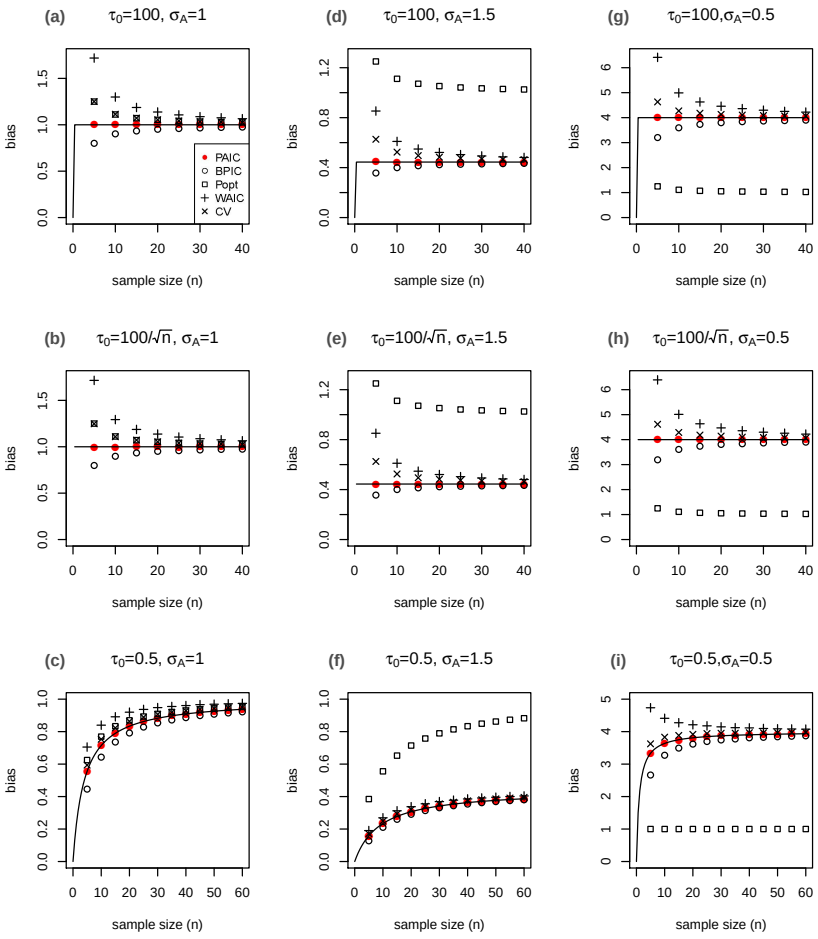

Figure 1. Performance of the bias estimators for $n \times E_y(\hat{\eta} - \eta)$. The top panels are under a relatively non-informative prior with $\tau_0^2 = 10^4$; the middle panels are under the case that the prior distribution grows with sample size with $\tau_0^2 = 10^4/n$; the bottom panels are under an informative prior with $\tau_0^2 = 0.25$. The left panels (**a**–**c**) are under the scenario of $\sigma_A^2 = \sigma_T^2 = 1$, i.e., the true distribution is contained in the candidate models. The middle panels (**d**–**f**) are under the scenario of $\sigma_A^2 = 2.25$ and right panels (**g**–**i**) are under the scenario of $\sigma_A^2 = 0.25$ when the proposed model is misspecified from $\sigma_T^2 = 1$. The true bias b_μ is curved by (—) as a function of sample size n. The averages of the different bias estimators are marked by (•) for PAIC; (○) for BPIC; (□) for p_{opt}; (+) for WAIC_2; and (×) for cross-validation. Each mark represents the mean of the estimated bias of 100,000 replications of $\mathbf{y}$.

4.2. Bayesian Logistic Regression

Consider frequencies $\mathbf{y} = \{y_1, \ldots, y_N\}$, which are independent observations from binomial distributions with respective true probabilities $\xi_1^T, \ldots, \xi_N^T$, and sample sizes, $n_1, \ldots, n_N$. To draw the inference of the ξ's, we assume that the logits

$$\beta_i = \text{logit}(\xi_i) = \log \frac{\xi_i}{1 - \xi_i}$$

are random effects that follow the normal distribution $\beta_i \sim N(\mu, \tau^2)$. The weakly-informative joint prior distribution $N(\mu; 0, 1000^2) \cdot Inv\text{-}\chi^2(\tau^2; 0.1, 10)$ is proposed on the hyper-parameter (μ, τ^2) so that the BPIC is properly defined and computable. The posterior distribution is asymmetric due to the logistic transformation.

We compare the performance of four asymptotically unbiased bias estimators in this hierarchical, asymmetric setting. The true bias η does not have an analytical form. We estimate it through numerical computation using independent simulation from the same data generating process, assuming the underlying true values of $\mu = 0$ and $\tau = 1$. The simulation scheme is as follows:

1. Draw $\beta_{T,i} \sim N(0, 1)$, $y_i \sim Bin(n_i, \text{logit}^{-1}(\beta_{T,i}))$, $i = 1, \ldots, N$ from the true distribution.
2. Simulate the posterior draws of $(\beta, \mu, \tau)|\mathbf{y}$.
3. Estimate $\hat{b}_\beta^{PAIC}$, $\hat{b}_\beta^{BPIC}$, $\hat{b}_\beta^{WAIC_2}$, and $\hat{b}_\beta^{CV}$.
4. Draw $\mathbf{z}^{(j)} \sim Bin(n, \text{logit}^{-1}(\beta_0^T))$, $j = 1, \ldots, J$, for approximation of true η.
5. Compare each $\hat{b}_\beta$ with true bias $b_\beta = \hat{\eta} - \eta$.
6. Repeat steps 1–5.

Table 1 summarizes the bias and standard deviation of the estimation error when we choose $N = 15$ and $n_1 = \ldots = n_N = 50$, and the β's are independently simulated from the standard normal distribution assuming the true hyper-parameter mean $\mu = 0$ and variance $\tau^2 = 1$. The simulation is repeated for 1000 scenarios, each with $J = 20{,}000$ for out-of-sample η estimation. PAIC and BPIC were calculated based on definition; leave-one-out cross-validation and WAIC_2 were estimated using *R* package *loo* v2.5.1 [31]. The actual error, mean absolute error, and mean square error were considered to assess the estimation error using the bias correction estimates. With respect to all three different metrics, the bias estimation of PAIC is consistently superior to other methods. Compared to BPIC, the second best performed model selection criterion, the bias, and the mean squared error of PAIC are reduced by about 40%, while the absolute bias is reduced by about one quarter, which matches our expectation that the natural estimate $\frac{1}{n}\sum_i E_{\theta|\mathbf{y}}[\log g(y_i|\theta)]$ will estimate the posterior averaged K-L discrepancy more precisely than plug-in estimate $\frac{1}{n}\sum_i \log g(y_i|\hat{\theta})$ when the posterior distribution is asymmetric and correlated. Compared to WAIC_2, the bias, absolute error, and mean square error of PAIC are dramatically reduced by at least 60%. In practice, we expect the improvement is even larger when proposed models have more complicated hierarchical structures.

Table 1. The estimation error of bias correction: the mean and standard deviation (in parentheses) from 1000 replications.

Criterion	Actual Error $\hat{\eta} - \eta - \hat{b}_\beta$	Mean Absolute Error $\lvert\hat{\eta} - \eta - \hat{b}_\beta\rvert$	Mean Square Error $(\hat{\eta} - \eta - \hat{b}_\beta)^2$
PAIC	**0.160 (0.238)**	**0.206 (0.199)**	**0.082 (0.207)**
BPIC	0.259 (0.244)	0.272 (0.229)	0.127 (0.267)
CV	0.840 (0.285)	0.840 (0.285)	0.786 (0.633)
$WAIC_2$	0.511 (0.248)	0.511 (0.248)	0.323 (0.389)

As suggested by reviewers, we also assessed PAIC in bias estimation with different priors, including the commonly used $Inv\text{-}Gamma^2(\tau^2; 0.001, 0.001)$ [36]. Although these

priors may produce different posterior distributions, we found almost identical results in terms of bias estimation error to Table 1, suggesting the robustness of the proposed method. Furthermore, we examined the BPIC and PAIC for uncorrelated posterior distributions of βs, by fixing the hyperparameters (μ, τ^2) either at its true value or at the posterior mode. In the simulation replications containing extreme observations (i.e., $\exists$ $i \in \{1, \ldots, N\}$, such that either $y_i = 0$ or $y_i = n_i$), we observed a large deviation of the plug-in estimate $1/N \log L(\hat{\theta}|\mathbf{y})$ to η, which cannot be properly recovered by BPIC's bias correction term in Equation (3) and yields significant estimation error; meanwhile, the plug-in estimand $E_{\tilde{y}}[\log g(\tilde{y}|\hat{\beta})]$ was also much more vulnerable to the observed data than $\eta = E_{\tilde{y}}[E_{\beta|\mathbf{y}} \log g(\tilde{y}|\beta)]$ given the extreme value, suggesting that the latter (the posterior averaged discrepancy) could be a better choice for model assessment.

5. Application

This is a variable selection example that uses real data to illustrate the practical difference between criteria proposed in either the explanatory or predictive perspective. We explore the problem of finding the best model to predict the selling of new accounts in branches of a large bank. The data were introduced in example 5.3 of George & McCulloch [37], analyzed with their method, the stochastic search variable selection (SSVS) technique to select the promising subsets of predictors. Their report on the 10 most frequently selected models after 10,000 iterations of Gibbs sampling for potential subsets, is listed in the first column of Table 2.

The original data consist of the numbers of new accounts sold in some time period as the outcome $\mathbf{y}$, together with 15 predictor variables X in each of 233 branches. Multiple linear regressions are employed to fit the data in the form of

$$y_i|\beta^{(m)}, \sigma_y^2 \sim N(X^{(m)}\beta^{(m)}, \sigma_y^2)$$

with prior $\beta_i^{(m)} \sim N(0, 1000^2)$ and $\sigma_y^2 \sim Inv\text{-}Gamma(0.001, 0.001)$, when m indicates the specific model with a subset of predictor $X^{(m)}$.

Table 2. Comparison of model performance using K-L based model selection criteria for SSVS example. The first row indicates the independent variables (x) to be excluded in each model. The mid rule separates the models most frequently appeared using SSVS method (above) vs. the models with lower PAIC (below).

Exclusion	SSVS	LOO-CV	KCV	$PL_{p^e_{opt}}$	BPIC	PAIC
4, 5	827	2603.85	2580.74	2527.32	2528.89	2529.60
2, 4, 5	627	2572.98	2564.92	2544.77	2533.90	2534.44
3, 4, 5, 11	595	2583.63	2572.59	2545.23	2539.79	2540.20
3, 4, 5	486	2593.10	2579.97	2567.85	2541.75	2542.32
3, 4	456	2590.36	2571.76	2538.80	2533.37	2533.97
4, 5, 11	390	2589.76	2573.04	**2526.77**	**2527.94**	**2528.58**
2, 3, 4, 5	315	2576.66	2577.17	2561.57	2553.29	2553.77
3, 4, 11	245	2579.53	2566.28	2565.22	2532.87	2533.42
2, 4, 5, 11	209	**2564.67**	**2559.36**	2540.41	2533.60	2534.03
2, 4	209	2741.46	2741.17	2737.46	2740.42	2740.51
5, 10, 12	n/a	2602.23	2572.86	**2519.41**	2525.07	2525.61
4, 12	n/a	2596.51	2570.94	2520.52	2524.31	2524.94
5, 12	n/a	2595.86	2570.32	2520.51	**2524.19**	**2524.90**
4, 5, 12	n/a	2596.67	2574.73	2525.65	2526.19	2526.86
4, 10, 12	n/a	2603.05	2573.80	2520.62	2525.17	2525.70
4, 5, 10, 12	n/a	2603.51	2577.86	2526.53	2527.06	2527.56

Several model selection estimators for $-2n \cdot \eta$, including the leave-one-out cross-validated estimator (LOO-CV), K-fold cross-validated estimator (KCV), the expected deviance penalized loss with p^e_{opt}, BPIC, and PAIC, are calculated based on a large number

of MCMC draws of the posterior distribution for model selection inference. In KCV, the original data are randomly partitioned for the K-fold cross-validation with a common choice of $K = 10$. All the posterior samples are simulated from three parallel chains based on MCMC techniques for model selection inference. To generate 15,000 effective draws of the posterior distribution, only one out of five iterations after convergence are kept to reduce the serial correlation.

The results are presented in Table 2, in which the models that have the smallest estimation value by each criterion are highlighted. The first 10 models with SSVS frequencies were originally picked by SSVS as shown in George & McCulloch [37]. An interesting finding is that the favored models selected by the K-L based criteria and SSVS are quite different. All of the K-L based criteria are developed in a predictive perspective, whereas SSVS is a variable selection method to pursue the model that best describes the given set of data. This illustrates that with different modeling purposes, either explanatory or predictive, the 'best' models found may not coincide. The estimated $PL_{p^e_{opt}}$, BPIC, and PAIC values for every candidate model are quite close to each other; whereas the cross-validation estimators are noisy due to the simulation error and tendency to overestimate the value. It is worth mentioning that the estimators of LOO-CV, K-fold-CV, and $PL_{p^e_{opt}}$ are relatively unstable, even with 15,000 posterior draws. Those methods have been much more computationally intensive than BPIC and PAIC.

6. Discussion

A clearly defined model selection criterion or score usually lies at the heart of any statistical selection and decision procedure. It facilitates the comparison of competing models through the assignment of some sort of preference or ranking to the alternatives. One of the typical scores is the K-L divergence, a non-symmetric measure of the difference between two probability distributions. By further acknowledging uncertainty in parameters and randomness in data, frequentist statistics theoretically employing K-L divergence into parametric model selection emerged during the 1970s. Since then, the development of related theories and applications has rapidly accelerated.

A good assessment measure helps establish attractive properties. To guide the Bayesian method development, two important questions should be first investigated.

1. What is a good estimand, based on K-L discrepancy, to evaluate Bayesian models?
2. What is a good estimator to estimate the estimand for K-L based Bayesian model selection?

The prevailing plug-in parameter methods, such as DIC, presume the candidate models are correct, and assess the goodness of each candidate model with a density function specified by the plug-in parameters. However, from a Bayesian perspective, it is inherent to examine the performance of a Bayesian model over the entire posterior distribution, as stated by (Celeux et al. [18], p. 703): "...we concede that using a plug-in estimate disqualifies the technique from being properly Bayesian." Accordingly, statistical approaches to estimate the K-L discrepancy as evaluated by averaging over the posterior distribution are of great interest.

We have proposed PAIC, a versatile model selection technique for Bayesian models under regularity assumptions, to address this problem. From a predictive perspective, we consider the asymptotic unbiased estimation of a K-L discrepancy, which averages the conditional density of the observable data against the posterior knowledge about the unobservable data. Empirically, the proposed PAIC measures the similarity of the fitted model and the underlying true distribution, regardless of whether or not the approximating distribution family contains the true model. The range of applications of the proposed criterion can be quite broad.

PAIC and BPIC are similar in many aspects. In addition to all the asymptotic properties and similar computational costs both methods share, PAIC has some unique features, mainly because it employs the natural posterior-averaged estimator. For example, PAIC

can be well applied even if the prior distribution of the parameters degenerates, in which case BPIC becomes uninterpretable. In the illustrative experiments, we focused on the comparison of estimation accuracy between the proposed criterion and other Bayesian model selection criteria including BPIC and $WAIC_2$. PAIC showed the least bias and variance to estimate the posterior averaged discrepancy.

Because the regularity condition assumes twice continuously differentiability and non-singularity, it could be problematic if the posterior mode is on the boundary of the parameter space Θ. For example, as pointed out by one reviewer, $\hat{\tau} = 0$ in the famous eight-school example [8]. This is a common concern for K-L based model selection since the method derivation relies on the the Taylor series expansion. However, in practice, a reparameterization may help. In the eight-school example, we can introduce the uniform prior $\phi = \log \tau \sim Unif(0,1)$ to pair with the weakly informative prior $\mu \sim N(0,100)$, which yields a posterior mode for $\hat{\tau} = 1.125$.

There are some future directions for the current work. In the current simulation setting, we made a default assumption that the estimand, i.e., the posterior-averaged out-of-sample log-likelihood, can be distinguished between candidate models. A more comprehensive comparison of Bayesian predictive methods for empirical model selection can be investigated by taking into account the likely over-fitting in the selection phase, similar to [38]. Because the users of PAIC and BPIC have to specify the first and second derivatives of the posterior distribution in their modeling, development of advanced computational tools for simultaneous calculations will be helpful. In singular learning machines, the regularity conditions can be relaxed to singular in a sense that the mapping from parameters to probability distributions is not necessarily one-to-one. Although here we focused on only the regular models, it is also possible to generalize PAIC to singular settings with a modified bias correction term, after an algebraic geometrical transformation of the singular parameter space to a real d-dimensional manifold. Finally, other metrics for comparing the distance or dissimilarity between two distributions, such as Hellinger distance [39] or Jensen–Shannon divergence [40], may be explored further and employed as alternative metrics in Bayesian model assessment.

Funding: This research was funded part by Columbia University GSAS Faculty Fellowship and NIH/NCI Grant CA100632.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author would like to acknowledge Ciprian Giurcaneanu, four anonymous referees, one associate editor and editor for careful reviews and constructive comments that substantially improved the article. The author is also grateful to David Madigan and Andrew Gelman for helpful discussions, and to Lee Ann Chastain for editorial assistance.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIC	Akaike information criterion
BPIC	Bayesian predictive information criterion
DIC	Deviance information criterion
K-L	Kullback–Leibler
PAIC	Posterior averaging information criterion
WAIC	Watanabe–Akaike information criterion

Appendix A. Supplementary Materials for Proof of Theorem 1

Appendix A.1. Some Important Notations

By the law of large numbers we have $\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\} \to E_{\tilde{y}}[\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}]$ as n tends to infinity. Denote θ_0, $\hat{\theta}$ the expected and empirical posterior mode of the log unnormalized posterior density $\log\{L(\theta|\mathbf{y})\pi(\theta)\}$, *i.e.*,

$$\begin{aligned}\theta_0 &= \arg\max_{\theta} E_{\tilde{y}}[\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}] \\ \hat{\theta} &= \arg\max_{\theta} \frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\},\end{aligned}$$

and let $I(\theta)$ and $J(\theta)$ denote the Bayesian Hessian matrix and Bayesian Fisher information matrix

$$I(\theta) = E_{\tilde{y}}\left(\frac{\partial \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}\frac{\partial \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta'}\right)$$

and

$$J(\theta) = -E_{\tilde{y}}\left(\frac{\partial^2 \log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta\partial\theta'}\right).$$

Appendix A.2. Proof of Lemmas

We start with a few lemmas to support the proofs of Theorem 1.

Lemma A1. *Under the same regularity conditions of Theorem 1, $\sqrt{n}(\hat{\theta}-\theta_0)$ is asymptotically distributed as $N(0, J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0))$.*

Proof. Consider the Taylor expansion of $\frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}}$ at θ_0,

$$\begin{aligned}\frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}} &\simeq \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} + \frac{\partial^2 \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\theta_0}(\hat{\theta}-\theta_0) \\ &= \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} - nJ_n(\theta_0)(\hat{\theta}-\theta_0).\end{aligned}$$

Note that $\hat{\theta}$ is the mode of $\log\{L(\mathbf{y}|\theta)\pi(\theta)\}$ and satisfies $\frac{\partial \log\{L(\mathbf{y}|\theta)\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}} = 0$. Plug it into the above equation, we have

$$nJ_n(\theta_0)(\hat{\theta}-\theta_0) \simeq \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0}. \tag{A1}$$

From the central limit theorem, the right-hand-side (RHS) of Equation (A1) is approximately distributed as $N(0, nI(\theta_0))$ when $E_y\frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \to 0$. Therefore,

$$\sqrt{n}(\hat{\theta}-\theta_0) \sim N(0, J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)).$$

□

Lemma A2. *Under the same regularity conditions of Theorem 1, $\sqrt{n}(\theta-\hat{\theta}) \sim N(0, J_n^{-1}(\hat{\theta}))$.*

Proof. Taylor-expand the logarithm of $L(\theta|\mathbf{y})\pi(\theta)$ around the posterior mode $\hat{\theta}$

$$\log L(\theta|\mathbf{y})\pi(\theta) = \log L(\hat{\theta}|\mathbf{y})\pi(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})'\frac{1}{n}J_n^{-1}(\hat{\theta})(\theta-\hat{\theta}) + o_p(n^{-1}) \tag{A2}$$

where $J_n(\hat{\theta}) = -\frac{1}{n}\frac{\partial^2 \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}}$.

Consider the RHS of Equation (A2) as a function of θ: the first term is a constant, whereas the second term is proportional to the logarithm of a normal density. It yields the approximation of the posterior distribution for θ:

$$p(\theta|\mathbf{y}) \approx N(\hat{\theta}, \frac{1}{n} J_n^{-1}(\hat{\theta})),$$

which completes the proof.

Alternatively, though less intuitive, this lemma can also be proved by applying the Berstein–Von Mises theorem. □

Lemma A3. *Under the same regularity conditions of Theorem 1,* $E_{\theta|\mathbf{y}}(\theta_0 - \hat{\theta})(\hat{\theta} - \theta)' = o_p(n^{-1})$.

Proof. First we have

$$\frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} = \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta}|_{\theta=\hat{\theta}} - nJ_n(\hat{\theta})(\theta - \hat{\theta}) + O_p(1).$$

Since $\hat{\theta}$ is the mode of $\log\{L(\theta|\mathbf{y})\pi(\theta)\}$, it satisfies $\frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta}|_{\theta=\hat{\theta}} = 0$. Therefore, $(\hat{\theta} - \theta) = n^{-1} J_n^{-1}(\hat{\theta}) \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} + O_p(n^{-1})$. Note that

$$\begin{aligned} E_{\theta|\mathbf{y}} \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} &= \int \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} \frac{L(\theta|\mathbf{y})\pi(\theta)}{p(\mathbf{y})} d\theta \\ &= \int \frac{1}{L(\theta|\mathbf{y})\pi(\theta)} \frac{\partial\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} \frac{L(\theta|\mathbf{y})\pi(\theta)}{p(\mathbf{y})} d\theta \\ &= \frac{1}{p(\mathbf{y})} \int \frac{\partial\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} d\theta \\ &= \frac{1}{p(\mathbf{y})} \frac{\partial}{\partial \theta} \int L(\theta|\mathbf{y})\pi(\theta) d\theta = \frac{\partial}{\partial \theta} 1 = 0. \end{aligned}$$

Because of assumption (C1), the equation holds when we change the order of the integral and derivative. Therefore,

$$E_{\theta|\mathbf{y}}(\hat{\theta} - \theta) = n^{-1} J_n^{-1}(\hat{\theta}) E_{\theta|\mathbf{y}} \frac{\partial \log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial \theta} + O_p(n^{-1}) = O_p(n^{-1}).$$

Together with $\theta_0 - \hat{\theta} = O_p(n^{-1/2})$ derived from Lemma A1, we complete the proof. □

Lemma A4. *Under the same regularity conditions of Theorem 1,* $E_{\theta|\mathbf{y}}(\theta_0 - \theta)(\theta_0 - \theta)' = \frac{1}{n} J_n^{-1}(\hat{\theta}) + \frac{1}{n} J_n^{-1}(\theta_0) I(\theta_0) J_n^{-1}(\theta_0) + o_p(n^{-1})$.

Proof. $E_{\theta|\mathbf{y}}(\theta_0 - \theta)(\theta_0 - \theta)'$ can be rewritten as $(\theta_0 - \hat{\theta})(\theta_0 - \hat{\theta})' + E_{\theta|\mathbf{y}}(\hat{\theta} - \theta)(\hat{\theta} - \theta)' + 2E_{\theta|\mathbf{y}}(\theta_0 - \hat{\theta})(\hat{\theta} - \theta)$. Applying Lemmas A1–A3, we complete the proof. □

Lemma A5. *Under the same regularity conditions of Theorem 1,*

$$\begin{aligned} E_{\theta|\mathbf{y}} \frac{1}{n} \log\{L(\boldsymbol{y}|\theta)\pi(\theta)\} \simeq & \frac{1}{n} \log\{L(\theta_0|\boldsymbol{y})\pi(\theta_0)\} \\ & + \frac{1}{2n}(tr\{J_n^{-1}(\theta_0) I(\theta_0)\} - tr\{J_n^{-1}(\hat{\theta}) J_n(\theta_0)\}) + O_p(n^{-1}). \end{aligned}$$

Proof. The posterior mean of the log joint density distribution of $(\mathbf{y}, \theta)$ can be Taylor-expanded around θ_0 as

$$\begin{aligned} E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\} &= \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} + E_{\theta|\mathbf{y}}(\theta-\theta_0)'\frac{1}{n}\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad +\frac{1}{2}E_{\theta|\mathbf{y}}(\theta-\theta_0)'\frac{1}{n}\frac{\partial^2\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\theta_0}(\theta-\theta_0)+o_p(n^{-1}) \\ &= \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} + E_{\theta|\mathbf{y}}(\theta-\theta_0)'\frac{1}{n}\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad -\frac{1}{2}E_{\theta|\mathbf{y}}(\theta-\theta_0)'J_n(\theta_0)(\theta-\theta_0)+o_p(n^{-1}). \end{aligned} \tag{A3}$$

Expand $\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}}$ around θ_0 to the first order, we obtain

$$\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}} = \frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} - nJ_n(\theta_0)(\hat{\theta}-\theta_0)+O_p(n^{-1}). \tag{A4}$$

Because the posterior mode $\hat{\theta}$ is the solution of $\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}=0$, Equation (A4) can be re-written as

$$\frac{1}{n}\frac{\partial\log\{L(\theta|\mathbf{y})\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} = J_n(\theta_0)(\hat{\theta}-\theta_0)+O_p(n^{-1}).$$

Substituting it into the second term of (A3), the expansion of $E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\}$ becomes:

$$\begin{aligned} E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\theta|\mathbf{y})\pi(\theta)\} &= \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} + E_{\theta|\mathbf{y}}(\theta-\theta_0)'J_n(\theta_0)(\hat{\theta}-\theta_0) \\ &\quad -\frac{1}{2}E_yE_{\theta|\mathbf{y}}(\theta-\theta_0)'J_n(\theta_0)(\theta-\theta_0)+o_p(n^{-1}) \\ &= \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} + tr\{E_{\theta|\mathbf{y}}[(\hat{\theta}-\theta_0)(\theta-\theta_0)']J_n(\theta_0)\} \\ &\quad -\frac{1}{2}tr\{E_{\theta|\mathbf{y}}[(\theta-\theta_0)(\theta-\theta_0)']J_n(\theta_0)\}+o_p(n^{-1}) \\ &= \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} + tr\{E_{\theta|\mathbf{y}}[(\theta-\theta_0)(\hat{\theta}-\theta_0)']J_n(\theta_0)\} \\ &\quad -\frac{1}{2}tr\{\frac{1}{n}[J_n^{-1}(\hat{\theta})+J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)]J_n(\theta_0)\}+o_p(n^{-1}) \end{aligned}$$

where in the last line we replace $E_{\theta|\mathbf{y}}[(\theta-\theta_0)(\theta-\theta_0)']$ with the result of Lemma A4. $E_{\theta|\mathbf{y}}[(\theta-\theta_0)(\hat{\theta}-\theta_0)']$ in the second term of the expansion can be rewritten as $E_{\theta|\mathbf{y}}[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)'] + E_{\theta|\mathbf{y}}[(\theta-\hat{\theta})(\hat{\theta}-\theta_0)']$, where the former term is asymptotically equal to $\frac{1}{n}J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)$ by Lemma A1, and the latter is negligible with higher order $o_p(n^{-1})$, as shown in Lemma A3. Therefore, the expansion can be finally simplified as

$$\begin{aligned} E_{\theta|\mathbf{y}}\frac{1}{n}\log\{L(\mathbf{y}|\theta)\pi(\theta)\} &\simeq \frac{1}{n}\log\{L(\theta_0|\mathbf{y})\pi(\theta_0)\} \\ &\quad +\frac{1}{2n}(tr\{J_n^{-1}(\theta_0)I(\theta_0)\}-tr\{J_n^{-1}(\hat{\theta})J_n(\theta_0)\})+O_p(n^{-1}). \end{aligned}$$

□

Appendix B. Supplementary Materials for Derivation of Equation (3)

We start from Equation (2), which rewrites Equation (5) in Ando [7].

$$
\begin{aligned}
\hat{\eta}^{BPIC} &= \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - \frac{1}{n}[\, E_{\theta|\mathbf{y}}\log\{\pi(\theta)L(\theta|\mathbf{y})\} - \log\{\pi(\hat{\theta})L(\hat{\theta}|\mathbf{y})\} \\
&\quad + tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} + \frac{K}{2}] \\
&= \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - \frac{1}{n}[\, E_{\theta|\mathbf{y}}\log\pi(\theta) + E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - \log\pi(\hat{\theta}) - \log L(\hat{\theta}|\mathbf{y}) \\
&\quad + tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} + \frac{K}{2}] \\
&= \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) - \frac{1}{n}\,E_{\theta|\mathbf{y}}\log\pi(\theta) - \frac{1}{n}E_{\theta|\mathbf{y}}\log L(\theta|\mathbf{y}) + \frac{1}{n}\log\pi(\hat{\theta}) + \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) \\
&\quad - \frac{1}{n}tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} - \frac{K}{2n} \\
&= \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) - \frac{1}{n}\,E_{\theta|\mathbf{y}}\log\pi(\theta) + \frac{1}{n}\log\pi(\hat{\theta}) - \frac{1}{n}tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} - \frac{K}{2n} \\
&= \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) - \frac{1}{n}[\, E_{\theta|\mathbf{y}}\log\pi(\theta) - \log\pi(\hat{\theta}) + tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} + \frac{K}{2}] \\
&\triangleq \frac{1}{n}\log L(\hat{\theta}|\mathbf{y}) - BC_2.
\end{aligned}
$$

References

1. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer Series in Statistics; Springer: New York, NY, USA, 1998; pp. 267–281.
2. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
3. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
4. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86. [CrossRef]
5. Geisser, S.; Eddy, W.F. A predictive approach to model selection. *J. Am. Stat. Assoc.* **1979**, *74*, 153–160. [CrossRef]
6. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van der Linde, A. Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. B* **2002**, *64*, 583–639. [CrossRef]
7. Ando, T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* **2007**, *94*, 443–458. [CrossRef]
8. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed.; CRC Press: London, UK, 2003.
9. Hurvich, C.; Tsai, C. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [CrossRef]
10. Konishi, S.; Kitagawa, G. Generalised information criteria in model selection. *Biometrika* **1996**, *83*, 875–890. [CrossRef]
11. Takeuchi, K. Distributions of information statistics and criteria for adequacy of models. *Math. Sci.* **1976**, *153*, 15–18. (In Japanese)
12. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference*, 2nd ed.; Springer: New York, NY, USA, 2002.
13. Laud, P.W.; Ibrahim, J.G. Predictive model selection. *J. R. Stat. Soc. B* **1995**, *57*, 247–262. [CrossRef]
14. San Martini, A.; Spezzaferri, F. A predictive model selection criterion. *J. R. Stat. Soc. B* **1984**, *46*, 296–303. [CrossRef]
15. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; Van der Linde, A. The deviance information criterion: 12 years on. *J. R. Stat. Soc. B* **2002**, *76*, 485–493. [CrossRef]
16. Spiegelhalter, D.J.; Thomas, A.; Best, N.G. *WinBUGS Version 1.2 User Manual*; MRC Biostatistics Unit: Cambridge, UK, 1999.
17. Meng, X.L.; Vaida, F. Comments on 'Deviance Information Criteria for Missing Data Models'. *Bayesian Anal.* **2006**, *70*, 687–698.
18. Celeux, G.; Forbes, F.; Robert, C.P.; Titterington, D.M. Deviance information criteria for missing data models. *Bayesian Anal.* **2006**, *70*, 651–676. [CrossRef]
19. Liang, H.; Wu, H.; Zou, G. A note on conditional AIC for linear mixed-effects models. *Biometrika* **2009**, *95*, 773–778. [CrossRef]
20. Vaida, F.; Blanchard, S. Conditional Akaike information for mixed effects models. *Biometrika* **2005**, *92*, 351–370. [CrossRef]
21. Donohue, M.C.; Overholser, R.; Xu, R.; Vaida, F. Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika* **2011**, *98*, 685–700. [CrossRef]
22. Plummer, M. Penalized loss functions for Bayesian model comparison. *Biostatistics* **2008**, *9*, 523–539. [CrossRef]
23. Efron, B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316–331. [CrossRef]
24. Lenk, P.J. The logistic normal distribution for Bayesian non parametric predictive densities. *J. Am. Stat. Assoc.* **1988**, *83*, 509–516. [CrossRef]
25. Walker, S.; Hjort, N.L. On bayesian consistency. *J. R. Stat. Soc. B* **2001**, *63*, 811–821. [CrossRef]
26. Hodges, J.S.; Sargent, D.J. Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika* **2001**, *88*, 367–379. [CrossRef]
27. Gelfand, A.E.; Ghosh, S.K. Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **1998**, *85*, 1–11. [CrossRef]

28. Vehtari, A.; Lampinen, J. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Comput.* **2002**, *14*, 1339–2468. [CrossRef] [PubMed]
29. Gelman, A.; Hwang, J.; Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **2014**, *24*, 997–1016. [CrossRef]
30. Vehtari, A.; Gelman, A.; Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **2017**, *27*, 1413–1432. [CrossRef]
31. Vehtari, A.; Gabry, J.; Yao Y.; Gelman, A. loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models. R Package Version 2.5.1. 2018. Available online: https://CRAN.R-project.org/package=loo (accessed on 28 August 2022).
32. Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **2010**, *11*, 3571–3594.
33. Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: Cambridge, UK, 2009.
34. Watanabe, S. A formula of equations of states in singular learning machines. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 2098–2105.
35. Stone, M. Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Stat. Soc. B* **1974**, *36*, 111–147. [CrossRef]
36. Gelman, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper) *Bayesian Anal.* **2006**, *1*, 515–534. [CrossRef]
37. George, E.I.; McCulloch, R. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 881–889. [CrossRef]
38. Piironen, J.; Vehtari, A. Comparison of Bayesian predictive methods for model selection. *Stat. Comput.* **2017**, *27*, 711–735. [CrossRef]
39. Beran, R. Minimum Hellinger distance estimates for parametric models. *Ann. Stat.* **1977**, *5*, 445–463. [CrossRef]
40. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef]