Bayesian Model Selection in terms of Kullback-Leibler discrepancy

Shouhao Zhou

Submitted in partial fulfillment of the
Requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

Abstract

Bayesian Model Selection in terms of Kullback-Leibler discrepancy

Shouhao Zhou

In this article we investigate and develop the practical model assessment and selection methods for Bayesian models, when we anticipate that a promising approach should be objective enough to accept, easy enough to understand, general enough to apply, simple enough to compute and coherent enough to interpret. We mainly restrict attention to the Kullback-Leibler divergence, a widely applied model evaluation measurement to quantify the similarity between the proposed candidate model and the underlying true model, where the true model is only referred to a probability distribution as the best projection onto the statistical modeling space once we try to understand the real but unknown dynamics/mechanism of interest. In addition to review and discussion on the advantages and disadvantages of the historically and currently prevailing practical model selection methods in literature, a series of convenient and useful tools, each designed and applied for different purposes, are proposed to asymptotically unbiasedly assess how the candidate Bayesian models are favored in terms of predicting a future independent observation. What's more, we also explore the connection of the Kullback-Leibler based information criterion to the Bayes factors, another most popular Bayesian model comparison approaches, after seeing the motivation through the developments of the Bayes factor variants. In general, we expect to provide a useful guidance for researchers who are interested in conducting Bayesian data analysis.

Contents

1 Introduction **1**

2 Bayesian Generalized Information Criterion **9**

2.1 Kullback-Leibler divergence: an objective criterion for model comparison 9

2.2 K-L based Bayesian predictive model selection criteria 11

2.3 Bayesian Generalized Information Criterion 17

2.4 A simple linear example 22

3 Posterior Averaging Information Criterion **27**

3.1 Model evaluation with posterior distribution 27

3.2 Posterior Averaging Information Criterion 29

3.3 Simulation Study 38

3.3.1 A simple linear example 38

3.3.2 Bayesian hierarchical logistic regression 41

3.3.3 Variable selection: a real example 44

4 Predictive Bayes factor **47**

4.1 Bayes Factors 47

4.2 Predictive Bayes factor 49

4.2.1 Models under comparison: Original or Fitted? 49

4.2.2 Predictive Bayes factor 51
4.3 Posterior Predictive Information Criterion 53
4.3.1 Asymptotic estimation for K-L discrepancy 54
4.3.2 A simple simulation study 61
4.4 Interpretation of predictive Bayes factor 64
4.5 Simulation Study 66

5 Conclusion and Discussion **69**
5.1 Conclusion 69
5.2 Discussion 70

References **74**

List of Figures

2.1 Comparison of criteria with plug-in estimators: $\sigma_A^2 = \sigma_T^2$ 24

2.2 Comparison of criteria with plug-in estimators: $\sigma_A^2 \neq \sigma_T^2$ 25

3.1 Comparison of criteria with posterior averaging: $\sigma_A^2 = \sigma_T^2$ 39

3.2 Comparison of criteria with posterior averaging: $\sigma_A^2 \neq \sigma_T^2$ 40

4.1 Comparison of bias and its PPIC estimator: $\sigma_A^2 = \sigma_T^2$ 62

4.2 Comparison of bias and its PPIC estimator: $\sigma_A^2 \neq \sigma_T^2$ 63

List of Tables

3.1 Criteria comparison for Bayesian logistic regression 42

3.2 Variable explanation of SSVS example 43

3.3 Criteria value comparison for SSVS example 44

4.1 Jeffreys' scale of evidence in favor of model M_1. (Jeffreys, 1961) . . . 65

4.2 Scale of evidence in favor of model M_1 by Kass and Raftery (1995). . 65

4.3 The interpretation of PrBF and difference of PPICs with respect to the posterior probability in favor of model $M_1(y)$. 66

4.4 Beetles Killed after Exposure to Carbon Disulfide (Bliss 1935) 67

4.5 Various information criteria for Beetles data (Bliss 1935) 68

4.6 Various Bayes factors comparison for Beetles data (Bliss 1935) 68

Acknowledgments

First and for most, I would like to thank the Almighty God for his abounding grace that helped me to endure the arduous task of doctoral study. Without His unfailing grace, a successful completion of this challenging journey would not have been possible.

I owe sincere and earnest gratitude to Prof. David Madigan, who stepped in as my advisor late during one of the toughest time in my dissertation process after Prof. Andrew Gelman took his academic leave to France. I have been amazingly fortunate to have all the insightful comments, constructive suggestions, enlightening motivations, and endurable encouragements from Prof. Madigan, who also consistently pushed me through all the writings. I have benefited significantly from his invaluable guidance and support that I will cherish forever. I am also truly indebted to Prof. Andrew Gelman, who shared his insights with me, trained me of my research skills, sparked my interest in Bayesian model selection and thus inspired my dissertation's work throughout discussions during many years. His influence will remain with me.

I would like to give a special thanks to Prof. Zhiliang Ying, my committee chair who steadily provided his encouragement and help when many have questioned whether I would finish my dissertation. I am also grateful to the remaining members of my dissertation committee for their constructive suggestions and willingness to help; the members of the statistics department, especially Dood Kalicharan, for making my graduate experience a truly memorable one; and all my friends at CCBSG for their prayers and support.

Of course no acknowledgments would be complete without giving heartfelt thanks to my parents, who have loved me without any reservation, instilled many decent qualities in me and given me a good foundation with which to meet life, as well as to my wife - my precious helper gifted from God, my patient soul mate co-walking

through my happiness and sufferings, and my faithful editor of my numerous poorly-written drafts. She has been a constant source of love, support, and strength in all these years.

Chapter 1

Introduction

The choice of an appropriate model to characterize the underlying distribution for the given set of data is essential for applied statistical practice. There has been considerable discussion over the past half century and numerous theoretical works have been contributed to its development. Just for multiple linear regression, a partial list of the model assessment tools is composed of adjusted R^2 (Wherry, 1931), Mallow's C_p (Mallows, 1973, 1995), Akaike information criterion (AIC, Akaike, 1973, 1974), prediction sum of squares (PRESS, Allen, 1974), generalized cross-validation (GCV, Craven and Wahba, 1979), minimum description length (MDL, Rissanen, 1978), S_p criterion (Breiman and Freedman, 1983), Fisher information criterion (FIC, Wei, 1992), risk inflation criterion (RIC, Foster and George, 1994), L-criterion (Laud and Ibrahim, 1995), generalized information criterion (GIC, Konishi and Kitagawa, 1996), covariance inflation criterion (CIC, Tibshirani and Knight, 1999) and focused information criterion (FIC, Claeskens and Hjort, 2003), to name but a few. (On the topic of the subset variable selection in regression, see Hocking (1976) for the review of early works, George (2000) for the recent development and Miller (1990, 2002) for the comprehensive introduction and bibliography.) For the criteria corresponding to time series modeling, some important findings are final prediction error

(FPE, Akaike, 1969), autoregressive transfer function criterion (CAT, Parzen, 1974), Hannan-Quinn's criterion (HQ, Hannan and Quinn, 1979) and corrected AIC (AIC_c, Hurvich and Tsai, 1989), while an introduction on the time series model selection techniques is given by McQuarrie and Tsai (1998). One special kind of model selection technique related is the automatic regression procedures, by which the choice of explanatory variables is carried out according to a specific criterion, such as those mentioned above. It includes all possible subsets regression (Garside, 1965) and stepwise regression, the latter of which consists of forward selection (Efroymson, 1966) and backward elimination (Draper and Smith, 1966).

Compared with the abundance of model selection proposals in the frequentist domain, Bayesian methods also have drawn a large amount of attention. The availability of both fast computers and advanced numerical methods in recent years enables the empirical popularity of Bayesian modeling, which allows the additional flexibility to incorporate the information out of the data, represented by the prior distribution. The fundamental assumption of Bayesian inference is also quite different, for the unknown parameters are treated as random variables, in the form of a probability distribution. Taking the above into account, it is important to have the selection techniques specially designed for Bayesian modeling. In the literature, most of the key model selection tools for Bayesian models can be classified into two categories:

1. methods with respect to posterior model probability, including Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995), Schwarz information criterion (SIC, Schwarz, 1978), posterior Bayes factors (Aitkin, 1991), fractional Bayes factors (O'Hagan, 1995) and intrinsic Bayes factors (Berger and Pericchi, 1996), etc.

2. methods with respect to Kullback-Leibler divergence, including deviance information criterion (DIC, Spiegelhalter et al., 2002), conditional AIC (cAIC, Vaida and Blanchard, 2005), Bayesian predictive information criterion (BPIC, Ando,

2007), deviance penalized loss (Plummer, 2008), etc.

There are also one kind of generally applicable procedures, such as cross-validation (Stone, 1974; Geisser, 1975) and Bootstrap (Efron, 1979; Efron and Tibshirani, 1993), requiring a loss/discrepancy function to be specified in advance for model performance assessment, in accordance with either a frequentist or Bayesian philosophy. A list of widely accepted discrepancy functions is provided in Linhart and Zucchini (1986). Stone (1979) shows that the cross-validation method employs the Kullback-Leibler discrepancy and AIC is asymptotically equivalent in the order of $o_p(1)$.

Explanatory vs Predictive

From either a frequentist or a Bayesian perspective, it is essential to distinguish the ultimate goal of modeling when confronting a statistical data analysis project. Geisser and Eddy (1979) challenge research workers 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 which generated the given set of data?

The former, which cares about how accurately a model describes the current data in the explanatory point of view, has been the problem of empirical science for many years; whereas the latter, which focuses on predicting future data as accurately as possible in the predictive perspective, is more crucial and difficult to answer and has drawn more attentions in recent decade.

If an infinitely large quantity of data is available, the predictive perspective and the explanatory perspective might not differ significantly. With only limited number of observations in practice, it is a more difficult task for predictive model selection

methods to achieve an optimal balance between goodness of fit and parsimony. A central issue for the predictive methods is to avoid the impact from the 'double use' of the data, i.e. the whole set of data is used both in the parameter estimation stage and in the model evaluation stage. One solution is to split the data into two independent subsets, using one as the training set to fit the model and the other as the testing set to assess the validity of the model. Subsequently, it is crucial to implement it with either cross-validation or bootstrap, for the data-split approach obviously reduces the data usage efficiency and intermediately raises the question how to make the proper data separation. However, it is quite computer-intensive to apply those numerical procedures, especially for Bayesian modeling. On the contrary, it is computationally more efficient to take the alternative approach by evaluating each model with an ad hoc penalized estimator of the out-of-sample discrepancy.

The goal of the study

In this article we investigate and develop the practical model assessment and selection methods for Bayesian models, when we expect that a promising methodology should be objective enough to accept, easy enough to understand, general enough to apply, simple enough to compute and coherent enough to interpret. We mainly restrict attention to the Kullback-Leibler divergence, a widely applied model evaluation measurement to quantify the similarity between the proposed candidate model and the underlying true model, where the true model is only referred to a model as the best projection unto the statistical modeling space once we try to understand the real but unknown dynamics/mechanism of interest. In addition to review and discussion on the advantages and disadvantages of the historically and currently prevailing practical model selection methods in literature, a series of convenient and useful tools, each applied for different purposes, are proposed to asymptotically unbi-

asedly assess how the candidate Bayesian models are favored in terms of predicting a future independent observation. What's more, we also explore the connection of the Kullback-Leibler based information criterion to the Bayes factors, another most popular Bayesian model comparison approaches, after seeing the motivation through the developments of the Bayes factor variants. In general, we expect to provide a useful guidance for researchers who are interested in conducting Bayesian data analysis.

The structure of the article

In Chapter 2, we first introduce the Kullback-Leibler divergence and give a short literature review how it is applied in the frequentist paradigm for model selection. Among the various criteria proposed in the past a few decades, the generalized information criterion (GIC, Konishi and Kitagawa, 1996) is the most promising one in terms of the generality by relaxing the two restrictive assumptions of Akaike Information Criterion (AIC, Akaike, 1973). Considering the fact that many statisticians also evaluate the Bayesian models with point estimators, we review the prevailing Bayesian methods and propose the Bayesian generalized information criterion (BGIC) as a general tool to choose Bayesian models estimated with distinct plug-in parameters. Theoretically, BGIC inherits all the attractive properties of GIC, including the asymptotic unbiasedness and applicable generality. BTIC, the Bayesian version of Takeuchi's information criterion (TIC, Takeuchi, 1976), is illustrated as a special case when we consider the posterior mode as a proper plug-in estimator. Heuristically, the posterior mode plays a similar role as maximum likelihood estimator in the frequentist's setting. A simulation study is conducted to compare the bias correction of BTIC together with other prevalent criteria, such as cross-validation, DIC (Spiegelhalter et al., 2002) and plug-in deviance penalized loss (Plummer, 2008), when the sample size is small and prior distribution is either weakly or strongly informative.

In Chapter 3, we shift our attention to the K-L based predictive criterion for models evaluated by averaging over the posterior distributions of parameters. After reviewing the available criteria, such as the Bayesian predictive information criterion (BPIC, Ando, 2007) and the expected deviance penalized loss (Plummer, 2008), we propose a generally applicable method for comparing different Bayesian statistical models, developed by correcting the asymptotic bias of the posterior mean of the log likelihood as an estimator of its expected log likelihood. Under certain standard regularity conditions, we prove the asymptotic unbiasedness of the proposed criterion even when the candidate models are misspecified. In addition to its appealing large sample properties, we present some numerical comparisons in both normal and binomial cases to investigate the small sample performance. A real data variable selection example is also provided to exhibit the possible difference between the explanatory and predictive approaches.

In Chapter 4, we re-visit the philosophy underneath the Bayes factors after taking a close look at the candidate Bayesian models for pairwise comparison. We demonstrate that, when the standard Bayes factor and its derivatives compare the proposed original models, it is of more interest for Bayesian researchers to make comparisons among the fitted models. Taking the above into account, the predictive Bayes factor is proposed on top of the posterior predictive information criterion (PPIC), both of which are assessed in terms of the Bayesian posterior predictive density. Through the theoretical link between predictive Bayes factor and PPIC, we investigate the significance level of one model outperforming another in accordance to the difference between their information criterion values. For illustrative purpose, we perform the numerical comparison of the predictive Bayes factor with the standard Bayes factor to emphasize the empirical difference.

Conclusion is drawn in the final chapter. Furthermore, we will discuss a few

interesting topics frequently encountered in data analysis when applying Bayesian model selection criteria, and give our suggestions to select the proper Bayesian model.

Notation

Before we go through any technical details, we provide a brief explanation about some of the notation used in this article.

Let $y_1, y_2, \cdots, y_n$ be n independent observations on a random vector y generated from probability distribution F with density function $f(\tilde{y})$, and $\tilde{y}$ a future observation generated from the same true density f, independent of the random vector y. Let y_{-i} denote the leave-one-out random vector $y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n$. An approximating model k is proposed with density $g_k(\tilde{y}|\theta^k)$ among a list of potential models $k = 1, 2, \cdots, K$, and the likelihood function can be written as $L(\theta^k|y) = \prod_{i=1}^{n} g_k(y_i|\theta^k)$. Under the Bayesian setting, the prior distribution of model k is denoted by $\pi_k(\theta^k)$, and the posterior is

$$p_k(\theta^k|y) = \frac{L(\theta^k|y)\pi_k(\theta^k)}{\int L(\theta^k|y)\pi_k(\theta^k)d\theta^k}.$$

Given the prior probabilities $P(M_k)$ for each model, the data y produce the posterior probabilities

$$p_k(\theta^k, M_k|y) = p_k(\theta^k|y)P(M_k|y)$$

For notational purposes, we ignore the model index m hereinafter when there is no ambiguity.

Define $\log \pi_0(\theta) = \lim_{n\to\infty} n^{-1} \log \pi(\theta)$. By the law of large numbers we have $\frac{1}{n} \log\{L(\theta|y)\pi(\theta)\} \to E_{\tilde{y}}[\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}]$ as n tends to infinity. Without specification, the notation $E_{\tilde{y}}$ and E_y in this article exclusively denote the expectation with respect to the underlying true distribution f. Let θ_0, $\hat{\theta}$ denote the expected and

empirical posterior mode of the log unnormalized posterior density $\log\{L(\theta|y)\pi(\theta)\}$,

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

Last, the empirical matrices

$$J_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta\partial\theta'}\right), \tag{1.1}$$

$$I_n(\theta) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\partial \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}\frac{\partial \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta'}\right) \tag{1.2}$$

are considered in our article to unbiasedly estimate the Bayesian Hessian matrix and Bayesian Fisher information matrix

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

Chapter 2

Bayesian Generalized Information Criterion

2.1 Kullback-Leibler divergence: an objective criterion for model comparison

Kullback and Leibler (1951) introduce an information measure, termed as Kullback-Leibler divergence, to assess the directed 'distance' between any two distributions. If we assume $f(\tilde{y})$ and $g(\tilde{y})$ respectively represent the probability density distributions of the 'true model' and the 'approximating model' on the same measurable space, Kullback-Leibler divergence is defined by

$$I(f,g) = \int f(\tilde{y}) \cdot \log \frac{f(\tilde{y})}{g(\tilde{y})} d\tilde{y} = E_{\tilde{y}}[\log f(\tilde{y})] - E_{\tilde{y}}[\log g(\tilde{y})]. \tag{2.1}$$

Note that such a quantity is always non-negative, reaching the minimum value 0 when f is the same as g almost surely, and interpretable as the 'information' lost when g is used to approximate f. Namely, the smaller (and hence the closer to 0) the value of $I(f,g)$, the closer we consider the model g to be to the true distribution.

Without the full knowledge of true distribution f, only the second term of $I(f, g)$ is relevant to compare different possible models in practice. This is because the first term, $E_{\tilde{y}}[\log f(\tilde{y})]$ is a function of f but independent of the proposed model g, and negligible in model comparison for given data $y = (y_1, y_2, \cdots, y_n)$.

As n increases to infinity, the average of log-likelihood

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

tends to $E_{\tilde{y}}[\log g(\tilde{y}|\theta)]$ by the law of large numbers, which gives us hints on how to estimate the second term of $I(f, g)$. Here $\tilde{y}$ is supposed to be an unknown but potentially observable quantity coming from the same distribution f and independent of y, and the second term of $I(f, g)$ is the Kullback-Leibler discrepancy (if interested, see Linhart and Zucchini, 1986, for the discussion why it is improper to call it Kullback-Leibler loss or Kullback-Leibler risk).

The model selection based on Kullback-Leibler divergence is obvious in the simplest case when all the candidate models are parameter-free probability distributions, i.e., $g(\tilde{y}|\theta) = g(\tilde{y})$ when models with large empirical log-likelihood $\frac{1}{n} \log g(y_i)$ are favored. When some unknown parameters θ are contained in the distribution family $g(\tilde{y}|\theta)$, a general procedure is to perform the model fitting first so that we may know what values the parameters most probably will take given the data, and make the model comparison thereafter.

In the frequentist setting, the general model selection procedure starts from choosing one 'best' candidate model specified by some point estimate $\hat{\theta}$ based on a certain statistical principle such as maximum likelihood. There have been a considerable number of references addressing this problem theoretically. For example, assuming the fitted model with MLE $\hat{\theta}$ as the best for family $\mathcal{G} = \{g(\tilde{y}|\theta), \theta \in \Theta\}$, Akaike

(1973) proves that under the assumption $f(\cdot) \in \mathcal{G}$, asymptotically,

$$\frac{1}{n}\sum_{i=1}^{n} \log g(y_i|\hat{\theta}) - K/n \cong E_{\tilde{y}}[\log g(\tilde{y}|\hat{\theta})], \tag{2.2}$$

where the number of parameters K can be considered as the penalty of over-estimating the out of sample log-likelihood. Akaike information criterion (AIC) is defined to be the estimator of (2.2) multiplied by $-2n$. It favors candidate models with small AIC values for the purpose of model selection. Hurvich and Tsai (1989) study the second-order bias adjustment under the normality assumption in small samples. The above two criteria assume that the true model is contained in the candidate class under consideration, an assumption relaxed by Takeuchi's information criterion (TIC) (Takeuchi, 1976). In addition to that, Konishi and Kitagawa (1996) propose the generalized information criterion (GIC) when the parameter estimate $\hat{\theta}$ is not necessarily to be MLE. Meanwhile, Murata et al. (1994) generalize TIC to network information criterion (NIC) Meanwhile, Murata et al.(1994) generalize TIC to network information criterion (NIC) by introducing a discrepancy function to measure the difference between the proposed distribution and the underlying true distribution, where K-L divergence can be considered as a special case. A comprehensive review is given by Burnham and Anderson (2002), when the theoretical discussions on asymptotic efficiency of the AIC-type criteria can be found in Shibata (1981, 1984) and Shao (1997).

2.2 K-L based Bayesian predictive model selection criteria

While there are an abundance of theoretical works in the frequentist's framework, there were no generally applicable K-L based predictive model selection criteria specifically designed for Bayesian modeling in the last century. It used to be prevailing to

apply frequentist criteria such as AIC directly for Bayesian model comparison. However, if we seriously think of the difference of the underlying philosophies between Bayesian and frequentist statistical inference, it is dangerous to make such kind of direct applications by discounting the information within the prior distribution.

The prior works on approaches to use Kullback-Leilber divergence for Bayesian model selection have been considered over last 30 years, for example, see Geisser and Eddy (1979), San Martini and Spezzaferri (1984) and Laud and Ibrahim (1995), while a detailed review is given in Kadane and Lazar (2004) for most of them. However, those methods are either limited in the scope of methodology or computationally infeasible for general Bayesian models, especially when parameters are in hierarchical structures. To find out a good Kullback-Leibler based criterion for Bayesian models, we focus the literature review in this section on most recent or widely applied methods by which model evaluation in terms of plug-in parameter estimators was conducted. The criteria on model evaluation with respect to averaging over parameter posterior distribution will be discussed in next chapter.

DIC

Spiegelhalter et al. (2002) is the most popular paper on this topic, in which they define the deviance information criterion (DIC)

$$DIC = D(\hat{\theta}, y) + 2p_D$$

as an adaptation of the Akaike information criterion for Bayesian models after arguing the plausibility to consider p_D

$$p_D = E_{\theta|y}[D(\theta, y)] - D(\hat{\theta}, y)$$

to estimate the 'effective number of parameters' as a model complexity measure, where the deviance function $D(\theta, y) = -2\sum_i \log g(y_i|\theta)$. Here $\hat{\theta}$ could be either the

posterior mean or mode instead of MLE since the full model specification of Bayesian statistics contains a prior specification $\pi(\theta)$ in addition to the likelihood, and the inference can only be derived from the posterior distribution $p(\theta|y) \propto L(\theta|y)\pi(\theta)$.

Spiegelhalter et al. (2002) heuristically demonstrate that, as a model selection criterion, $-DIC/2n$ estimates the expected expected out-of-sample log-likelihood $\eta_1 = E_{\theta^t} E_{\tilde{y}|\theta^t} \left[\log g(\tilde{y}|\hat{\theta})\right]$, where θ^t are considered as the true parameters after assuming that the proposed model encompasses the true model with $\hat{\theta} \to \theta^t$. However, the estimation is pointed out to be lack of theoretical foundation by various researchers, for instance, Meng and Vaida (2006) in a radical prognosis and Celeux et al. (2006b) in an agreement.

In practice, DIC is simple to calculate after deriving the posterior samples by using Markov chain Monte Carlo (MCMC) method. An open-source software developed for its computation is BUGS (Spiegelhalter et al., 1994; 2003), while JAGS (Plummer, 2007) provides an alternative approach of estimation by using importance sampling method.

cAIC

To evaluate the goodness of the Gaussian linear mixed-effects models for clustered data under the normality assumption

$$\begin{aligned} g(y|\beta, b) &= N(X\beta + Zb; \sigma^2) \\ p(b) &= N(0; \Sigma), \end{aligned}$$

Vaida and Blanchard (2005) propose the conditional Akaike's information criterion (cAIC). One of their important assumptions is that the true model is within the class of approximating parametric probability distributions, that is, there exist a pair of β_0 and random effects u with distribution $p(u)$ which satisfy $f(y|u) = g(y|\beta_0, u)$. The

conditional Akaike information

$$\eta_2 = E_u E_{y|u} E_{\tilde{y}|u} \left[\log g(\tilde{y}|\hat{\beta}, \hat{b}) \right]$$

is treated as the adjusted target function of model selection criterion when replacing the true density in the expected out-of-sample log likelihood by its parametric estimation, where $\hat{\beta}$, $\hat{b}$ are the maximum likelihood estimator for β and the empirical Bayes estimator for b respectively, and the expectation is over the posterior predictive distribution $p(\tilde{y}|y) = \int g(\tilde{y}|\hat{\beta}, b)p(b|y)db$.

When the variance σ^2 and covariance matrix Σ are known,

$$cAIC = -2 \sum_i \log g(y_i|\hat{\beta}, \hat{b}) + 2\rho$$

is proved to be unbiased for $-2n \cdot \eta_2$. where $\rho = tr(H)$, and H is the 'hat' matrix mapping the observed y onto the fitted $\hat{y}$ such that $\hat{y} = X\hat{\beta} + Z\hat{b} = Hy$. It is worth mentioning that ρ is considered to be a measure of degrees of freedom for mixed effects models by Hodges and Sargent (2001). With an interpretation in terms of subspace geometrical projection, they argue that the complexity of the random effects is a 'fraction' of the number of parameters because of the constraints from the hyper-level covariance.

Liang et al. (2009) generalize cAIC by removing the assumption on variance. They prove that, when both the variance and the covariance matrix of the linear mixed-effects model are unknown, an unbiased estimator for $-2n \cdot \eta_2$ is

$$GcAIC = -2 \sum_i \log g(y_i|\hat{\beta}, \hat{b}) + 2\Phi_0(y),$$

where $\Phi_0(y) = \sum_{i=1}^{N} \partial \hat{y}_i / \partial y_i$ is inherent to the generalized degrees of freedom defined by Ye (1998).

Cross-validation

Cross-validation (Stone, 1974; Geisser, 1975) is an algorithm to assess the out-of-sample performance of a discrepancy function, which could be either the log-likelihood evaluated with plug-in estimator or averaging over the posterior in our case. For example, Geisser and Eddy (1979) provide a cross-validative approach to estimate posterior predictive density for Bayesian model selction. Stone (1977) shows that cross-validation is asymptotically equivalent to the AIC in the order of $o_p(1)$. The comprehensive review on the recent development of cross-validation in the frequentist paradigm is given by Arlot and Celisse (2010), whereas Vehtari and Lampinen (2002) explore the application of the cross-validation procedure to estimate expected utilities for Bayesian models.

The computation of the cross-validation estimate is always challenging for Bayesian modeling. For leave-one-out cross-validation, every candidate model need to be re-fitted for n times to generate a series of the posterior $p(\theta|y_{-i})$, each with a single observation i deleted. The process will be unfeasibly time-consuming for iterative computation. An alternative is to use importance sampling when the posterior given the full data $p(\theta|y)$ is chosen as the sampling proposal; however, this is still not a good solution because the weights, $1/g(y_i|\theta)$, are unbounded, making the importance-weighted estimate unstable. The k-fold cross-validation can reduce the number of re-fitting, but the variance of the expected utility estimate will increase for small k even after the additional higher-order bias correction, and the error of K-folded cross-validation is generally over-estimated (Vehtari and Lampinen, 2002).

Plug-in/Expected Deviance Penalized Loss

With the definition of 'optimism'

$$p_{opt_i} = E\{L(y_i, y_{-i}) - L(y_i, y)|y_{-i}\}$$

for observation i, Plummer (2008) shows that the penalized loss

$$L(y_i, y) + p_{opt_i}$$

has the same conditional expectation over the predictive density $p(y_i|y_{-i})$ as the cross-validation loss $L(y_i, y_{-i})$. After considering 2 special loss functions each based on distinct treatment to the deviance function: the 'plug-in deviance'

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

and the 'expected deviance'

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

the total penalized loss

$$L(y, y) + p_{opt} = \sum_{i=1}^{n} \{L(y_i, y) + p_{opt_i}\}$$

is in the form of a K-L based model selection criterion, where p_{opt} is the bias correction term. It is worth to mention that by employing the predictive density $p(y_i|y_{-i})$ in the conditional expectation, the assumption of the true model contained in the approximating family is added compared with the cross-validation method. In principle, the expected deviance penalized loss is a special case of the predictive discrepancy measure (Gelfand and Ghosh 1998).

A numerical solution similar to BUGS but designed to estimate both the DIC and the deviance penalized loss is provided by JAGS (Plummer, 2007). Other than Gibbs sampler and Metropolis algorithm, JAGS uses importance sampling to draw samples from full posterior $p(\theta|y)$ for leave-one-out posterior $p(\theta|y_{-i})$. One caveat is that importance sampling algorithm may cause inaccurate estimation in practice if some observation y_i was influential, as illustrated in the discussion of cross-validation.

2.3 Bayesian Generalized Information Criterion

Note that maximization of the expected out-of-sample log likelihood is equivalent to minimization of the Kullback-Leilber divergence. To estimate the expected out-of-sample log likelihood, all the approaches listed above consider the empirical log likelihood as a proxy. However, different approaches employ different bias corrections to compensate the double use of the data for both model estimation and evaluation.

Besides the computationally costly cross-validation method, the estimations of the bias correction term from the rest methods are derived under the assumption that the candidate model is not mis-specified, i.e., the true model f is contained in the proposed parametric family. It makes the usage of those approaches very challenging, for such a strong assumption is almost impossible to verify. If that assumption were not met, the interpretation of the estimated criteria values could significantly mislead the conclusion.

Another key element in the bias correction term estimation is about the selection of the plug-in estimator in model assessment. From the literature reviews in the previous section, one may find that some potential candidates of plug-in estimators are posterior mean and posterior mode. However, no theoretical foundation has been built and no agreement settled on the choice of the plug-in estimator for the parametric distribution.

To develop a generalized model selection criterion for Bayesian modeling without those shortages, we consider the estimation of bias correction based on functional-type estimators and the corresponding functional Taylor series expansion (Huber, 1981; Hampel et al., 1986), when the idea to employ the functional estimator in model selection is introduced by Konishi and Katagawa (1996) for frequentist modeling. In the following theorem, a bias estimator of the discrepancy between the true model against the fitted model is proposed and its asymptotic unbiasedness is proved.

Theorem 2.1. *Let $y = (y_1, y_2, \cdots, y_n)$ 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 not necessarily containing the true distribution f, where $\theta = (\theta_1, ..., \theta_p)'$ is the p-dimensional vector of unknown parameters, with prior distribution $\pi(\theta)$. Assume a statistical functional $T(\cdot)$ is both second-order compact differentiable at F and Fisher consistent, i.e., $T(G) = \theta$ for all $\theta \in \Theta$. The asymptotic bias of*

$$\hat{\eta} = \int \log g(\tilde{y}|\hat{\theta}) d\hat{F}(\tilde{y}) = \frac{1}{n} \sum_i \log g(y_i|\hat{\theta})$$

in the estimation of

$$\eta = \int \log g(\tilde{y}|\hat{\theta}) dF(\tilde{y}) = E_{\tilde{y}} g(\tilde{y}|\hat{\theta})$$

can be unbiasedly approximated by

$$E_y(\hat{\eta} - \eta) = b(F) + o_p(n^{-1}), \tag{2.3}$$

where

$$b(F) = \frac{1}{n} tr\{E_{\tilde{y}}[T^{(1)}(\tilde{y}; F)' \frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial \theta}|_{T(F)}]\} \tag{2.4}$$

and $T^{(1)}(\tilde{y}; F) = (T_1^{(1)}(\tilde{y}; F), \ldots, T_p^{(1)}(\tilde{y}; F))'$ is the influence function of a p-dimensional functional $T(F)$ at the distribution F.

The derivation of Theorem 2.1 in spirit is similar to Theorem 2.1 of Konishi and Katagawa (1996) but now adjusted to the setting for a Bayesian probability model.

Proof of Theorem . The functional Taylor series expansion for vector $\hat{\theta} = T(\hat{F})$ is, up to order n^{-1},

$$\hat{\theta} = T(F) + \frac{1}{n} \sum_{i=1}^{n} T^{(1)}(y_i; F) + \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} T^{(2)}(y_i, y_j; F) + o_p(n^{-1}) \tag{2.5}$$

where $T^{(k)}(y_1, \ldots, y_k; F)$ is the same as defined on p.888 of Konishi and Katagawa (1996) with property

$$E_{y_1,\ldots,y_k} T^{(k)}(y_1, \ldots, y_k; F) = 0.$$

It is simple to derive the asymptotic property of $\hat{\theta}$ through (2.5):

$$\begin{aligned} E_{\tilde{y}}(\hat{\theta}) &= T(F) + \xi/n + o_p(n^{-1}), \\ Cov(\hat{\theta}) &= \frac{1}{n}\Sigma(T(F)), \end{aligned}$$

where $\xi = \frac{1}{2}E_{\tilde{y}}T^{(2)}(\tilde{y},\tilde{y};F)$ and $\Sigma(T(F)) = E_{\tilde{y}}[T^{(1)}(\tilde{y};F)T^{(1)}(\tilde{y};F)']$.

By expanding $\log\{g(\tilde{y}|\hat{\theta})\pi^{\frac{1}{n}}(\hat{\theta})\}$ in a Taylor series around $\theta = T(G)$ and substituting (2.5) in the resulting expansion, the stochastic expansions for η and $\hat{\eta}$ are given as follows:

$$\begin{aligned} \eta &= E_{\tilde{y}}\log\{g(\tilde{y}|\hat{\theta})\pi^{\frac{1}{n}}(\hat{\theta})\} - \frac{1}{n}\log\pi(\hat{\theta}) \\ &= E_{\tilde{y}}\log\{g(\tilde{y}|T(F))\pi^{\frac{1}{n}}(T(F))\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} T^{(1)}(y_i;F)'\kappa + \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} T^{(2)}(y_i,y_j;F)'\kappa \\ &\quad - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} T^{(1)}(y_i;F)'J(T(F))T^{(1)}(y_j;F) - \frac{1}{n}\log\pi(\hat{\theta}) + o_p(n^{-1}), \\ \hat{\eta} &= \frac{1}{n}\sum_{i=1}^{n}\log\{g(y_i|\hat{\theta})\pi^{\frac{1}{n}}(\hat{\theta})\} - \frac{1}{n}\log\pi(\hat{\theta}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\{g(y_i|T(F))\pi^{\frac{1}{n}}(T(F))\} \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} T^{(1)}(y_i;F)'\frac{\partial\log\{g(y_j|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}|_{T(F)} \\ &\quad + \frac{1}{2n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\{T^{(2)}(y_i,y_j;F)'\frac{\partial\log\{g(y_k|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}|_{T(F)} \\ &\quad + T^{(1)}(y_i;F)'\frac{\partial^2\log\{g(y_k|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta\partial\theta'}|_{T(F)}T^{(1)}(y_j;F)\} - \frac{1}{n}\log\pi(\hat{\theta}) + o_p(n^{-1}) \end{aligned}$$

Taking expectations term by term yields

$$\begin{aligned} E_y\ \eta &= E_{\tilde{y}} \log\{g(\tilde{y}|T(F))\pi^{\frac{1}{n}}(T(F))\} + \frac{1}{n}[\xi'\kappa - \frac{1}{2}tr\{\Sigma(F)J(T(F))\}] \\ &\quad -\frac{1}{n}E_y \log \pi(\hat{\theta}) + o_p(n^{-1}) \qquad (2.6) \\ E_y\ \hat{\eta} &= E_{\tilde{y}} \log\{g(\tilde{y}|T(F))\pi^{\frac{1}{n}}(T(F))\} + \frac{1}{n}E_{\tilde{y}}[T^{(1)}(\tilde{y};F)'\frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}|_{T(F)}] \\ &\quad +\frac{1}{n}[\xi'\kappa - \frac{1}{2}tr\{\Sigma(F)J(T(F))\}] - \frac{1}{n}E_y \log \pi(\hat{\theta}) + o_p(n^{-1}) \qquad (2.7) \end{aligned}$$

where κ and $J(\theta)$ are given by

$$\kappa = E_{\tilde{y}}\frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}|_{T(F)}, \qquad J(\theta) = -E_{\tilde{y}}\frac{\partial^2 \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta\partial\theta'}.$$

By directly comparing (2.6) and (2.7), we complete the proof. □

In practice, an estimator of the true bias in (2.4) is $b(\hat{F})$ when replacing the unknown true distribution F by the empirical distribution $\hat{F}$. Subsequently we have an information criterion based on the bias corrected log likelihood as follows:

$$\mathrm{BGIC}(y;\hat{F}) := -2\sum_{i=1}^{n}\log g(y_i|\hat{\theta}) + \frac{2}{n}\sum_{i=1}^{n} tr\{T^{(1)}(y_i;\hat{F}(y))\frac{\partial \log\{g(y_i|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta}|_{\hat{\theta}}\},$$

where $T^{(1)}(y_i;\hat{F}) = (T_1^{(1)}(y_i;\hat{F}),\ldots,T_p^{(1)}(y_i;\hat{F}))'$ is the p-dimensional empirical influence function defined by

$$T_j^{(1)}(y_i;\hat{F}) = \lim_{\varepsilon\to 0}[T_j((1-\varepsilon)\hat{F} + \varepsilon\delta(y_i)) - T_j(\hat{F})]/\varepsilon,$$

with $\delta(y_i)$ being the Dirac delta function with the point mass at y_i. When choosing among different models, we select the model for which the value of the information criterion $\mathrm{BGIC}(y;\hat{F})$ is small.

The new criterion is a model selection device specially designed to evaluate Bayesian models, when the prior distribution is properly incorporated into the bias correction. Benefiting from adopting the functional Taylor series expansion, it asymptotically unbiasedly estimates the over-estimation bias of the empirical log likelihood. The

criterion is widely applicable for models specified by any functional-type estimator $\hat{\theta}$, even when candidate models are misspecified.

Regularly, a functional estimator of interest for Bayesian models is the posterior mode $\hat{\theta} = T_m(\hat{F})$. In this case, the influence function vector is

$$T_m^{(1)}(\tilde{y}; F) = J^{-1}(T_m(F)) \frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial \theta}|_{T_m(F)} \tag{2.8}$$

where

$$J(\theta) = -E_{\tilde{y}} \frac{\partial^2 \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial \theta \partial \theta'}.$$

Substituting the influence function $T_m^{(1)}(\tilde{y}; F)$ given by (2.8) into the result of (2.4) yields the asymptotic bias $b_m(F) = tr\{J^{-1}(T_m(F))I(T_m(F))\}$, where

$$I(\theta) = E_{\tilde{y}}[\frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial \theta} \frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial \theta'}],$$

which induces the following corollary:

Corollary 2.2. *Let $y = (y_1, y_2, \cdots, y_n)$ 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 not necessarily containing the true distribution f, where $\theta = (\theta_1, ..., \theta_p)'$ is the p-dimensional vector of unknown parameters, with prior distribution $\pi(\theta)$. Under the regularity conditions:*
C1: Both the log density function $\log g(\tilde{y}|\theta)$ and the log unnormalized posterior density $\log\{L(\theta|y)\pi(\theta)\}$ 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 asymptotic bias of $\hat{\eta} = \frac{1}{n}\sum_i \log g(y_i|\hat{\theta})$ for $\eta = E_{\tilde{y}} \log g(\tilde{y}|\hat{\theta})$ can be unbiasedly approximated by $\frac{1}{n} tr\{J^{-1}(\theta_0)I(\theta_0)\}$.

Correspondingly, we derive a Bayesian version of Takeuchi information criterion (BTIC):

$$-2\sum_i \log g(y_i|\hat{\theta}) + 2tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} \tag{2.9}$$

when the candidate models with small criterion values are preferred on the purpose of model selection. Here $\hat{\theta}$ is the posterior mode which minimizes the posterior distribution $\propto \pi(\theta)\prod_{i=1}^n g(y_i|\theta)$, and matrices $J_n(\theta)$ and $I_n(\theta)$ are empirical estimators of Bayesian Hessian matrix $J(\theta)$ and Bayesian Fisher information matrix $I(\theta)$, respectively.

2.4 A simple linear example

The following simple simulation example demonstrates both the importance of introducing K-L based criteria for Bayesian modeling and the efficiency of the proposed criterion in the estimation of bias correction.

Suppose observations $y = (y_1, y_2, ..., y_n)$ 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 is 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.

It is easy to derive the posterior distribution of μ which 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, we have

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

To eliminate the estimation error caused by the sampling of the observations y, we average the bias $\hat{\eta}-\eta$ over y with its true density $N(\mu_T,\sigma_T^2)$,

$$\begin{aligned} b_\mu &= E_y(\hat{\eta}-\eta) = E_y\{\frac{\sigma_T^2}{2\sigma_A^2}+\frac{(\mu_T-\hat{\mu})^2}{2\sigma_A^2}-\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})^2}{2\sigma_A^2}\} \\ &= \frac{\sigma_T^2}{2\sigma_A^2}+\frac{(\mu_T-\mu_0)^2/\tau_0^4+n\sigma_T^2/\sigma_A^4}{2\sigma_A^2(\frac{1}{\tau_0^2}+\frac{n}{\sigma_A^2})^2} \\ &\quad -\frac{1}{2\sigma_A^2(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2})^2}\left[\frac{(\mu_T-\mu_0)^2+\sigma_T^2}{\tau_0^4}+\frac{2(n-1)\sigma_T^2}{\tau_0^2\sigma_A^2}+\frac{n(n-1)\sigma_T^2}{\sigma_A^4}\right] \\ &= \frac{\sigma_T^2}{2\sigma_A^2}+\frac{-\sigma_T^2/\tau_0^4-2(n-1)\sigma_T^2/\tau_0^2\sigma_A^2+n(2-n)\sigma_T^2/\sigma_A^4}{2\sigma_A^2(1/\tau_0^2+n/\sigma_A^2)^2} \\ &= \sigma_T^2\hat{\sigma}^2/\sigma_A^4. \end{aligned}$$

Here we compare the bias estimator of BTIC, b_μ^{BTIC} with 6 other bias estimators: b_μ^{AIC} Akaike (1973), b_μ^{TIC} Takeuchi (1976), b_μ^{DIC} (Spiegelhalter et al., 2002), b_μ^{cAIC} (Vaida and Blanchard, 2005), $b_\mu^{PL^p}$ (Plummer, 2008) and b_μ^{CV} (Stone, 1974).

$$\begin{aligned} b_\mu^{BTIC} &= \frac{1}{n-1}\hat{\sigma}^2\sum_{i=1}^{n}((\mu_0-\hat{\mu})/(n\tau_0^2)+(y_i-\hat{\mu})/\sigma_A^2)^2 \\ b_\mu^{AIC} &= 1 \\ b_\mu^{TIC} &= \frac{1}{n-1}\sum_{i=1}^{n}((y_i-\hat{\mu})^2)/\sigma_A^2 \\ b_\mu^{DIC} &= b_\mu^{cAIC}=\hat{\sigma}^2/\sigma_A^2 \\ b_\mu^{PL^p} &= \frac{1}{2n}p_{opt}^p=(\hat{\sigma}^2+1/(1/\tau_0^2+(n-1)/\sigma_A^2))/\sigma_A^2/2 \\ b_\mu^{CV} &= \hat{\eta}-\sum_{i=1}^{n}(y_i-(\mu_0/\tau_0^2+\sum_{j\neq i}y_j/\sigma_A^2)/(1/\tau_0^2+(n-1)/\sigma_A^2))^2/n/\sigma_A^2/2 \end{aligned}$$

under the settings of 6 different scenarios:

1. $\sigma_T=1$; τ_0=100, $\sigma_A=1$;

2. $\sigma_T=1$; τ_0=0.5, $\sigma_A=1$;

3. $\sigma_T=1$; τ_0=100, $\sigma_A=1.5$;

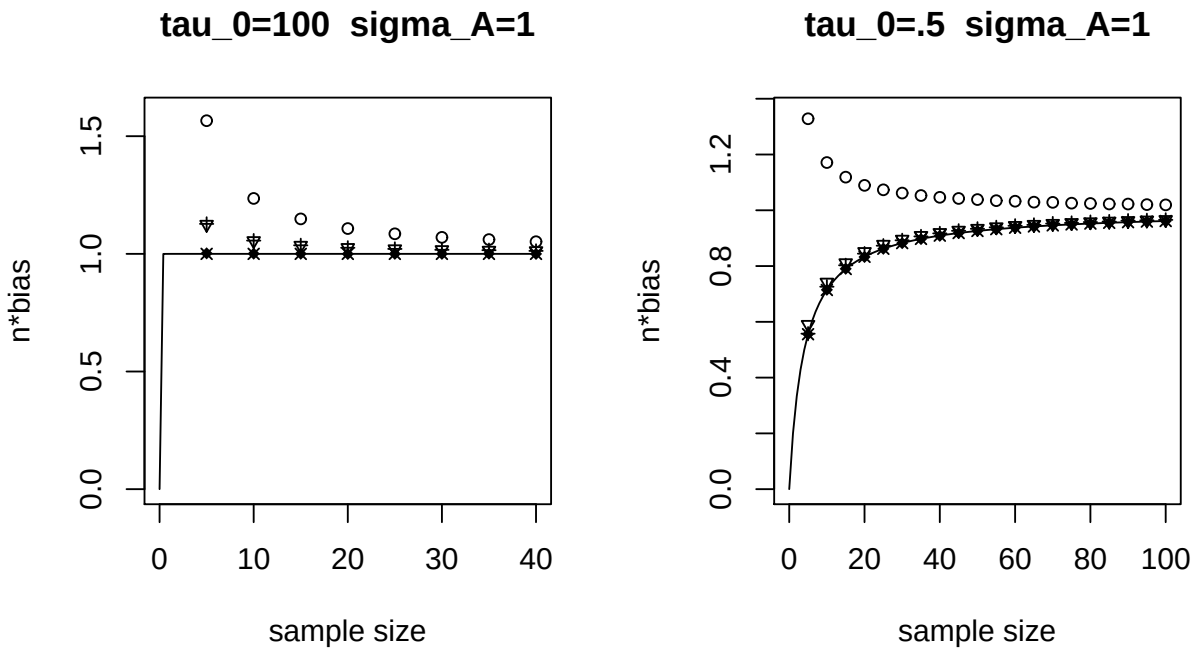

Figure 2.1: Performance of the estimators for $E_y(\hat{\eta} - \eta)$ when $\sigma_A^2 = \sigma_T^2 = 1$, i.e., the true distribution is contained in the candidate models. The left plot is under a relatively non-informative prior with $\tau_0 = 100$; the right plot is under a relatively informative prior with $\tau_0 = 0.5$. The true bias is curved by (——) as a function of sample size n. The averages of different bias estimators are marked by: ($\bullet$) for BTIC; ($\circ$) for TIC; ($\times$) for DIC and cAIC; (∇) for PL^p; (+) for cross-validation. Each mark represents the mean of estimated bias of 250,000 replications.

4. $\sigma_T = 1$; τ_0=0.5, $\sigma_A = 1.5$;

5. $\sigma_T = 1$; τ_0=100, $\sigma_A = 0.5$;

6. $\sigma_T = 1$; τ_0=0.5, $\sigma_A = 0.5$.

which include the cases with models exposed to either very weakly-informative or informative priors, and the true model may or may not contained in the approximating distribution family.

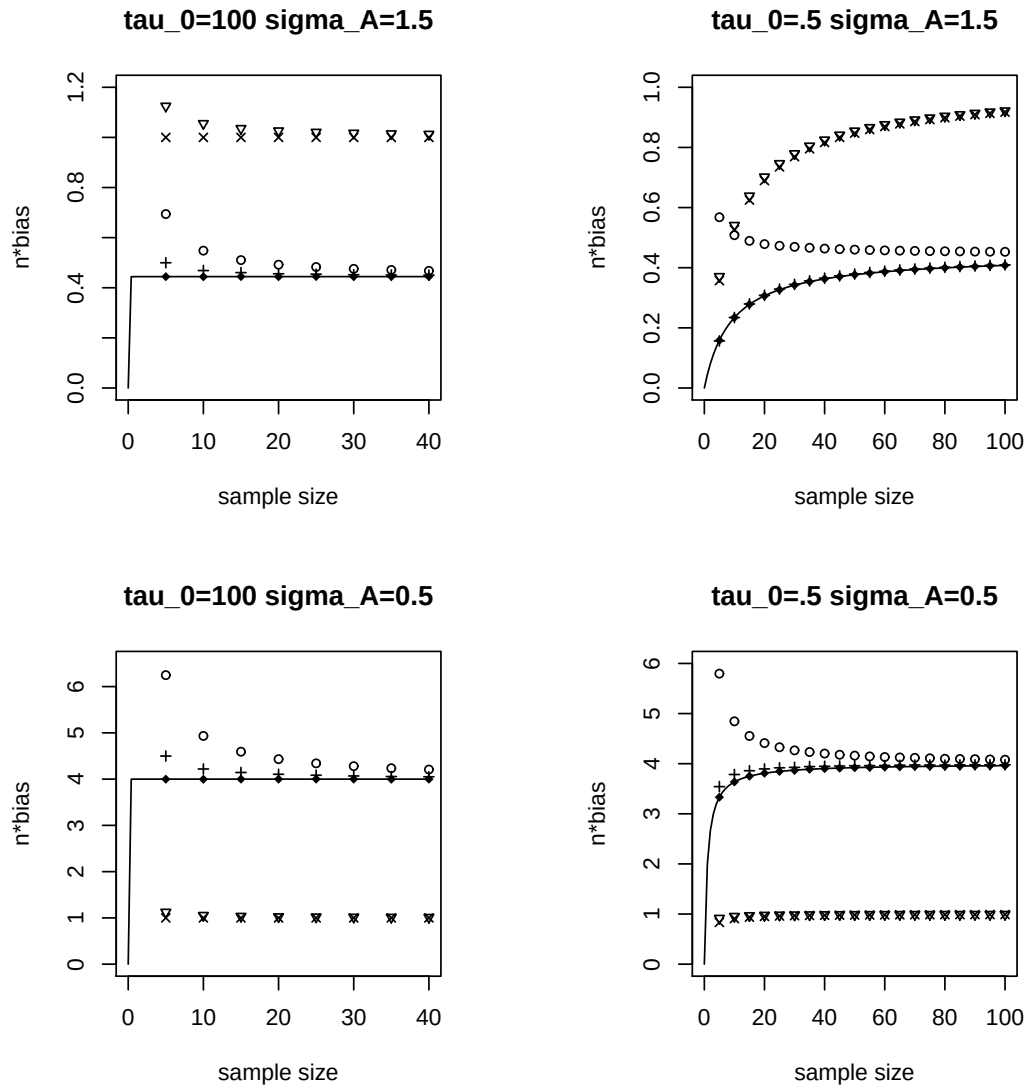

Figure 2.2: Performance of the estimators for $E_y(\hat{\eta} - \eta)$ when true model is not contained in the candidate distributions. The left two plots are under a relatively non-informative prior with $\tau_0 = 100$; the right ones are under a relatively informative prior with $\tau_0 = 0.5$. The true bias is curved by (——) as a function of sample size n. The averages of different bias estimators are marked by: ($\bullet$) for BTIC; ($\circ$) for TIC; ($\times$) for DIC and cAIC; (∇) for PL^p; ($+$) for cross-validation. Each mark represents the mean of estimated bias of 250,000 replications.

Computationally, we simply replicate the process:

1. Draw a vector of length n observations y from the true distribution $N(\mu_T, \sigma_T^2)$.

2. Generate the posterior draws from the posterior distribution of $\mu|y$.

3. Estimate b_μ^{BTIC}, b_μ^{TIC}, b_μ^{DIC}, b_μ^{cAIC}, $b_\mu^{PL^p}$ and b_μ^{CV}. (Here b_μ^{AIC} is constant 1.)

The true mean and the prior mean are set to be $\mu_T = 0$ and $\mu_0 = 0$, respectively, and the prior variances are set to be either the informative $\tau_0^2 = (.5)^2$ or non-informative $\tau_0^2 = (100)^2$. After 250,000 replications for each pre-specified n and the averages of the bias estimators are plotted in Figure 2.1 for the case $\sigma_A^2 = \sigma_T^2$ and Figure 2.2 when the equality between σ_A^2 and σ_T^2 does not hold.

The results, are in accordance with theory. All of the 7 estimates are close to the true bias-correction values when $\sigma_A^2 = \sigma_T^2 = 1$, especially when the sample size becomes moderately large. The estimated criterion values based on the AIC, TIC, DIC and BTIC are consistently closer to the true values than cross-validation and plug-in deviance penalized loss, which overestimate the bias, especially when sample size is small. When the models are misspecified, it is not surprising that in all of the plots given in Figure 2.2, the estimates based on DIC, cAIC and plug-in deviance penalized loss all miss the target even asymptotically since their assumption is violated, whereas both the BTIC and cross-validation converge to b_μ. Generally, BTIC out-performs the others.

Chapter 3

Posterior Averaging Information Criterion

'... we concede that using a plug-in estimate disqualifies the technique from being properly Bayesian.'

Celeux et al. (2006) p.703.

3.1 Model evaluation with posterior distribution

In this chapter we focus on Kullback-Leibler divergence based model selection methods with respect to Bayesian models evaluated by averaging over the posterior distribution. Unlike the preceding methods which assess the model performance in terms of the similarity between the true distribution f with the model density function specified by the plug-in parameters, approaches developed to estimate the posterior averaged K-L discrepancy, i.e., the expected out-of-sample log likelihood $E_{\tilde{y}} \log g(\tilde{y}|\theta)$ averaged over the posterior distribution $p(\theta|y)$, are investigated.

The attention to the posterior averaged K-L discrepancy has been paid by some Bayesian researchers in recent years. Ando (2007) makes an important contribu-

tion to the literature by proposing an estimator for the discrepancy measure $\eta_3 = E_{\tilde{y}}[E_{\theta|y} \log g(\tilde{y}|\theta)]$ in terms of K-L divergence. Plummer's paper (Plummer, 2008), which was reviewed in the previous chapter, introduces the expected deviance penalized loss in a cross-validation perspective. The standard cross-validation method can also be applied in this circumstance to estimate η_3, simply by considering the K-L discrepancy as the utility function of Vehtari and Lampinen (2002). The estimation of bootstrap error correction $\eta_3^{(b)} - \hat{\eta}_3^{(b)}$ with bootstrap analogues

$$\eta_3^{(b)} = E_{\tilde{y^*}}[E_{\theta|y^*} \log g(\tilde{y}|\theta)]$$

and

$$\hat{\eta}_3^{(b)} = E_{\tilde{y^*}}[n^{-1} E_{\theta|y^*} \log L(\theta|y^*)]$$

for $\eta_3 - \hat{\eta}_3$ has been discussed in Ando (2007) as a Bayesian adaption to frequentist's model selection (Konish and Kitagawa, 1996).

The application of the posterior averaging approaches for Bayesian model comparison is consistent with Bayesian philosophy. Other than some unknown but fixed values implied in frequentist's inference, the vector of parameters in the Bayesian perspective are considered as a (multi-variate) random variable represented by a probability distribution. A comprehensive representation of the Bayesian inference should be based on posterior distribution. As the starting sentence in Gelman et al. (2003) states:

'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.'*

Correspondingly, instead of considering model specified by a point estimate, the 'goodness' of a Bayesian candidate model in terms of prediction should be measured against the posterior distribution, in which case η_3 is much more favorable.

Usually the computation of the posterior averaged K-L discrepancy is quite intensive, especially in the case that a large set of posterior samples are in need for numerical averaging; whereas the computation of the K-L discrepancy specified by the plug-in estimators is relatively straightforward. However, we consider it as a worthy price, mainly with regard to the methodology of Bayesian model selction, when the computational cost becomes more and more acceptable due to the popularity of modern computers.

For notational simplicity, we rename η_3 to η thereinafter within this chapter.

3.2 Posterior Averaging Information Criterion

Bayesian statistical conclusions about a parameter θ, or unobserved data $\tilde{y}$, are made in terms of the probability statements. Accordingly, it is natural to consider the posterior average over the K-L discrepancy $\eta = E_{\tilde{y}}[E_{\theta|y} \log g(\tilde{y}|\theta)]$ to measure the deviation of the approximating model from the true model.

One substantial proposal in literature on this topic is the Bayesian predictive information criterion (BPIC). Under certain regularity conditions, Ando (2007) proves that an asymptotic unbiased estimator of η is

$$\hat{\eta}^{BPIC} = \frac{1}{n}[\log\{\pi(\hat{\theta})L(\hat{\theta}|y)\} - E_{\theta|y} \log \pi(\theta) - tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\} - \frac{K}{2}]. \tag{3.1}$$

Here $\hat{\theta}$ is the posterior mode, K is the cardinality of θ, and matrices J_n and I_n are empirical estimators for Bayesian Hessian matrix

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

and Bayesian Fisher information matrix

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

BPIC is introduced as $-2n \cdot \hat{\eta}^{BPIC}$ and model with minimum BPIC values is favored.

Compared with other numerical estimators of η, (3.1) is fast to compute and applicable when the true model is not necessarily in the specified family of probability distributions. However, it has the following unpleasant features in practice.

- BPIC is undefined when the prior distribution $\pi(\theta)$ is degenerate, a situation commonly occurred in Bayesian analysis when objective non-informative prior is selected.

- The natural estimator for η is not $n^{-1} \log L(\hat{\theta}|y)$, but $n^{-1} E_{\theta|y} \log L(\theta|y)$. The usage of $n^{-1} L(\hat{\theta}|y)$ will reduce the estimation efficiency if the posterior distribution is asymmetric, which occurs in a majority of cases in Bayesian modeling.

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

Note that the entire data y are used for both model fitting and model selection, $\hat{\eta}$ always over-estimates η. In order to correct the estimation bias, the following theorem is derived for the data over-usage.

Theorem 3.1. *Let $y = (y_1, y_2, \cdots, y_n)$ 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 not necessarily containing the true distribution f, where $\theta = (\theta_1, ..., \theta_p)'$ is the p-dimensional vector of unknown parameters, with prior distribution $\pi(\theta)$. Under the regularity conditions: C1: Both the log density function $\log g(\tilde{y}|\theta)$ and the log unnormalized posterior density $\log\{L(\theta|y)\pi(\theta)\}$ 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;

asymptotically the bias of $\hat{\eta}$ *for* η *can be unbiasedly approximated by*

$$E_y(\hat{\eta} - \eta) = b_\theta \approx \frac{1}{n} tr\{J_n^{-1}(\hat{\theta}) I_n(\hat{\theta})\}, \tag{3.2}$$

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

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

With the above result, a new predictive criterion for Bayesian model, the posterior averaging information criterion (PAIC) is proposed as:

$$-2\sum_i E_{\theta|y}[\log g(y_i|\theta)] + 2tr\{J_n^{-1}(\hat{\theta}) I_n(\hat{\theta})\} \tag{3.3}$$

The candidate models with small criterion values are preferred on the purpose of model selection.

The proposed criterion has many attractive properties. It is an objective model selection criterion consistent with Bayesian philosophy. It is asymptotically unbiased for the out-of-sample log-likelihood, a measure in terms of K-L divergence for the similarity of the fitted model and the underlying true distribution. The estimation averaged over the posterior is more precise and robust than any plug-in based estimators especially when the posterior distribution of parameters is asymmetric, a normal situation especially when parameters are hierarchical. Because it is derived free of the assumption on the approximating distributions containing the truth, our criterion is generally applicable. Unlike BPIC, the new criterion is well-defined and can cope with degenerate non-informative prior distribution for parameters.

In contrast to frequentist modeling, it is inevitable to include a prior distribution for parameters in each Bayesian model, either informative or non-informative, representing the current believing on the parameters independent of the given set of data. Subsequently, the ad hoc statistical inference depends on the posterior distribution $p(\theta|y) \propto L(\theta|y)\pi(\theta)$ other than the likelihood function $L(\theta|y)$ alone; the choice of the prior distribution may cause a strong impact. Specifically, that impact for model selection in our case is not limited to the posterior averaging over the discrepancy function, but to the extent how much the error of the in-sample estimator is corrected. Especially when the prior knowledge is substantive from reliable resources, the specification of $J_n(\theta)$ and $I_n(\theta)$ may depend on $\pi(\theta)$ significantly, as well as the posterior mode on which both matrices are assessed.

The proof of Theorem 3.1 is given in the rest of the section. We start with a few lemmas to support the main proof.

Lemma 1a. Under the same regularity conditions of Theorem 3.1, $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically approximated by $N(0, J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0))$.

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

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

If the parameter $\hat{\theta}$ is the mode of $\log\{L(Y|\theta)\pi(\theta)\}$ and satisfies $\frac{\partial \log\{L(Y|\theta)\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}} = 0$, then

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

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

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

□

Lemma 1b. Under the same regularity conditions of Theorem 3.1, $\sqrt{n}(\theta - \hat{\theta}) \sim N(0, J_n^{-1}(\hat{\theta}))$.

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

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

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

Consider it as a function of θ, the first term of RHS is a constant, whereas the second term is proportional to the logarithm of a normal density, yielding the approximation of the posterior distribution for θ:

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

Note that a formal but less intuitive proof can be obtained by applying Berstein-Von Mises theorem directly. □

Lemma 1c. Under the same regularity conditions of Theorem 3.1, $E_{\theta|y}(\theta_0 - \hat{\theta})(\hat{\theta} - \theta)' = o_p(n^{-1})$.

Proof. First we have

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

$\hat{\theta}$, the mode of $\log\{L(\theta|y)\pi(\theta)\}$, satisfies $\frac{\partial \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}} = 0$, yielding

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

Note that

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

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

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

Together with $\theta_0 - \hat{\theta} = O_p(n^{-1/2})$ derived from lemma 1a, we get the desired result. □

Lemma 1d. Under the same regularity conditions of Theorem 3.1, $E_{\theta|y}(\theta_0 - \theta)(\theta_0-\theta)' = \frac{1}{n}J_n^{-1}(\hat{\theta}) + \frac{1}{n}J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0) + o_p(n^{-1})$.

Proof. $E_{\theta|y}(\theta_0-\theta)(\theta_0-\theta)'$ can be rewritten as $(\theta_0-\hat{\theta})(\theta_0-\hat{\theta})' + E_{\theta|y}(\hat{\theta}-\theta)(\hat{\theta}-\theta)' + 2E_{\theta|y}(\theta_0-\hat{\theta})(\hat{\theta}-\theta)$. Using Lemma 1a, 1b and 1c, we obtain the desired result. □

Lemma 1e. Under the same regularity conditions of Theorem 3.1,

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

Proof. $1/n$ of the posterior mean over the log joint density distribution of (y,θ) can be Taylor-expanded around θ_0 as:

$$\begin{aligned} E_{\theta|y}\frac{1}{n}\log\{L(\theta|y)\pi(\theta)\} &= \frac{1}{n}\log\{L(\theta_0|y)\pi(\theta_0) \\ &\quad + E_{\theta|y}(\theta-\theta_0)'\frac{1}{n}\frac{\partial \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad + \frac{1}{2}E_{\theta|y}(\theta-\theta_0)'\frac{1}{n}\frac{\partial^2 \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\theta_0}(\theta-\theta_0) \\ &\quad + o_p(n^{-1}) \\ &= \frac{1}{n}\log\{L(\theta_0|y)\pi(\theta_0)\} \\ &\quad + E_{\theta|y}(\theta-\theta_0)'\frac{1}{n}\frac{\partial \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\ &\quad - \frac{1}{2}E_{\theta|y}(\theta-\theta_0)'J_n(\theta_0)(\theta-\theta_0) + o_p(n^{-1}). \end{aligned} \tag{3.4}$$

We also expand $\frac{\partial \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta}|_{\theta=\hat{\theta}}$ around θ_0 by Taylor's theorem to the first order term and obtain

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

Considering that the posterior mode $\hat{\theta}$ is the solution of $\frac{\partial \log\{L(\theta|y)\pi(\theta)\}}{\partial\theta} = 0$, we get

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

Substitute it into the second term of (3.4), the expansion of $E_{\theta|y}\frac{1}{n}\log\{L(\theta|y)\pi(\theta)\}$ becomes:

$$\begin{aligned}
E_{\theta|y}\frac{1}{n}\log\{L(\theta|y)\pi(\theta)\} &= \frac{1}{n}\log\{L(\theta_0|y)\pi(\theta_0)\} + E_{\theta|y}(\theta-\theta_0)'J_n(\theta_0)(\hat{\theta}-\theta_0) \\
&\quad -\frac{1}{2}E_yE_{\theta|y}(\theta-\theta_0)'J_n(\theta_0)(\theta-\theta_0) + o_p(n^{-1}) \\
&= \frac{1}{n}\log\{L(\theta_0|y)\pi(\theta_0)\} + tr\{E_{\theta|y}[(\hat{\theta}-\theta_0)(\theta-\theta_0)']J_n(\theta_0)\} \\
&\quad -\frac{1}{2}tr\{E_{\theta|y}[(\theta-\theta_0)(\theta-\theta_0)']J_n(\theta_0)\} + o_p(n^{-1}) \\
&= \frac{1}{n}\log\{L(\theta_0|y)\pi(\theta_0)\} + tr\{E_{\theta|y}[(\theta-\theta_0)(\hat{\theta}-\theta_0)']J_n(\theta_0)\} \\
&\quad -\frac{1}{2}tr\{\frac{1}{n}[J_n^{-1}(\hat{\theta}) + J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)]J_n(\theta_0)\} \\
&\quad +o_p(n^{-1}), \qquad (3.5)
\end{aligned}$$

where in (3.5) we replace $E_{\theta|y}[(\theta-\theta_0)(\theta-\theta_0)']$ with the result of Lemma 1d. $E_{\theta|y}[(\theta-\theta_0)(\hat{\theta}-\theta_0)']$ in the second term of (3.5) can be rewritten as $E_{\theta|y}[(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)'] + E_{\theta|y}[(\theta-\hat{\theta})(\hat{\theta}-\theta_0)']$, where the former term asymptotically equals to $\frac{1}{n}J_n^{-1}(\theta_0)I(\theta_0)J_n^{-1}(\theta_0)$ by Lemma 1a and the latter is negligible with higher order $o_p(n^{-1})$ as shown in Lemma 1c. Therefore, the expansion of $E_{\theta|y}\frac{1}{n}\log\{L(y|\theta)\pi(\theta)\}$ is finally simplified as

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

□

Proof of Theorem 3.1. Recall that the quantity of interest is $E_{\tilde{y}}E_{\theta|y}\log g(\tilde{y}|\theta)$. To estimate that, we first look at $E_{\tilde{y}}E_{\theta|y}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\} = E_{\tilde{y}}E_{\theta|y}\{\log g(\tilde{y}|\theta)+\log\pi_0(\theta)\}$ and expand it about θ_0,

$$\begin{aligned}
E_{\tilde{y}}E_{\theta|y}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\} &= E_{\tilde{y}}\log\{g(\tilde{y}|\theta_0)\pi_0(\theta_0)\} \\
&\quad +E_{\theta|y}(\theta-\theta_0)'\frac{\partial E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\
&\quad +\frac{1}{2}E_{\theta|y}[(\theta-\theta_0)'\frac{\partial^2 E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta\partial\theta'}|_{\theta=\theta_0}(\theta-\theta_0)] \\
&\quad +o_p(n^{-1}) \\
&= E_{\tilde{y}}\log\{g(\tilde{y}|\theta_0)\pi_0(\theta_0)\} \\
&\quad +E_{\theta|y}(\theta-\theta_0)'\frac{\partial E_{\tilde{y}}\log\{g(\tilde{y}|\theta)\pi_0(\theta)\}}{\partial\theta}|_{\theta=\theta_0} \\
&\quad -\frac{1}{2}E_{\theta|y}[(\theta-\theta_0)'J(\theta_0)(\theta-\theta_0)]+o_p(n^{-1}) \\
&\triangleq I_1+I_2+I_3+o_p(n^{-1}) \qquad (3.6)
\end{aligned}$$

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

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

where the last equation holds due to Lemma 1e.

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 1d, the third term I_3 can be rewritten as

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

By substituting each term in equation (3.6) and neglecting the residual term, we obtain

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

Recall that we have defined $\log\pi_0(\theta)=\lim_{n\to\infty}n^{-1}\log\pi(\theta)$, so that we have $\log\pi_0(\theta_0)-\frac{1}{n}\log\pi(\theta_0)\simeq 0$ and $E_{\theta|y}\log\{\pi_0(\theta)\}-E_{\theta|y}\frac{1}{n}\log\{\pi(\theta)\}\simeq 0$ asymptotically.

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

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

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 $E_{\theta|y}\frac{1}{n}\log\{L(\theta|y)\}-\frac{1}{n}tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\}$ as the asymptotically unbiased estimate for $E_{\tilde{y}}E_{\theta|y}\log\{g(\tilde{y}|\theta)\}$.

□

3.3 Simulation Study

In this section, we present numerical results to study the behavior of the proposed method under small and moderate sample sizes. In the first two simulation experiments, we estimate the true expected bias η either analytically (§ 3.3.1) or numerically by averaging $E_{\theta|y}[\log g(\tilde{y}|\theta)]$ over a large number of extra independent draws of $\tilde{y}$ when there is no closed form for integration (§ 3.3.2). The third example is a variable selection problem using real data to illustrate the practical difference between criteria proposed in an either explanatory and predictive perspective. To have BPIC well-defined for comparison, only the proper prior distributions are considered.

3.3.1 A simple linear example

The setting of the simulation study is the same as the example in section 3.3. As for the posterior average over the log-likelihood, we have

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

To eliminate the estimation error caused by the sampling of the observations y, we average the bias $\hat{\eta} - \eta$ over y with its true density $N(\mu_T, \sigma_T^2)$,

$$\begin{aligned}
b_\mu &= E_y(\hat{\eta} - \eta) = E_y\{\frac{\sigma_T^2}{2\sigma_A^2} + \frac{(\mu_T - \hat{\mu})^2}{2\sigma_A^2} - \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu})^2}{2\sigma_A^2}\} \\
&= \sigma_T^2\hat{\sigma}^2/\sigma_A^4.
\end{aligned}$$

Here we compare the bias estimate defined in Theorem 3.1, b_μ^{PAIC} with 3 other

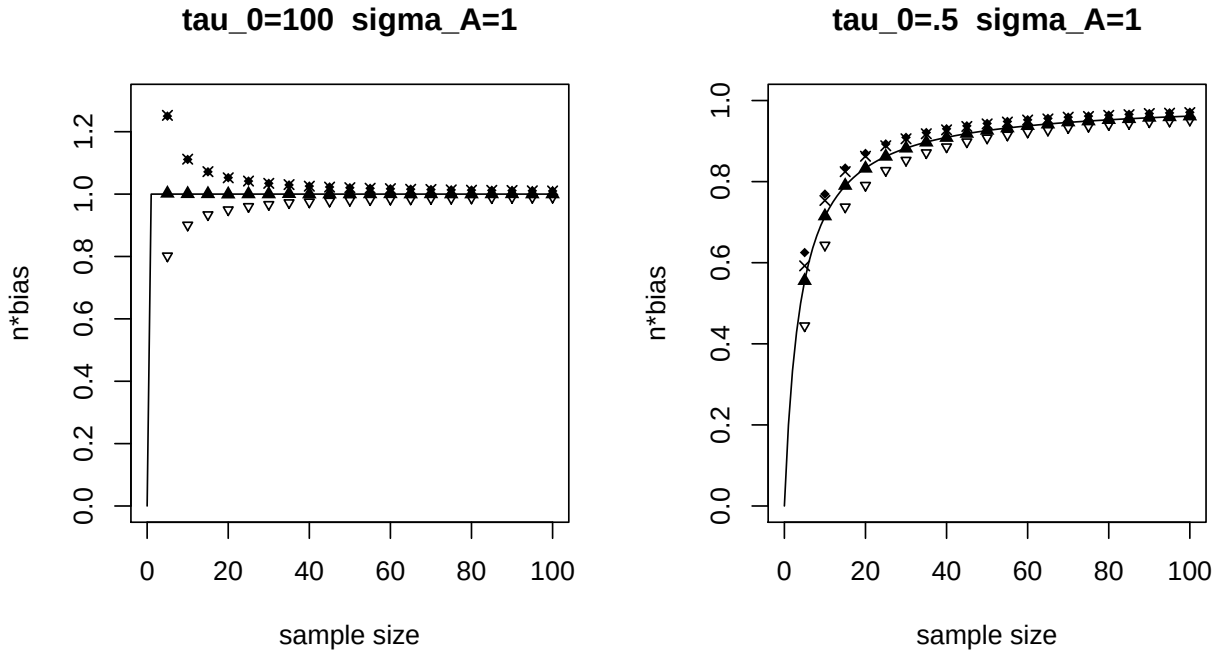

Figure 3.1: Performance of the estimators for $E_y(\hat{\eta} - \eta)$ when $\sigma_A^2 = \sigma_T^2 = 1$, i.e., the true distribution is contained in the candidate models. The left plot is under a relatively non-informative prior with $\tau_0 = 100$; the right plot is under a relatively informative prior with $\tau_0 = 0.5$. The true bias is curved by (——) as a function of sample size n. The averages of different bias estimators are marked by: (▲) for PAIC; (∇) for BPIC; (•) for PL^e; ($\times$) for cross-validation. Each mark represents the mean of estimated bias of 250,000 replications.

bias estimators: b_μ^{BPIC} (Ando, 2007), $b_\mu^{p^e_{opt}}$ (Plummer, 2008) and b_μ^{CV} (Stone, 1974).

$$
\begin{aligned}
b_\mu^{PAIC} &= \frac{1}{n-1}\hat{\sigma}^2 \sum_{i=1}^{n}((\mu_0 - \hat{\mu})/(n\tau_0^2) + (y_i - \hat{\mu})/\sigma_A^2)^2 \\
b_\mu^{BPIC} &= \frac{1}{n}\hat{\sigma}^2 \sum_{i=1}^{n}((\mu_0 - \hat{\mu})/(n\tau_0^2) + (y_i - \hat{\mu})/\sigma_A^2)^2 \\
b_\mu^{PL^e} &= \frac{1}{2n}p^e_{opt} = 1/(1/\tau_0^2 + (n-1)/\sigma_A^2)/\sigma_A^2 \\
b_\mu^{CV} &= \hat{\eta} - (\sum_{i=1}^{n}(y_i - (\mu_0/\tau_0^2 + \sum_{j\neq i} y_j/\sigma_A^2)/(1/\tau_0^2 + (n-1)/\sigma_A^2))^2/n + \hat{\sigma}^2)/\sigma_A^2/2
\end{aligned}
$$

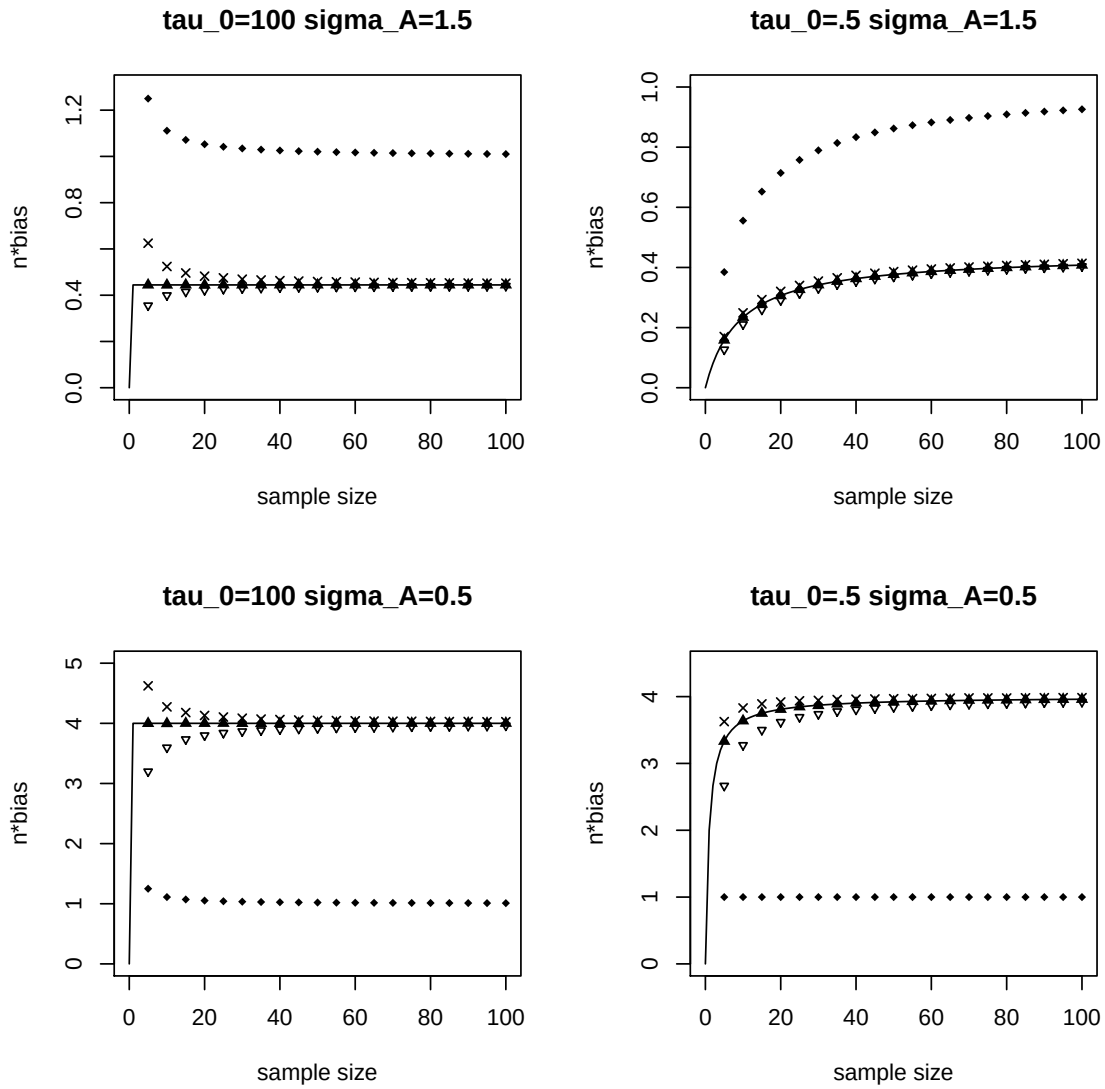

Figure 3.2: Performance of the estimators for $E_y(\hat{\eta} - \eta)$ when true model is not contained in the candidate distributions. The left two plots are under a relatively non-informative prior with $\tau_0 = 100$; the right ones are under a relatively informative prior with $\tau_0 = 0.5$. The true bias is curved by (——) as a function of sample size n. The averages of different bias estimators are marked by: ($\blacktriangle$) for PAIC; (∇) for BPIC; ($\bullet$) for PL^e; ($\times$) for cross-validation. Each mark represents the mean of estimated bias of 250,000 replications.

The results, are in accordance with theory. All of the 4 estimates are close to the true bias-correction values when $\sigma_A^2 = \sigma_T^2 = 1$, especially when the sample size becomes moderately large. However, the estimated values based on the PAIC are consistently closer to the true values than those based on Ando's method, which underestimate the bias, or the cross-validation or expected deviance penalized loss, which overestimate the bias, especially when sample size is small. When the models are misspecified, it is not surprising that in all of the plots given in Figure 3.2, only the expected deviance penalized loss misses the target even asymptotically since its assumption is violated, whileas all of the PAIC, BPIC and cross-validation converge to b_μ. In conclusion, PAIC achieves the best overall performance.

3.3.2 Bayesian hierarchical logistic regression

Consider frequencies $y_1, \ldots, y_N$ 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 inference of ξ's, we assume that the logits

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

are random effects which follow the normal distribution

$$\beta_i \sim N(\mu, \tau^2).$$

The very weakly-informative joint prior distribution $N(\mu; 0, 1000^2) \cdot Inv\text{-}\chi^2(\tau^2; 0.1, 10)$ is proposed on hyper-parameter (μ, τ^2) so that BPIC is properly defined and computable. The posterior distribution is asymmetric, due to both the logistic transformation and the hierarchical structure of parameters.

In this example, the true bias η does not have an analytical form. We estimate it through numerical computation. The simulation scheme is as follows:

1. Draw $\beta_i^T \sim N(0, 1)$, $i = 1, \ldots, N$; $y \sim Bin(n, logit^{-1}(\beta^T))$.

2. Simulate the posterior draws of $(\beta, \mu, \tau)|y$;

3. Estimate $\widehat{b_\beta^{PAIC}}$ and $\widehat{b_\beta^{BPIC}}$.

4. Draw $z^{(j)} \sim Bin(n, logit^{-1}(\beta^T))$, $j = 1, \ldots, J$;

5. Estimate $\widehat{b_\beta} = \hat{\eta} - \eta$ numerically with $\{z^{(j)}\}$, $j = 1, \ldots, J$.

6. Repeat steps 1-5 for 1000 times.

Table 3.1: The estimation error of bias correction: the mean and standard deviation (in parentheses) from 1000 replications.

	$\hat{\eta} - \eta - \widehat{b_\beta}$	$\lvert\hat{\eta} - \eta - b_\beta\rvert$	$(\hat{\eta} - \eta - b_\beta)^2$
b_β^{PAIC}	0.159	0.205	0.079
	(0.232)	(0.192)	(0.161)
b_β^{BPIC}	0.258	0.270	0.122
	(0.235)	(0.221)	(0.206)

Table 3.1 summarizes the simulation bias and standard deviation of the estimation error when we choose $N = 15$ and $n_1 = \ldots = n_N = 50$ and β's are independently simulated from the standard normal distribution. With respect to all three different metrics our bias estimation is consistently superior to that of BPIC, which matches our expectation that the natural estimate $\frac{1}{n}\sum_i E_{\theta|y}[\log g(y_i|\theta)]$ will estimate η_3 more precisely than $\frac{1}{n}\sum_i \log g(y_i|\hat{\theta})$ when the posterior distribution is asymmetric. Compared to BPIC, the bias and the average mean squared error are reduced by about 40%, while the absolute bias are reduced by about one quarter.

Y	the number of new accounts sold in a given time period
X_1	number of households serviced
X_2	number of people selling the new account
X_3	1 if the branch is in Manhattan and 0 otherwise
X_4	1 if the branch is in the boroughs and 0 otherwise
X_5	1 if the branch is in the suburbs and 0 otherwise
X_6	demand deposits balance
X_7	number of demand deposit
X_8	now accounts balance
X_9	number of now accounts
X_{10}	balance of money market accounts
X_{11}	number of money market accounts
X_{12}	passbook saving balance
X_{13}	other time balance
X_{14}	consumer loans
X_{15}	shelter loans

Table 3.2: Explanation of Data: The numbers of new accounts sold in some time period with 15 predictor variables in each of 233 branches. (George and McCulloch, 1993)

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

Table 3.3: Results from K-L based model selection criteria

3.3.3 Variable selection: a real example

In the last example we explore the problem of finding the best model to predict the selling of new accounts in branches of a large bank. The data is first introduced in the example 5.3 of George and McCulloch (1993), analyzed with their SSVS (Stochastic Search Variable Selection) technique to select the promising subsets of predictors. Their report on the most 10 frequently selected models after 10,000 iterations of Gibbs sampling for potential subsets, is listed in the first column of Table 3.3.

The original data consists of the numbers of new accounts sold in some time period as the outcome y, together with 15 predictor variables X in each of 233 branches. The description of the data is given in Table 3.2. The 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 $\sigma_y^2 \sim Inv\text{-}Gamma(.001, .001)$, when m indicates the specific model with a subset of predictor $X^{(m)}$.

Several model selection estimators for $-2n \cdot \eta$, including the leave-one-out cross-validated estimator, K-fold cross-validated estimator, the expected deviance penalized loss with p^e_{opt}, BPIC and PAIC, are calculated based on a large amount of MCMC draws of the posterior distribution for model selection inference. Here the original data is randomly partitioned for the K-fold cross-validation with a common choice $K = 10$. All the posterior samples are simulated from 3 parallel chains based on MCMC techniques for model selection inference. To generate 15000 effective draws of the posterior distribution, only one out of five iterations after convergence are kept to reduce the serial correlation.

The result is presented in Table 3.3 when the models having the smallest estimation value by each criterion are highlighted. An interesting finding is that the favored model selected by K-L based criteria and SSVS are quite different. Note that all of the K-L based criteria are developed in a predictive perspective, whereas SSVS is a variable selection method to pursue the model best describing the given set of data. This illustrates that with different modeling purpose, 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, when cross-validation estimators are noisy due to the simulation error and tend to over-estimate. It is worth to mention that the estimators of LOO-CV, K-fold-CV and $PL_{p^e_{opt}}$ are

relatively unstable even with 15000 posterior draws, as though those methods have been much more computationally intensive than BPIC and PAIC.

Chapter 4

Predictive Bayes factor

If we consider model selection as a problem of statistical decision, a natural way to formulate it within the Bayesian framework is Bayes factor.

4.1 Bayes Factors

Suppose we are considering a group of K candidate models, each specified by the density distribution $g_k(\cdot|\theta^k)$ with parameter prior distribution $\pi_k(\theta^k)$, $k = 1, 2, ..., K$. Given the prior probabilities $P(M_k)$ for each model, the posterior probabilities of M_k, $k = 1, 2, ..., K$, are given by

$$P(M_k|y) = \frac{p(y|M_k)P(M_k)}{p(y)} = \frac{p(y|M_k)P(M_k)}{\sum_{j=1}^{K} p(y|M_j)P(M_j)},$$

where $p(y|M_k) = \int \pi_k(\theta^k) \prod_i g_k(y_i|\theta^k) d\theta^k$ is usually called the (vector) prior posterior distribution (Gelman et al. 2003) or integrated likelihood (Madigan and Raftery, 1994).

When the candidate models are compared pairwisely, the denominator $p_k(y)$ cancels out, and the odds of posterior probabilities in favor of model M_k over alternative

model M_j is

$$\frac{P(M_k|y)}{P(M_j|y)} = \frac{p(y|M_k)}{p(y|M_j)} \frac{P(M_k)}{P(M_j)}. \tag{4.1}$$

It reveals the key role of the ratio of integrated likelihood, defined as the (standard) Bayes factor (Kass and Raftery 1995),

$$B_{kj} = \frac{p(y|M_k)}{p(y|M_j)} = \frac{\int \pi_k(\theta^k) \prod_i g_k(y_i|\theta^k) d\theta^k}{\int \pi_j(\theta^j) \prod_i g_j(y_i|\theta^j) d\theta^j}, \tag{4.2}$$

in the mechanism of changing the posterior odds of model M_k from its priors. It is one of the most widely used Bayesian model selection measure, which can be dated to Jeffreys (1939) in the name of 'tests of significance', with respect to comparative support for the two models from the data y.

Laplace's method (Tierney and Kadane, 1986; Tierney et al., 1989) is traditionally employed in the approximation of marginal distribution. However, it may be challenging or even impossible when parameter spaces are high-dimensional, and the same difficulty also applies to Markov chain Monte Carlo (MCMC) algorithms. Han and Carlin (2001) review and compare five different simulation methods for computing Bayes factors under proper parameter priors and suggested using the marginal likelihood methods (Chibs, 1995) for its accuracy.

In the literature, the strength and weakness of Bayes factor have been actively debated (for instance, see Jeffreys, 1961; Kass, 1993; Gilks et al., 1996; Berger and Pericchi, 2001, on its attractive features and difficulties). Generally speaking, standard Bayes factor is intuitive in a Bayesian nature and easy to interpret, but has drawbacks such as Lindley's paradox, a case that the nested models may result in support of the reduced model in spite of the data when the prior is proper and sufficiently diffuse (Lindley, 1957; Shafer, 1982).

The values of Bayes factors may strongly depend on the choices of diffuse prior information on the model parameters. Especially, the improper non-informative priors, which are commonly used in Bayesian analysis for an objective purpose, will make

Bayes factor non-interpretable since the denominator of the Bayes factor becomes zero. Hill (1982) addresses this problem with a review of interesting historical comments. Some efforts has been made to resolve this difficulty such as intrinsic Bayes factors (Berger and Pericchi, 1996) and fractional Bayes factors (O'Hagan, 1995). Their general idea is to set aside part of the data or information (the likelihood function) to update the prior distribution to avoid the weak prior distribution and use the remainder of the data for model discrimination.

4.2 Predictive Bayes factor

'Surely we do not require that the experimenters return to their prior densities for θ_j, given their information about the particular value of θ_j that actually applied in this experiment, nor that they generate independent data from a new experiment, to settle the issue of which model is better supported by the previous experiment.'

Aitkin (1991) p.141.

4.2.1 Models under comparison: Original or Fitted?

As we can tell from its definition in (4.2), the standard Bayes factor evaluates the goodness of the candidate models when model fitting is not in need. That property is considered as a significant advantage. However, it also indicates that the model comparison with respect to standard Bayes factor would be made among the original models, rather than the fitted models whose parameter distribution has been updated to the posterior.

A review of the general class of various Bayes factors (Gelfand and Dey, 1994)

may help us understand this difference. The formal conditional distribution

$$\begin{aligned} p(y_{S_1}|y_{S_2}, M_k) &= \int p(y_{S_1}|\theta^k, M_k)p(\theta^k|y_{S_2})d\theta^k \\ &= \frac{\int p(y_{S_1}|\theta^k, M_k)p(y_{S_2}|\theta^k, M_k)\pi_k(\theta^k)d\theta^k}{\int p(y_{S_2}|\theta^k, M_k)\pi_k(\theta^k)d\theta^k} \end{aligned} \tag{4.3}$$

is defined as a predictive density which averages the joint density of y_{S_1} against the prior for θ^k updated by y_{S_2}, where S_1, S_2 are arbitrary subsets of the universe $U_n = \{1, \ldots, n\}$ and $y = y_{U_n}$. From a cross-validation perspective, y_{S_1} can be viewed as the testing sample whereas y_{S_2} as the training sample.

When $S_1 = U_n$ and $S_2 = \emptyset$, (4.3) yields the prior predictive density of the data used in the standard Bayes factor

$$BF_{k,k'} = \frac{p(y|M_k)}{p(y|M_{k'})},$$

explaining the prediction power implied in the model M_k:

$$\begin{aligned} \tilde{y} &\sim g_k(\tilde{y}|\theta^k) \\ \theta^k &\sim \pi_k(\theta^k). \end{aligned} \tag{4.4}$$

The state of knowledge within the models subject to comparison only rely on the prior; all of the observations y are retained to test the adequacy of the candidate models. Subsequently, standard Bayes factor demonstrates the relative evidence in probability to support model M_k against $M_{k'}$ when describing the observed data y.

In contrast to considering prior predictive density in model assessment, Aitkin (1991) proposes the posterior Bayes factor

$$PoBF_{k,k'} = \frac{p(y|y, M_k)}{p(y|y, M_{k'})}$$

in terms of posterior predictive density where $S_2 = U_n$ to replace $S_2 = \emptyset$ in the standard Bayes factor when $S_1 = U_n$ is unchanged. Given both the model and the

full data have been seen, here the fitted models in light of the current data y are compared.

Because the entire dataset is used twice (first to convert the prior into the posterior, and then to compute the realized discrepancy), posterior predictive density in tradition is merely employed as a conservative tool in Bayesian model monitoring and checking (Rubin, 1984; Robins et al., 2000). Without any penalty for data overusage, the application of posterior predictive density for model evaluation may lead to some counterintuitive results. This and some other criticisms have been pointed out by Dawid, Fearn, Goldstern, Lindley, and Whittakerthe in the discussion of Aitkin (1991).

4.2.2 Predictive Bayes factor

Given the observed data y, it is of general interest for Bayesian researchers to make the comparison among the fitted models other than original models.

In order to illustrate the relationship between a candidate model under comparison and data y, we first change the notation of the original model M_k to $M_k(\emptyset)$ hereinafter, and let $M_k(y)$ denote the fitted model in light of data y:

$$\begin{aligned} \tilde{y} &\sim g_k(\tilde{y}|\theta^k) \\ \theta^k &\sim p_k(\theta^k|y) \propto \pi_k(\theta^k) \prod_i g_k(y_i|\theta^k). \end{aligned} \tag{4.5}$$

From a predictive perspective, next we pay major attention to model selection approaches in terms of model probabilities to compare models $M_k(y)$, $k = 1, 2, ..., K$ against a future observable $\tilde{y}$.

In the class of Gelfand and Dey (1994), if we expand the universe to $U_{n+1} = \{1, \ldots, n+1\}$ where $y_{n+1} = \tilde{y}$ denoting a future independent observation, $S_1 = \{n+1\}$

and $S_2 = U_n$ yields the posterior predictive distribution of the testing sample $\tilde{y}$,

$$p(\tilde{y}|M_k(y)) = \int g_k(\tilde{y}|\theta^k)p_k(\theta^k|y)d\theta^k, \tag{4.6}$$

when all of the observations y are employed as the training sample to update the knowledge of the parameters. The unobserved quantity $\tilde{y}$ is presumed to be generated from f, the underlying true distribution. Taking that into account, we evaluate the goodness of a model $M_k(y)$ through the similarity of the distribution (4.6) to f.

The posterior predictive distribution of $\tilde{y}$ can be empirically assessed on behalf of the observable y. In order to avoid the double use of the data, a numerical solution is to employ the cross-validation method. For each single observation $j \in U_n$, $S_1 = \{j\}$ and $S_2 = U_n/\{j\}$ for (4.3) yields the observation-i-deleted cross-validative predictive density (Geisser, 1975); their product $\prod_{i=1}^{n} p(y_i|M_k(y_{-i}))$ is suggested as the pseudo-predictive distribution to replace $p(M_k(\emptyset), y)$ for model selection, on which the pseudo-Bayes factor (Geisser and Eddy, 1979) is defined as

$$PsBF_{k,k'} = \frac{\prod_i p(y_i|M_k(y_{-i}))}{\prod_i p(y_i|M_{k'}(y_{-i}))}.$$

Mathematically, the logarithm of pseudo-predictive distribution is exactly the first-order leave-one-out cross-validation estimator for $E_{\tilde{y}} \log p(\tilde{y}|M_k(y))$. Note that even with numerical approximation, it is computationally challenging to apply leave-one-out cross-validation method for Bayesian modeling.

One alternative approach is to approximate $E_{\tilde{y}} \log p(\tilde{y}|M_k(y))$ directly with penalized empirical posterior predictive density $\frac{1}{n}\sum_i \log p(y_i|M_k(y)) + b_k$, where b_k is the bias of over-estimation of the empirical log posterior predictive for the 'double use' of y. Subsequently, we can define the predictive Bayes factor (PrBF):

$$PrBF_{k,k'} = \frac{\prod_i p(y_i|M_k(y))}{\prod_i p(y_i|M_{k'}(y))} \cdot \frac{\exp(n \cdot b_k)}{\exp(n \cdot b_{k'})}$$

as a measure of the weight of sample evidence in favor of $M_k(y)$ compared with $M_{k'}(y)$.

An asymptotic unbiased estimator of the bias b_k is $-\frac{1}{n}tr\{J_{n,k}^{-1}(\hat{\theta}^k)I_{n,k}(\hat{\theta}^k)\}$. The details of the derivation are given in the Theorem 4.1 of next section. Hence, empirically we present PrBF as

$$PrBF_{k,k'} = \frac{\prod_i p_k(y_i|y)}{\prod_i p_{k'}(y_i|y)} \cdot \frac{\exp(-tr\{J_{n,k}^{-1}(\hat{\theta}^k)I_{n,k}(\hat{\theta}^k)\})}{\exp(-tr\{J_{n,k'}^{-1}(\hat{\theta}^{k'})I_{n,k'}(\hat{\theta}^{k'})\})}$$

The posterior predictive density of $\tilde{y}$ indicates what a future observation would look like, given the updated model fully refined by the entire data. By employing the posterior predictive distribution rather than the prior predictive distribution, it reduces the sensitivity to variations in the prior distribution and avoids the degeneration of the integrated likelihood as well as the Lindley paradox. Compared with the cross-validative predictive densities in pseudo Bayes factor, the natural estimator of the posterior predictive distribution $E_{\tilde{y}} \log p(\tilde{y}|M_k(y))$

$$\frac{1}{n}\sum_i \log p_k(y_i|y)$$

is fast to compute and steady in estimation. Unlike posterior Bayes factor, predictive Bayes factor penalizes the over-estimation asymptotically unbiasedly. In addition, the predictive Bayes factor inherits the property of coherence, i.e. the Bayes factor between, say, models A and C equal the Bayes factor between models A and B multiplied by the Bayes factor between models B and C. Coherence is important for result interpretation when more than 2 candidate models are under comparison.

4.3 Posterior Predictive Information Criterion

In this section we derive an asymptotically unbiased estimator for the expected out-of-sample log posterior predictive density $E_{\tilde{y}} \log p(\tilde{y}|y)$, on which the theoretical foundation of the predictive Bayes factor is built. The quantity, $E_{\tilde{y}} \log p(\tilde{y}|y)$, can be considered as a special K-L discrepancy function, a distance measure to compare the

posterior predictive density of a future observation with the underlying true model, where the true model is only referred to a model as the best projection unto the statistical modeling space once we try to understand the real but unknown dynamics/mechanism of interest. Based on the bias correction from an asymptotic estimator of $E_{\tilde{y}} \log p(\tilde{y}|y)$, we also propose an ad hoc information criterion in terms of the posterior predictive density for Bayesian evaluation.

4.3.1 Asymptotic estimation for K-L discrepancy

Theorem 4.1. *Let $y = (y_1, y_2, \cdots, y_n)$ 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 not necessarily containing the true distribution f, where $\theta = (\theta_1, ..., \theta_p)'$ is the p-dimensional vector of unknown parameters, with prior distribution $\pi(\theta)$. The asymptotic bias of $\hat{\eta} = \frac{1}{n}\sum_i \log p(y_i|y)$ for $\eta = E_{\tilde{y}} \log p(\tilde{y}|y)$ is estimated by*

$$E_y(\hat{\eta} - \eta) = b_\theta \approx \frac{1}{n} tr\{J^{-1}(\theta_0) I(\theta_0)\}, \tag{4.7}$$

where

$$\begin{aligned} J(\theta) &= -E_{\tilde{y}}\left(\frac{\partial^2 \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta\partial\theta'}\right), \\ I(\theta) &= E_{\tilde{y}}\left(\frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)}{\partial\theta} \frac{\partial \log\{g(\tilde{y}|\theta)\pi^{\frac{1}{n}}(\theta)\}}{\partial\theta'}\right) \end{aligned}$$

under the following regularity conditions:

C1: Both the log density function $\log g(\tilde{y}|\theta)$ and the log unnormalized posterior density $\log\{L(\theta|y)\pi(\theta)\}$ are twice continuously differentiable in the compact parameter space Θ, where $L(\theta|y) = \prod_i g(y_i|\theta)$;

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.

Accordingly, we propose the posterior predictive information criterion (PPIC)

$$-2\sum_i \log p(y_i|y) + 2 \cdot tr\{J_n^{-1}(\hat{\theta}) I_n(\hat{\theta})\}, \tag{4.8}$$

where $\hat{\theta} = \arg\max_\theta L(\theta|y)\pi(\theta)$ is the posterior mode and

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

Models with small PPIC values are favored when various candidate Bayesian models are under comparison. Actually, the difference of PPICs can be interpreted by Table 4.3, through the equation

$$PPIC_1 - PPIC_2 = 2\log_e PrBF_{12}.$$

The proposed criterion has many attractive properties. As an objective model selection criterion consistent with Bayesian philosophy, PPIC is developed by unbiasedly correcting the asymptotic bias of the log posterior against the ad hoc K-L discrepancy, which measures the similarity of the predictive distribution and the underlying true distribution. Without presuming that the approximating distributions contains the truth, our criterion is generally applicable for Bayesian model comparison. From a predictive perspective, the direct estimation penalized for 'double use of the data' makes PPIC much easier to adopt computationally than other numerically methods such as cross-validation, especially when the model structure is complicated. Note that all of those properties are also possessed by predictive Bayes factor.

In the literature, Konishi and Kitagawa (1996) propose a similar-looking criterion to PPIC in their section 3.4. From a frequentist's perspective, their attempt is to build up the asymptotic link between the log-likelihood, which was estimated at the MLE other than the posterior mode, and the predictive distribution. By neglecting

the information contained within the prior distribution $\pi(\theta)$, that kind of approach may cause significant bias even for large samples; for instance, see arguments in Appendix 2 of Ando (2007). What's more, the error correction term and its estimator is also affected through the incomplete definition of the matrices $J(\theta)$ and $I(\theta)$. It invariably induces to biased results when evaluating most Bayesian models, which partially explains why the error correction of their proposed criterion is only in the order of $O_p(n^{-1})$, against $o_p(n^{-1})$ in our proposal when we consider the posterior mode $\hat{\theta}$ in the evaluation of the predictive density.

To prove the Theorem 4.1, we note that the posterior predictive density can be expanded as

$$p(\tilde{y}|y) = \int g(\tilde{y}|\theta)p(\theta|y)d\theta = \frac{\int g(\tilde{y}|\theta)L(\theta|y)\pi(\theta)d\theta}{\int L(\theta|y)\pi(\theta)d\theta} \tag{4.9}$$

$$= \frac{g(\tilde{y}|\hat{\theta}(\tilde{y}))L(\hat{\theta}(\tilde{y})|y)\pi(\hat{\theta}(\tilde{y}))}{L(\hat{\theta}|y)\pi(\hat{\theta})} \left\{ \frac{\left|J_n(\hat{\theta})\right|}{\left|H_n(\tilde{y},\hat{\theta}(\tilde{y}))\right|} \right\}^{\frac{1}{2}} + O_p(n^{-2}) \tag{4.10}$$

by Laplace transformation (Bernardo and Smith, 1994 §5.5.1), where $(\hat{\theta}(\tilde{y}),\ H_n(\tilde{y},\theta))$ and $(\hat{\theta},\ J_n(\theta))$ are pairs of posterior modes and second derivative matrices of $-\frac{1}{n}\log\{g(\tilde{y}|\theta)L(\theta|y)\pi(\theta)\}$ and $-\frac{1}{n}\log\{L(\theta|y)\pi(\theta)\}$, respectively. For notational purpose, letting

$$h(\tilde{y},\theta;y) = \log\{g(\tilde{y}|\theta)L(\theta|y)\pi(\theta)\},$$

then we have

$$H_n(\tilde{y},\theta) = -\frac{1}{n}\frac{\partial^2 h(\tilde{y},\theta)}{\partial\theta\partial\theta'}.$$

With the definition of

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

we start with the proofs of a few lemmas to support the proof of Theorem 4.1.

LEMMA 1. *Under the same regularity conditions of Theorem 4.1,*

$$\theta_0 - \hat{\theta}(\tilde{y}) = O_p(n^{-1/2}); \quad \theta_0 - \hat{\theta}(y_i) = O_p(n^{-1/2}).$$

Proof. Expand $\frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})}$ at θ_0

$$\begin{aligned} \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})} &\simeq \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\theta_0} + \frac{\partial^2 h(\tilde{y},\theta)}{\partial \theta \partial \theta'}|_{\theta=\theta_0}(\hat{\theta}(\tilde{y}) - \theta_0) \\ &= \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\theta_0} - nH_n(\tilde{y},\theta_0)(\hat{\theta}(\tilde{y}) - \theta_0). \end{aligned}$$

The left-hand-side is 0 since $\hat{\theta}(\tilde{y})$ is the mode of $h(\tilde{y},\theta)$, and $\frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\theta_0}$ on the right-hand-side converges to $N(0, nI(\theta_0) + K(\theta_0))$. Therefore, we obtain

$$\sqrt{n}(\hat{\theta}(\tilde{y}) - \theta_0) \sim N(0, H_n^{-1}(\tilde{y},\theta_0)(I(\theta_0) + \frac{1}{n}K(\theta_0))H_n^{-1}(\tilde{y},\theta_0)),$$

or $\theta_0 - \hat{\theta}(\tilde{y}) = O_p(n^{-1/2})$.

Following the same procedure, we derive $\theta_0 - \hat{\theta}(y_i) = O_p(n^{-1/2})$. □

LEMMA 2. *Under the same regularity conditions of Theorem 4.1, both* $n(\hat{\theta}(\tilde{y}) - \hat{\theta})$ *and* $n(\hat{\theta}(y_i) - \hat{\theta})$ *are approximately distributed as* $N(0, J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta}))$.

Proof. Expand $\frac{\partial \log\{L(y|\theta)\pi(\theta)\}}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})}$ at $\hat{\theta}$

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

The left-hand-side $\frac{\partial \log\{L(y|\theta)\pi(\theta)\}}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})} = -\frac{\partial \log g(\tilde{y}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})}$ is approximately distributed as $N(0, K(\theta_0))$. Therefore we obtain

$$n(\hat{\theta}(\tilde{y}) - \hat{\theta}) \sim N(0, J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})).$$

Similarly, we can prove that the asymptotic distribution of $n(\hat{\theta}(y_i) - \hat{\theta})$ is

$$N(0, J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta}))$$

. □

LEMMA 3. *Under the same regularity conditions of Theorem 4.1,*

$$\hat{\theta}(\tilde{y}) - \hat{\theta}(y_i) = o_p(n^{-1}).$$

Proof. Expand $\frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})}$ at $\hat{\theta}(y_i)$

$$\begin{aligned} \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} &\simeq \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(\tilde{y})} + \frac{\partial^2 h(\tilde{y},\theta)}{\partial \theta \partial \theta'}|_{\theta=\hat{\theta}(\tilde{y})}(\hat{\theta}(y_i) - \hat{\theta}(\tilde{y})) \\ &= -nH_n(\tilde{y}, \hat{\theta}(\tilde{y}))(\hat{\theta}(y_i) - \hat{\theta}(\tilde{y})). \end{aligned} \tag{4.11}$$

The left-hand-side

$$\begin{aligned} \frac{\partial h(\tilde{y},\theta)}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} &= \frac{\partial h(y_i,\theta)}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} + \frac{\partial \log g(\tilde{y}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} - \frac{\partial \log g(y_i|\theta)}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} \\ &= \frac{\partial[\log g(\tilde{y}|\theta) - \log g(y_i|\theta)]}{\partial \theta}|_{\theta=\hat{\theta}(y_i)} \end{aligned}$$

converges to 0 as $n \to \infty$, and the right-hand-side of 4.11

$$\hat{\theta}(\tilde{y}) - \hat{\theta}(y_i) = o_p(n^{-1}).$$

□

LEMMA 4. *Under the same regularity conditions of Theorem 4.1,*

$$E_{\tilde{y}}E_y[H_n(y_i, \hat{\theta}(y_i)) - H_n(\tilde{y}, \hat{\theta}(\tilde{y}))] = o_p(n^{-1}).$$

Proof.

$$
\begin{aligned}
-E_y H_n(y_i, \hat{\theta}(y_i)) = & \frac{1}{n} E_y \frac{\partial^2 \log g(y_i|\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(y_i)} + \frac{1}{n} E_y \frac{\partial^2 \log L(\theta|y)\pi(\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(y_i)} \\
= & \frac{1}{n} E_y \frac{\partial^2 \log g(y_i|\theta)}{\partial\theta\partial\theta'}|_{\theta_0} + \frac{1}{n} E_y \frac{\partial^2 \log L(\theta|y)\pi(\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(y_i)} + o_p(n^{-1}) \qquad (4.12) \\
= & \frac{1}{n} E_{\tilde{y}} \frac{\partial^2 \log g(\tilde{y}|\theta)}{\partial\theta\partial\theta'}|_{\theta_0} + \frac{1}{n} E_y \frac{\partial^2 \log L(\theta|y)\pi(\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(y_i)} + o_p(n^{-1}) \qquad (4.13) \\
= & \frac{1}{n} E_{\tilde{y}} \frac{\partial^2 \log g(\tilde{y}|\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(\tilde{y})} + \frac{1}{n} E_{\tilde{y}} E_y \frac{\partial^2 \log L(\theta|y)\pi(\theta)}{\partial\theta\partial\theta'}|_{\hat{\theta}(\tilde{y})} + o_p(n^{-1}) \qquad (4.14) \\
= & -E_{\tilde{y}} E_y H_n(\tilde{y}, \hat{\theta}(\tilde{y})) + o_p(n^{-1}).
\end{aligned}
$$

Lemma 1 is used in (4.12) and (4.14), whereas (4.13) uses Lemma 3. □

LEMMA 5. *Under the same regularity conditions of Theorem 4.1,*

$$E_{\tilde{y}} E_y[J_n(\theta_0) - H_n(\tilde{y}, \theta_0)] = o_p(1); \quad E_{\tilde{y}} E_y[J_n(\theta_0) - H_n(y_i, \theta_0)] = o_p(1).$$

Proof. Compare the definition of $J_n(\theta)$, $H_n(\tilde{y}, \theta)$ and $H_n(y_i, \theta)$ directly, we obtained the desired result. □

LEMMA 6. *Under the same regularity conditions of Theorem 4.1, asymptotically*

$$E_y[\frac{1}{n}\sum_i \log g(y_i|\hat{\theta}) - E_{\tilde{y}} \log g(\tilde{y}|\hat{\theta})] = \frac{1}{n} tr\{J^{-1}(\theta_0) I(\theta_0)\}.$$

Proof. Essentially, it is a Bayesian adaption of TIC (Takeuchi, 1976). The proof can be derived directly from Theorem 3.1 of Zhou (2011a) by applying the posterior mode as the functional estimator $T(\hat{F})$. □

With all the above results, here we give a proof of Theorem 4.1.

Proof of Theorem 4.1. The log transformed predictive distributions are given as

$$\begin{aligned}
\log p(\tilde{y}|y) &= h(\tilde{y},\hat{\theta}(\tilde{y})) - \frac{1}{2}\log\left|H_n(\tilde{y},\hat{\theta}(\tilde{y}))\right| \\
&\quad - \log L(\hat{\theta}|y)\pi(\hat{\theta}) + \frac{1}{2}\log\left|J_n(\hat{\theta})\right| + o_p(n^{-1}); \qquad (4.15)\\
\frac{1}{n}\sum_i \log p(y_i|y) &= \frac{1}{n}\sum_i h(y_i,\hat{\theta}(y_i)) - \frac{1}{2n}\sum_i \log\left|H_n(y_i,\hat{\theta}(y_i))\right| \\
&\quad - \log L(\hat{\theta}|y)\pi(\hat{\theta}) + \frac{1}{2}\log\left|J_n(\hat{\theta})\right| + o_p(n^{-1}). \qquad (4.16)
\end{aligned}$$

Expanding $h(\tilde{y},\hat{\theta})$ in Taylor series around $\hat{\theta}(\tilde{y})$, and using Lemma 2, Lemma 1 and Lemma 5 in steps, we have

$$\begin{aligned}
h(\tilde{y},\hat{\theta}) &= h(\tilde{y},\hat{\theta}(\tilde{y})) + \frac{\partial h(\tilde{y},\theta)}{\partial\theta'}|_{\theta=\hat{\theta}(\tilde{y})}(\hat{\theta}-\hat{\theta}(\tilde{y})) \\
&\quad -\frac{1}{2}(\hat{\theta}-\hat{\theta}(\tilde{y}))' n H_n(\tilde{y},\hat{\theta}(\tilde{y}))(\hat{\theta}-\hat{\theta}(\tilde{y})) + o_p(n^{-1}) \\
&= h(\tilde{y},\hat{\theta}(\tilde{y})) - \frac{1}{2n} tr\{H_n(\tilde{y},\hat{\theta}(\tilde{y}))J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}) \\
&= h(\tilde{y},\hat{\theta}(\tilde{y})) - \frac{1}{2n} tr\{H_n(\tilde{y},\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}) \\
&= h(\tilde{y},\hat{\theta}(\tilde{y})) - \frac{1}{2n} tr\{J_n(\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}). \qquad (4.17)
\end{aligned}$$

Using a very similar argument as above,

$$\begin{aligned}
h(y_i,\hat{\theta}) &= h(y_i,\hat{\theta}(y_i)) + \frac{\partial h(y_i,\theta)}{\partial\theta'}|_{\theta=\hat{\theta}(y_i)}(\hat{\theta}-\hat{\theta}(y_i)) \\
&\quad -\frac{1}{2}(\hat{\theta}-\hat{\theta}(y_i))' n H_n(y_i,\hat{\theta}(y_i))(\hat{\theta}-\hat{\theta}(y_i)) + o_p(n^{-1}) \\
&= h(y_i,\hat{\theta}(y_i)) - \frac{1}{2n} tr\{H_n(y_i,\hat{\theta}(y_i))J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}) \\
&= h(y_i,\hat{\theta}(y_i)) - \frac{1}{2n} tr\{H_n(y_i,\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}) \\
&= h(y_i,\hat{\theta}(y_i)) - \frac{1}{2n} tr\{J_n(\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} + o_p(n^{-1}), \qquad (4.18)
\end{aligned}$$

Substitute (4.17) and (4.18) into (4.15) and (4.16) respectively,

$$\begin{aligned}
\log p(\tilde{y}|y) &= h(\tilde{y},\hat{\theta}) + \frac{1}{2n} tr\{J_n(\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\} - \frac{1}{2}\log\left|H_n(\tilde{y},\hat{\theta}(\tilde{y}))\right| \\
&\quad - \log L(\hat{\theta}|y)\pi(\hat{\theta}) + \frac{1}{2}\log\left|J_n(\hat{\theta})\right| + o_p(n^{-1}) \\
\frac{1}{n}\sum_i \log p(y_i|y) &= \frac{1}{n}\sum_i \{h(y_i,\hat{\theta}) + \frac{1}{2n} tr\{J_n(\theta_0)J_n^{-1}(\hat{\theta})K(\theta_0)J_n^{-1}(\hat{\theta})\}\} \\
&\quad - \frac{1}{2n}\sum_i \log\left|H_{n+1}(y_i,\hat{\theta}(y_i))\right| - \log L(\hat{\theta}|y)\pi(\hat{\theta}) + \frac{1}{2}\log\left|J_n(\hat{\theta})\right| \\
&\quad + o_p(n^{-1})
\end{aligned}$$

Taking expectations with respect to the underlying true distribution and using Lemma 4 and 6, we complete the proof. □

4.3.2 A simple simulation study

To give insight into PPIC, we first apply it to a simple simulation study of normal model with known variance.

Suppose observations $y = (y_1, y_2, ..., y_n)$ 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 is 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.

It is easy to derive the posterior distribution of μ which 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, we have

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

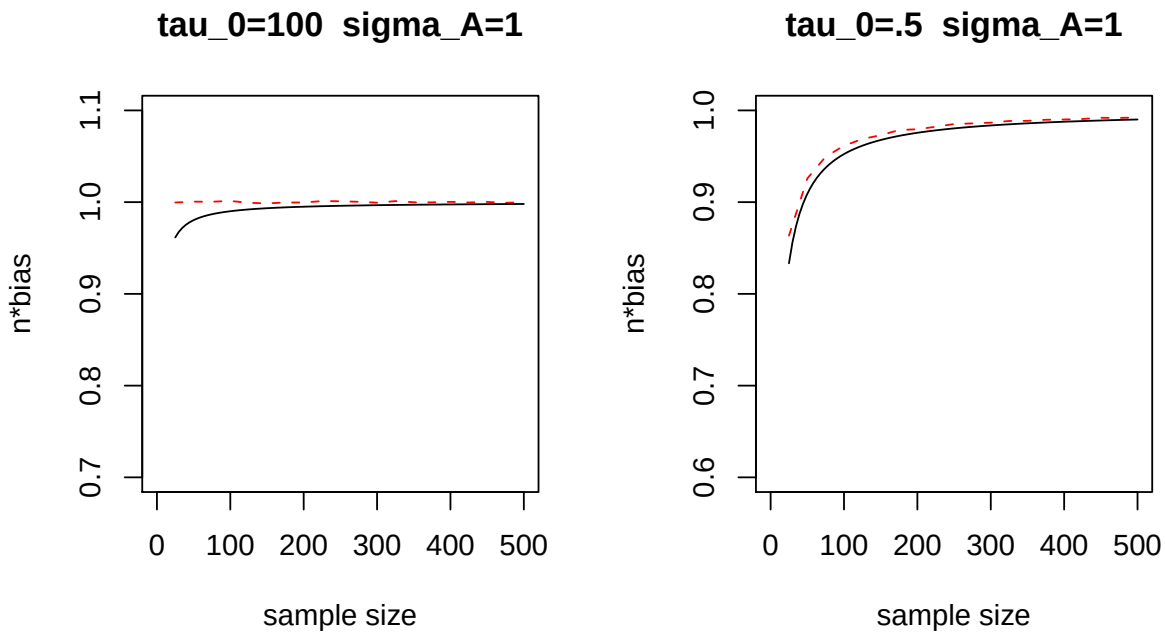

Figure 4.1: Comparison of the expected true bias nb_μ(———) and the bias estimated by PPIC $n\hat{b}_\mu^{PPIC}$(– – –) when $\sigma_A^2 = \sigma_T^2 = 1$. The left plot is under a relatively non-informative prior with $\tau_0 = 100$; the right plot is under a relatively informative prior with $\tau_0 = 0.5$.

To eliminate the estimation error caused by the sampling of the observations y, we average the bias $\hat{b}_\mu = \hat{\eta} - \eta$ over y with its true density $N(\mu_T, \sigma_T^2)$,

$$\begin{aligned} b_\mu &= E_y(\hat{\eta} - \eta) = E_y\{\frac{\sigma_T^2}{2(\sigma_A^2+\sigma_T^2)} + \frac{(\mu_T-\hat{\mu})^2}{2(\sigma_A^2+\sigma_T^2)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu})^2}{2(\sigma_A^2+\sigma_T^2)}\} \\ &= \frac{\sigma_T^2}{2(\sigma_A^2+\sigma_T^2)} + \frac{-\sigma_T^2/\tau_0^4 - 2(n-1)\sigma_T^2/\tau_0^2\sigma_A^2 + n(2-n)\sigma_T^2/\sigma_A^4}{2(\sigma_A^2+\sigma_T^2)(1/\tau_0^2+n/\sigma_A^2)^2} = \frac{\sigma_T^2\hat{\sigma}^2}{\sigma_A^2(\sigma_A^2+\sigma_T^2)}, \end{aligned}$$

whereas the asymptotic bias estimator (4.7) is

$$\hat{b}_\mu^{PPIC} = \frac{1}{n-1}\hat{\sigma}^2\sum_{i=1}^{n}((\mu_0-\hat{\mu})/(n\tau_0^2) + (y_i-\hat{\mu})/\sigma_A^2)^2.$$

The true mean and variance are arbitrarily set to be $\mu_T = 0$ and $\sigma_T = 1$, respectively, and the prior variances are set to be either the informative $\tau_0^2 = (.5)^2$ or

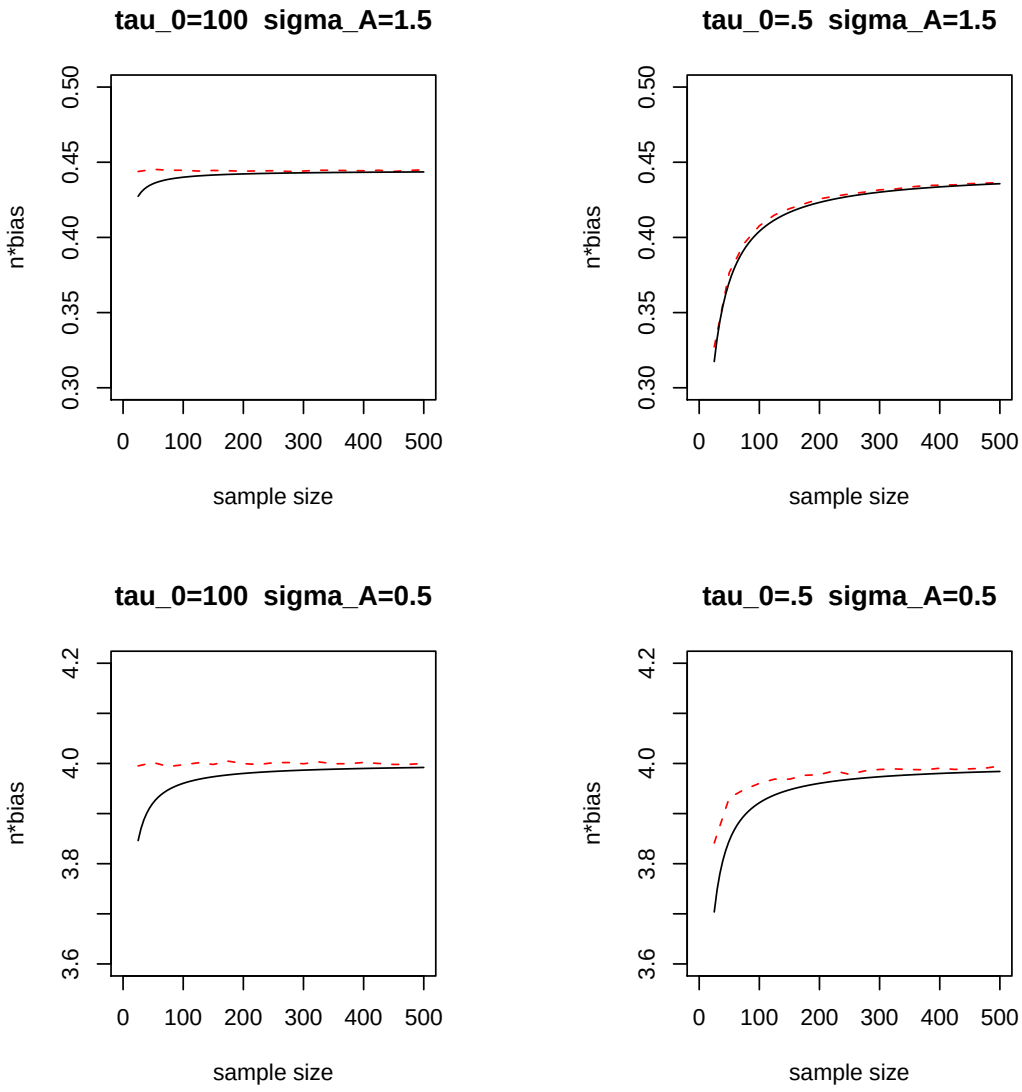

Figure 4.2: Comparison of the expected true bias nb_μ(———) and the bias estimated by PPIC $n\hat{b}_\mu^{PPIC}$(— — —) when $\sigma_A^2 \neq \sigma_T^2$. The left two plots are under a relatively non-informative prior with $\tau_0 = 100$; the right ones are under a relatively informative prior with $\tau_0 = 0.5$.

almost non-informative $\tau_0^2 = (100)^2$ with prior mean $\mu_0 = 0$. After a Monte Carlo simulation with 25,000 repetitions for each pre-specified n, curves of expected true bias b_μ against the bias estimates $\hat{b}_\mu^{PPIC}$ are plotted in either Figure 4.1 for the case $\sigma_A^2 = \sigma_T^2$, or Figure 4.2 when the equality between σ_A^2 and σ_T^2 does not hold.

The results are in accordance with theory. Regardless of whether or not the prior of the model is informative, the estimated asymptotic bias of $\hat{b}_\mu^{PPIC}$ is close to the true bias-correction values when $\sigma_A^2 = \sigma_T^2 = 1$, and that result is essentially unchanged when investigated under model misspecification.

4.4 Interpretation of predictive Bayes factor

Statistically, model probability itself represents a substantial measure to evaluate one statistical model against another. Bayes factor is a practical device to assign each candidate model with a conditional probability for model comparison. Table 4.1 presents the Jeffreys' proposal (pp.432, Jeffreys, 1961) to interpret the strength of evidence for standard Bayes factors in half units on the $\log_{10}$ scale, while Kass and Raftery (1995) considers the guideline by twice the natural logarithm, as shown in Table 4.2. Based on the evidence how the expected posterior probability of the model is supported by the data, here we propose a slightly modified calibration in Table 4.3 as the scale of evidence for interpretation of Bayes factors.

The difference of Table 4.3 from other proposals is not significant, however, we may use it to interpret the difference of PPICs. Generally, an individual K-L based criterion value, by itself, is not interpretable without knowing the constant $E_{\tilde{y}}[\log f(\tilde{y})]$ in equation (2.1). In practice, only the difference between the model selection criterion values is meaningful, which theoretically estimates the relative difference of the expected Kullback-Leibler divergences, a discrepancy measure of the similarity between the candidate model and the true distribution of the data. An important question for

Table 4.1: Jeffreys' scale of evidence in favor of model M_1. (Jeffreys, 1961)

B_{12}	$\log_{10} B_{12}$	$P(M_1\|y)$	Evidence
1 to 3.2	0 to 1/2	50% to 76%	Not worth more than a bare mention
3.2 to 10	1/2 to 1	76% to 90.9%	Substantial
10 to 31.6	1 to 1.5	90.9% to 96.9%	Strong
31.6 to 100	1.5 to 2	96.9% to 99%	Very Strong
> 100	> 2	$> 99\%$	Decisive

Table 4.2: Scale of evidence in favor of model M_1 by Kass and Raftery (1995).

B_{12}	$2\log_e B_{12}$	$P(M_1\|y)$	Evidence
1 to 3	0 to 2	50% to 75%	Not worth more than a bare mention
3 to 20	2 to 6	75% to 95.2%	Positive
20 to 150	6 to 10	95.2% to 99.3%	Strong
> 150	> 10	$> 99.3\%$	Very Strong

model selection is naturally raise: how big of a difference would be statistically meaningful, in the sense of when one model should no longer be considered competitive with the other?

Following

$$PrBF_{12} = \frac{exp\{\frac{1}{2} \cdot PPIC_1\}}{exp\{\frac{1}{2} \cdot PPIC_2\}} = \exp\{\frac{1}{2} \cdot (PPIC_1 - PPIC_2)\} \tag{4.19}$$

and

$$P(M_k(y)|\tilde{y}) = \frac{p(\tilde{y}|M_k(y))P(M_k(y))}{p(\tilde{y}|M_1(y))P(M_1(y)) + p(\tilde{y}|M_2(y))P(M_2(y))},\ k = 1, 2. \tag{4.20}$$

approximately we have the equation for the difference of the posterior predictive

Table 4.3: The interpretation of both predictive Bayes factor and difference of PPIC values with respect to the posterior probability in favor of model $M_1(y)$, where $PPIC_1 - PPIC_2 = 2\log_e PrBF_{12}$.

$PrBF_{12}$	$2\log_e PrBF_{12}$	$P(M_1(y)\|y)$	Evidence
1 to 3	-2.2 to 0	50% to 75%	Not worth more than a bare mention
3 to 19	-5.9 to -2.2	75% to 95%	Substantial
19 to 99	-9.2 to -5.9	95% to 99%	Strong
> 99	< -9.2	$> 99\%$	Decisive

information criterion values with

$$PPIC_1 - PPIC_2 \approx -2\log\{\frac{E(P(M_1(y)|\tilde{y}))}{E(P(M_2(y)|\tilde{y}))}\} = -2\log\{\frac{E(P(M_1(y)|\tilde{y}))}{1 - E(P(M_1(y)|\tilde{y}))}\} \quad (4.21)$$

when assuming that the prior probability $P(M_k(y))$ for each of the fitted model $M_k(y)$ satisfies $P(M_1(y)) = P(M_2(y)) = 1/2$, $k = 1, 2$. Or equivalently, for the expected probability of fitted model $M_1(y)$,

$$E(P(M_1(y)|\tilde{y})) \approx \text{logit}^{-1}\{-\frac{1}{2}(PPIC_1 - PPIC_2)\}.$$

Equation (4.21) demonstrates that the PPIC difference can be used as a summary of the evidence provided by the data for model preference in model comparison. Together with the Table 4.3, the level of model preference is quantified. What's more, we can make consistent model selection conclusion either in terms of Bayes factor or Kullback-Leibler discrepancy.

4.5 Simulation Study

Bliss (1935) reports the proportion of beetles killed after 5 hours of exposure at various concentrations of gaseous carbon disulphide in an experimental study. Here

we reprint the data in Table 4.4. After comparing the fitted probability of killed beetles as well as G^2 goodness-of-fit statistic of three generalized linear models each with a logit link, a probit link and a cloglog link, Agresti (2002) recommends the GLM with the cloglog link in an explanatory point of view.

Log Dose	1.691	1.724	1.755	1.784	1.811	1.837	1.861	1.884
Number of Beetles	59	60	62	56	63	59	62	60
Number Killed	6	13	18	28	52	53	61	60

Table 4.4: Beetles Killed after Exposure to Carbon Disulfide (Bliss 1935)

In this section we consider the same problem in the Bayesian settings, i.e., assuming a prior distribution $N(0, \tau^2)$ for each parameter of the generalized models. To predict the probability of beetles been killed, a weakly-informative prior is introduced with $\tau = 100$ so that the standard Bayes factor is well-defined and free from Lindley's paradox, as well as a strongly informative but partially mis-specified prior $\tau = 10$ since it gently deviates from what the data supports. For model comparison purpose, posterior predictive information criterion (PPIC), posterior average information criterion (PAIC, in Chapter 3), Bayesian Takeuchi information criterion (BTIC in Chaptor 2), standard Bayes factors (BF, Jeffreys 1939), posterior Bayes factors (PoBF, Aitkin 1991), pseudo Bayes factors (PsBF, Geisser and Eddy, 1979) and predictive Bayes factors (PrBF) are computed based on a very large amount of valid posterior samples ($> 100,000$) from Bugs (Spiegelhalter et al., 1994, 2003). Note that the posterior samples for pseudo Bayes factor are iteratively and independently generated for each cross-validative predictive distribution, rather than employing the importance sampling technique in which the unbounded weights may make the importance-weighted estimate unstable.

For each candidate model, we present the estimated information criteria values in Table 4.5. The result is consistent across the three Bayesian predictive criteria for

	$\tau = 100$			$\tau = 10$		
	cloglog	probit	logit	cloglog	probit	logit
PPIC	31.56	37.79	39.68	33.00	38.08	45.20
PAIC	32.32	39.52	41.12	32.27	39.31	41.08
BTIC	30.12	37.32	38.83	30.04	37.09	38.83

Table 4.5: Various Bayes factors under either weakly informative prior $\tau = 100$ or strongly informative but partially mis-specified prior $\tau = 10$.

$\tau = 100$	BF	PoBF	PsBF	PrBF
cloglog/probit	13.74 (93.2%)	12.91 (92.8%)	24.70 (96.1%)	22.56 (95.8%)
cloglog/logit	16.91 (94.4%)	30.77 (96.9%)	58.06 (98.3%)	58.11 (98.3%)
$\tau = 10$	BF	PoBF	PsBF	PrBF
cloglog/probit	11.84 (92.2%)	11.17 (91.8%)	14.15 (93.4%)	12.69 (92.7%)
cloglog/logit	2.3e11 (100.0%)	97.65 (99.0%)	986.5 (99.9%)	447.6 (99.8%)

Table 4.6: Various Bayes factors under either weakly informative prior $\tau = 100$ or strongly informative but partially mis-specified prior $\tau = 10$, as well as the corresponding probabilities that model 1 is preferred (in parentheses).

each selection of prior variance, while the GLM with cloglog link are significantly the best among all of three models.

Table 4.6 provides the comparison of various Bayes factors. All of the results indicate that the GLM with cloglog link are the best with quite confidence when conducting pairwise comparison. The pseudo Bayes factors and predictive Bayes factors are quite similar to each other, while posterior Bayes factors mildly underestimate the evidence to support the fitted model M_1 in this example. Different from the other three, the standard Bayes factors evaluate the original model specified with the prior distribution and disfavors the logit link extremely when prior variance $\tau = 10$.

Chapter 5

Conclusion and Discussion

5.1 Conclusion

In contrast to frequentist modeling, it is inevitable to include a prior distribution for parameters in each Bayesian model, either informative or non-informative, representing the antecedent believing on the parameters independent of the observed set of data. Subsequently, the ad hoc statistical inference depends on the posterior distribution $p(\theta|y) \propto L(\theta|y)\pi(\theta)$ rather than the likelihood function $L(\theta|y)$ alone; the choice of the prior distribution may cause a strong impact for models under consideration. In terms of model selection based on Kullback-Leilber divergence, it is reflected to the extent how precisely the error of in-sample estimator against out-of-sample target is corrected. Without incorporating the prior information into bias estimation, the usage of the frequentist criteria to compare Bayesian statistical models is risky, resulting in support for new Bayesian model selection proposals.

We have so far considered the evaluation of Bayesian statistical models estimated by the plug-in parameters, averaged over the posterior distributions and evaluated with respect to predictive distributions, for which BGIC, BAIC and PPIC are useful tools for model assessment. All of the new model selection criteria are proposed in

a predictive perspective for Bayesian models to asymptotically unbiasedly estimate the ad hoc Kullback-Leibler discrepancy for distinct purposes. Given some standard regularity conditions, those criteria can be widely implemented if the observations are independent and identically distributed, and generally applicable even in the case of model misspecification, i.e. $f(\tilde{y})$ does not come from the parametric family $\{g(\tilde{y}|\theta); \theta \in \Theta\}$. It is also worth to mention that the computational cost for those criteria is pretty low.

Through re-visiting the philosophy of Bayes factors, we illustrate an explanation for the cause of the AIC-type efficiency and BIC-type consistency. We also build up the link between the Kullback-Leibler discrepancy and predictive Bayes factor, by which to interpret the scale of significance for the information criterion relative values.

In our point of view, the information criteria values are good to use for reference of model performance. Unlike Akaike's minimum AIC procedure to select the 'best' model, it makes more sense to employ the proposed criteria to deselect models that are obviously poor, maintaining a subset for further consideration.

5.2 Discussion

What follows are a few related topics which are of interest to discuss for Bayesian modeling.

Bayesian Model Averaging

When the decision is not restricted to select a single model but to create a large mixture of models, Bayesian model averaging (BMA) (Draper, 1995; Raftery et al., 1997) is an approach by using individual model prior probabilities to describe model

uncertainty, weighting each single model prediction by the corresponding posterior model probability, which would be higher if the candidate model obtained the stronger support from the data.

To achieve the principle of parsimony, Madigan and Raftery (1994) propose a search strategy to exclude both the models with much smaller weights than the largest posterior probability and the complex models receiving less support from the data than their simpler counterparts. Usually only top $5\%-10\%$ of the models are selected.

Alternatively, the proposal of the predictive Bayes factors provides a set of natural weights may be used in the Bayesian model averaging, when the candidate models to consist of the final mixture have been updated by the data.

Missing Data

In the setting of missing data model, Celeux et al. (2006a) compare the performance of 8 distinct DICs, depending on various focuses of the hierarchical models and treatments on the missing variables. However, no conclusions have been drawn with respect to which of the DIC should be adopted for model selection.

In a Bayesian point of view, missing data is a special kind of unknown quantity, similar to parameters. Therefore, one solution is to conduct model selection based on the 'generalized parameters', i.e., the set containing both the missing data and parameters of interest, but treat the missing data as the ancillary.

The largest challenge of evaluating such a complicated structure is to properly measure the model complexity for bias correction when the missingness of data increases the . This problem can be properly solved by imposing our criteria.

Computation

In addition to the refinement of the theoretical methodology, it is also important to balance the efficiency and accuracy of computation in need for practical statistical analysis. Given a specific dataset and the corresponding candidate Bayesian model, the key components to apply our proposed criteria in computation consist of

1. the simulation of posterior distribution, for the posterior mean of the in-sample log-likelihood in PAIC or the log predictive posterior density in PPIC;

2. the mode of the posterior density, $\hat{\theta}$, which can be done by using methods such as conditional maximization or Newton-Raphson method;

3. the matrices $J_n(\theta)$ and $I_n(\theta)$ evaluated at the posterior mode.

The first two components are quite standard for Bayesian inference, as numerical methods played an important role in the development of Bayesian statistics. For instance, in spite of some simple non-hierarchical cases when the prior distributions are conjugate to the likelihood, it is difficult to draw the posterior distribution directly. Therefore, Markov chain Monte Carlo (MCMC) algorithms, especially those iterative simulation methods such as the Metropolis-Hasting algorithm and the Gibbs sampler (Metropolis and Ulam, 1949; Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; and Gelfand and Smith, 1990) are employed as important tools for simulation purpose. The advent of electronic computational equipment in the last a few decades has enhanced our ability to apply those computationally intensive techniques. Rather than writing specific codes to draw posterior samples with proper algorithm, some Bayesian computing software packages using MCMC algorithms are available for posterior simulation, including BUGS (Bayesian inference Using Gibbs Sampling) by Spiegelhalter et al. (1994, 2003) and JAGS (Just Another Gibbs Sampler) by Plummer (2009), both of which can be called from statistical software R after

installing corresponding libraries.

It is interesting to observe that the error correction term $tr\{J^{-1}(\theta_0)I(\theta_0)\}$ of the BGIC, PAIC and PPIC are the same, though deducted independently. In our proposals, we simply adopt the empirical $tr\{J_n^{-1}(\hat{\theta})I_n(\hat{\theta})\}$ as the estimator for the bias correction term $tr\{J^{-1}(\theta_0)I(\theta_0)\}$.

When in practice matrices $J_n(\hat{\theta})$ and $I_n(\hat{\theta})$ are difficult or tedious to determine analytically, a feasible approach is the numerical approximation using finite differences. §12.1 of Gelman et al. (2003) provides a detailed instruction to estimate the first and second derivatives of the log joint density of (y, θ). In addition, $J_n(\hat{\theta})$ is the Bayesian Hessian matrix in the optimization problem when seeking the $\hat{\theta}$ maximizing the log-posterior, a problem that there are many well-written software functions or packages are available to deal with. Furthermore, when the derivation of second derivative matrix is relatively simple, the matrix $I_n(\hat{\theta})$ can also be estimated by using the equation

$$\frac{\partial}{\partial\theta}\log p\frac{\partial}{\partial\theta'}\log p = \frac{1}{p}\frac{\partial^2}{\partial\theta\partial\theta'}p - \frac{\partial^2}{\partial\theta\partial\theta'}\log p. \tag{5.1}$$

Usually matrix $J_n(\theta)$ is fairly robust. However, the empirical Bayesian Fisher information matrix $I_n(\hat{\theta})$ might not be always computationally stable in practice, especially when the models are complex or the number of observations are small. The employment of robust estimators, for instance, proposed in Royall (1986), is valuable.

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis*, second edition. New York: John Wiley & Sons.

[2] Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society B* **53**, 111-142.

[3] Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243-247.

[4] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, ed. B. N. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado. Reprinted in *Breakthroughs in Statistics*, ed. S. Kotz, 610-624. New York: Springer-Verlag (1992).

[5] Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* **30**, 9-14.

[6] Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika* **66**, 237-242.

[7] Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16**, 125-127.

[8] Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* **94**, 443-458.

[9] Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**, 40-79.

[10] Bayarri, M. J. and Berger, J. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics* **6**, J. M. Bernardo, et. al. (Eds.) Oxford: Oxford University Press, 53-82.

[11] Bayarri, M. J. and Berger, J. O. (2000). P-values for composite null models (with discussion). *Journal of the American Statistical Association* **95** 1127-1142, 1157–1170.

[12] Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109-122.

[13] Berger, J. and Pericchi, L. (1998). Accurate and stable Bayesian model selection: the median intrinsic Bayes factor. *The Indian Journal of Statistics Society B* **60**, 1-18.

[14] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). *Model Selection.* IMS Lecture Notes - Monograph Series 38 (P. Lahiri, ed.) 135–207. IMS, Beachwood, OH

[15] Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology* **22**, 134-167.

[16] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52**, 345-370.

[17] Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78**, 131-136.

[18] Burnham K. P., Anderson, D. R. (2002). *Model selection and multimodel inference*, second edition. New York: Springer-Verlag.

[19] Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006a). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis* **70**, 651-676.

[20] Celeux, G., Forbes, F., Robert, C.P. and Titterington, D.M. (2006b). Rejoinder to 'Deviance Information Criteria for Missing Data Models'. *Bayesian Analysis* **70**, 701-706.

[21] Chen, J. and Chen, Z. (2009). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759-771.

[22] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313-1321.

[23] Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900-916.

[24] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377-403.

[25] Draper, D. (1995). Assessment and Propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57**, 45-97.

[26] Draper, N. R. and Smith, H. (1966). *Applied Regression Analysis*. Wiley, New York.

[27] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1-26.

[28] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

[29] Efroymson, M. A. (1966). Stepwise regression-a backward and forward look. presented at *the Eastern Region Meetings of the Institute of Mathematical Statistics*, Florham Park, New Jersey.

[30] Foster D. P. and George E. I. (1994) The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics* **22**, 1947-1975.

[31] Garside, M. J. (1965). The best subset in multiple regression analysis. *Applied Statistics* **14**, 196-200.

[32] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320-328.

[33] Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153-160.

[34] Gelfand, A. E. and Dey, D. (1994). Bayesian model choice: asymptotic and exact calculations. *Journal of the Royal Statistical Society B* **56**, 501-514.

[35] Gelfand, A. E. and Ghosh, S. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **85**, 1-11.

[36] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 398-409.

[37] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

[38] Gelman, A., Meng, X. L. and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**, 733-807.

[39] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

[40] George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304-1308.

[41] George, E. I. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881-889.

[42] Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B* **29**, 83-100.

[43] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A.(1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley.

[44] Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association* **96**, 1122-1132.

[45] Hannan, E. J., and Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society* **41**, 190-195.

[46] Hansen, M. and Yu, B. (2001). Model selection and Minimum Description Length principle. *Journal of the American Statistical Association* **96**, 746-774.

[47] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97-109.

[48] Hill, B. M. (1982). Lindley's paradox: Comment. *Journal of the American Statistical Association* **77**, 344-347.

[49] Hocking, R. R. (1976) The Analysis and Selection of Variables in Linear Regression. *Biometrics* **32**, 1-49.

[50] Hodges, J. S., and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367-379.

[51] Hoeting, J. A., D. M. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14**, 382-401.

[52] Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

[53] Hurvich, C. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

[54] Jeffreys, H. (1939). *Theory of Probability*, first edition. Cambridge, MA; New York: Oxford University Press.

[55] Jeffreys, H. (1961). *Theory of Probability*, third edition. Cambridge, MA; New York: Oxford University Press.

[56] Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association* **99**, 279-290.

[57] Kass, R.E. (1993) Bayes Factors in Practice, *The Statistician* **42**, 551-560.

[58] Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

[59] Kass, R. E., and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses And its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association* **90**, 928–934.

[60] Konishi, S., and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875-890.

[61] Kullback, S., and Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics* **22**, 79-86.

[62] Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society B* **57**, 247-262.

[63] Liang, H., Wu, H. and Zou, G. (2009). A note on conditional aic for linear mixed-effects models. *Biometrika* **95**, 773-778.

[64] Lindley, D.V. (1957). A Statistical Paradox. *Biometrika* **44**, 187-192.

[65] Linhart, H. and Zucchini, W. (1986). *Model selection*, New York: Wiley.

[66] Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89**, 1335-1346.

[67] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data *Internat. Statist. Rev.* **63**, 215-232.

[68] Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15**, 661-675.

[69] Mallows, C. L. (1995). More comments on Cp. *Technometrics* **37**, 362-372.

[70] McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection.* Singapore: World Scientific.

[71] Meng, X. L. and Vaida, F. (2006) Comments on 'Deviance Information Criteria for Missing Data Models'. *Bayesian Analysis* **70**, 687-698.

[72] Metropolis, N., Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association* **44**, 335-341.

[73] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21** , 1087-1092.

[74] Miller, A. J. (1990) *Subset selection in regression*, first edition. London: Chapman & Hall.

[75] Miller, A. J. (2002) *Subset selection in regression*, second edition. London: Chapman & Hall.

[76] Murata, N., Yoshizawa, S. and Amari, S. (1994). Network information criterion determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neural Networks* **5**, 865-872.

[77] Neath, A. A. and Cavanaugh, J. E. (1997). Regression And Time Series Model Selection Using Variants Of The Schwarz Information Criterion. *Communications in Statistics - Theory and Methods* **26**, 559-580.

[78] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* **12**, 758-65.

[79] O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society B* **57**, 99-138.

[80] Parzen, E. (1974). Some Recent Advances in Time Series Modeling. *IEEE Transactions on Automatic Control* **19**, 723-730.

[81] Plummer, M., (2007). JAGS: Just Another Gibbs Sampler. International Agency for Research on Cancer, Lyon, France. `http://www-fis.iarc.fr/ martyn/software/jags/`

[82] Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9**, 523-539.

[83] R Project (2002). The R project for statistical computing. `http://www.r-project.org/`

[84] Raftery, A. E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology* **25**, 111-196.

[85] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association* **92**, 179-191.

[86] Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica* **14**, 465-471.

[87] Robins, J. M., van der Vaart, A. and Ventura V. (2000). Asymptotic distribution of p-values in composite null models (with discussion). *Journal of the American Statistical Association* **95**, 1143-1156, 1171-1172.

[88] Royall, R. M. (1986). Model Robust Confidence Intervals Using Maximum Likelihood Estimators. *International Statistical Review* **54**, 221-226.

[89] Rubin, D. B. (1981). Estimation in parallel randomized experiments. *J. Educ. Statist* **6**, 377-400.

[90] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151-1172.

[91] San Martini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *Journal of the Royal Statistical Society B* **46**, 296-303.

[92] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* **6**, 461-464.

[93] Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association* **77** , 325-334.

[94] Shao, J. (1997), An Asymptotic Theory for Linear Model Selection. Statistica Sinica, 7, 221-264.

[95] Shibata, R. (1981), An Optimal Selection of Regression Variables. *Biometrika* **68**, 45-54.

[96] Shibata, R. (1984), Approximation Efficiency of a Selection Procedure for the Number of Regression Variables. *Biometrika* **71**, 43-49.

[97] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B* **64**, 583-639.

[98] Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2003). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. `www.mrc-bsu.cam.ac.uk/bugs/`

[99] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B* **36**, 111-147.

[100] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B* **39**, 44-47.

[101] Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models (in Japanese). *Mathematical Science* d**153**, 12-18.

[102] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**,267-288.

[103] Tibshirani R. and Knight, K. (1999). The Covariance Inflation Criterion for Adaptive Model Selection. *Journal Of The Royal Statistical Society Series B* **61**, 529-546.

[104] Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82-86.

[105] Tierney, L., Kass, R. E., and Kadane, J. B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association* **84**, 710-716.

[106] Vaida, F., and Blanchard, S. (2005). Conditional Akaike informaion for mixed effects models. *Biometrika* **92**, 351-370.

[107] Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation* **14**, 2339-2468.

[108] Wasserman (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**, 92-107.

[109] Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. *Journal of the American Statistical Association* **93**, 120-131.

[110] Wei, C. Z. (1992). On Predictive Least Squares Principles. *The Annals of Statistics* **20**, 1-42.

[111] Wherry, R. J. (1931). A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation. *Annals of Mathematical Statistics* **2** , 440-457.