Approximation of the Formal Bayesian Model Comparison using the Extended Conditional Predictive Ordinate Criterion

by

Md Rashedul Hoque

M.S. in Applied Statistics, University of Dhaka, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

August 2017

Abstract

The optimal method for Bayesian model comparison is the formal Bayes factor (BF), according to decision theory. The formal BF is computationally troublesome for more complex models. If predictive distributions under the competing models do not have a closed form, a cross-validation idea, called the conditional predictive ordinate (CPO) criterion can be used. In the cross-validation sense, this is a "leave-out one" approach. CPO can be calculated directly from the Monte Carlo (MC) outputs, and the resulting Bayesian model comparison is called the pseudo Bayes factor (PBF). We can get closer to the formal Bayesian model comparison by increasing the "leave-out size", and at "leave-out all" we recover the formal BF. But, the MC error increases with increasing "leave-out size". In this study, we examine this for linear and logistic regression models.

Our study reveals that the Bayesian model comparison can favour a different model for PBF compared to BF when comparing two close linear models. So, larger "leave-out sizes" are preferred which provide result close to the optimal BF. On the other hand, MC samples based formal Bayesian model comparisons are computed with more MC error for increasing "leave-out sizes"; this is observed by comparing with the available closed form results. Still, considering a reasonable error, we can use "leave-out size" more than one instead of fixing it at one. These findings can be extended to logistic models where a closed form solution is unavailable.

Lay Summary

The purpose of the model comparison is to find a useful model among some possible models. There are different model selection methods available in the literature, developed by the frequentist and Bayesian schools. The main goal of this thesis is to examine a model selection method based on cross-validation as an alternative to the formal Bayesian model comparison. We demonstrate the behavior of this model comparison tool for both simple and complex models in the Bayesian context. The major finding of the thesis suggests using a general version of the widely used cross-validation approach, but with a larger "leave-out size" than one, to get closer to the formal Bayesian model comparison.

Preface

I did the analyses in this thesis (Chapter 3 and Chapter 4) under the supervision of my supervisor, Professor Paul Gustafson. Professor Gustafson instructed me to work according to the objectives set for the thesis and made corrections to my simulation setup and choice of statistical tool. We have not submitted any manuscript for publication from this yet, but we are planning to do soon. The manuscript will summarize the major findings of the thesis incorporating concise content from Chapter 2, Chapter 3, and Chapter 4.

Table of Contents

List of Tables

List of Figures

Acknowledgments

At first, I would like to thank my supervisor Professor Paul Gustafson for his kind supervision. He not only supervised in this academic work but also guided me in some ups and downs of my study during my stay in UBC. That boosted my inspiration and helped me a lot to concentrate on the research. The faculty members in the Department of Statistics are very nice and friendly. Especially, I want to thank Professor Lang Wu for his valuable inputs as a second reader.

Also, I would like to thank my wife, Fatema Tuz Jhohura for her constant support during my writing. She is also a graduate student here in the Department of Statistics, UBC. The staff and graduate students are friendly and helped me a lot during this thesis. I discussed many ideas and possibilities for this thesis output with Joe Watson and Qiong Zhang. Both of them are graduate students in the Department of Statistics, and my office mates too. I am grateful to them. I would like to thank all of the graduate students in the Department who shared their ideas about my research.

Finally, I think I am lucky to be a part of the Department of Statistics, UBC. My stay is brief here, but I feel this is my second home. I am grateful to everyone in the Department for their kind and friendly support during my study and research.

Chapter 1

Introduction

Statisticians build models to establish some relationship from the available data, and that can be used to predict future data. There might be multiple possibilities for building models from an available existing data set. Then, the next task is to find the most useful model among the possible models using a model selection method. Model selection is very important in Statistics. Both frequentist and Bayesian schools have developed many selection methods for model comparisons. This thesis focuses on the Bayesian model selection methods.

1.1 Problem Statement

Some Bayesian model selection methods are available in the literature for comparing models (Gelfand and Dey, 1994). Among these, a popular choice is the formal Bayesian comparison which simply uses the Bayes factor. As an optimal model comparison tool, the Bayes factor is the desired model selection method in Bayesian paradigm. But, calculation of the Bayes factor is not easy for complex models. In that case, one can use alternatives to the Bayes factor that are easily computable. One of these alternatives is to use the cross-validation approach based on the so-called conditional predictive ordinate criterion (discussed in detail in Chapter 2); this can be computed easily from Monte Carlo posterior samples. The cross-validation approach with "leave-out one" is widely used as an approximation of the formal Bayes factor (Geisser and Eddy, 1979), whereas the formal

Bayes factor can be computed mathematically using this cross-validation approach with "leave-out all". Mathematically, increasing "leave-out sizes" will allow us to compute more closely to formal Bayesian comparison than the commonly used "leave-out one". This mathematical closeness comes with a price, the Monte Carlo error.

The cross-validation approach requires posterior samples to compute the (approximate) Bayes factors. As the cross-validation approach with "leave-out all" requires more computations, the results come with more Monte Carlo error than the "leave-out one". In general, the Monte Carlo error increases with increasing "leave-out sizes". So, we have an optimization problem here. We want to examine how the cross-validation approach works to approximate the optimal formal Bayesian comparison for some "leave-out sizes" from "leave-out one" to "leave-out all". Also, if possible we want to find a "leave-out size" at which we have closer to formal Bayesian comparison than the "leave-out one" with only a little increase in Monte Carlo error. According to this problem, we formulate the objectives of our study in the next section.

1.2 Objectives

According to the problem statement, we can break down the objectives of this study in several stages. We list those below:

1. At first, we want to find relevant pieces of literatures if any regarding this issue.

2. Our second objective is to examine the model comparisons for simple models (say based on normal distributions) where Monte Carlo is not needed. Then, our objective is to examine how closely the Monte Carlo samples compute the closed form results for the model comparison tools considered. For the simple models, we also want to examine how rapidly does Monte Carlo error "kills us" as "leave-out size" increases.

3. Our next objective is to examine how bad it is to not do an optimal formal Bayesian model comparison. Here our interest is to examine whether it is

common/rare to see a switch in the winning model as "leave-out size" increases.

4. Our final objective is to find a practical advice for more complex models where we really need to use Markov chain Monte Carlo output. Particularly, we want to examine whether we can look at different "leave-out sizes" and report Monte Carlo error.

In general, our over-arching goal is to examine the cross-validation approach as an alternative to the formal Bayesian model comparison. Also, we want to get some interesting insight about "leave-out size" (other than 1) for achieving a closer result to the optimal solution with small Monte Carlo error. Especially, we want to examine this for more complex models when the formal Bayesian model comparisons are hard to compute.

1.3 Organization of the Thesis Report

The general idea of the problem we are interested in and the objectives of our study have been discussed in this Chapter. In Chapter 2 we formulate the problems more mathematically and describe the model selection methods we use in this study. A literature review of the model comparison tools is also given there. Chapter 3 focuses on the Bayesian model comparison of linear regression models; the closed form and the Monte Carlo samples based results are documented, and we examine those according to the objectives. We discuss the Bayesian model comparisons for the generalized linear models, in particular, logistic regression models, in Chapter 4. Overall findings and some concluding remarks are discussed in brief in Chapter 5.

Chapter 2

Bayesian Model Comparison

2.1 Introduction

Researchers collect data on some variables to study the effect of these variables on some outcome of interest. Then the question arises of which variables are important to explain the variation in the outcome. Also, inclusion of interactions between the variables might be an interesting question to the researchers. These are model selection problems and statisticians have proposed many approaches to deal with the issue of model selection. Some popular and well-known model selection methods are Akaike information criterion (AIC), Mallows CP, likelihood ratio tests for nested models, stepwise selection procedures (backward or forward selection), cross-validation, different types of Bayes factors (intrinsic, partial, pseudo, posterior), Bayesian information criterion (BIC), and Bayesian model averaging. These methods work in different ways. For example, some of these methods are just algorithms for choosing a useful model (e.g. stepwise selection). Other methods are based on the criteria to judge the quality of a model (e.g. AIC, BIC).

2.2 Classical View of Model Selection

We discuss the classical approach to model selection here. Suppose we want to choose a model between two parametric models $M_i,\ i = 1,2$. These two models are denoted by the joint density $f(\boldsymbol{y}|\boldsymbol{\theta}_i;M_i)$ or likelihood $L(\boldsymbol{\theta}_i;\boldsymbol{y},M_i)$, $i = 1,2$. Here,

$\boldsymbol{\theta}_i$ is a $p_i \times 1$ parameter vector and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is a $n \times 1$ outcome vector.

Classical Neyman-Pearson theory is applied to nested models where models are compared pairwise for model selection. Suppose, our hypotheses are H_i: data $\boldsymbol{y}$ correspond to the model M_i, $i = 1, 2$. For example, we set H_1 as the null hypothesis. Then, the likelihood ratio test can be used to compare these models by specifying M_1 and M_2 as the reduced and full model, respectively. The reduced model is nested within the full model. With the estimated parameter vectors $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ from Models M_1 and M_2, the test statistic of the likelihood ratio test has the following form:

$$\lambda_n = \frac{L(\hat{\boldsymbol{\theta}}_1; \boldsymbol{y}, M_1)}{L(\hat{\boldsymbol{\theta}}_2; \boldsymbol{y}, M_2)}. \tag{2.1}$$

The null hypothesis H_1 can be rejected if $\lambda_n < c < 1$, where $0 < c < 1$ is a constant. Also, under H_1, $-2\log\lambda_n$ has an approximate $\chi^2_{p_2-p_1}$ distribution. Sometimes the reduced models are rejected though they are actually true, especially when λ_n tends to be very small, i.e.,

$$\lim_{n\to\infty} \Pr\left(\text{select}\, M_2 | M_1 \text{ True}\right) = \Pr(\chi^2_{p_2-p_1} > -2\log c) > 0.$$

However, how small turns out to be too small is dependent on the significance level of the test, that is how much tolerance is considered for the probability of Type I error. In general, the likelihood ratio test has a preference on the full model than the reduced model at a smaller level of significance.

To deal with this problem, many penalization techniques of the log-likelihood in the form of $\log L(\hat{\boldsymbol{\theta}}_i; \boldsymbol{y}, M_i) - k(n, p_i)$ have been proposed so that the largest penalized log-likelihood wins. Here, $k(n, p_i) > 0$ and this is a increasing function of n and p implying more penalization for the big model than the nested small model. Incorporating this, λ_n in equation 2.1 is extended to $\log(\lambda_n) + k(n, p_2) - k(n, p_1)$. Many model selection procedures, including AIC and BIC are different versions of this expression. We need $k(n, p_2) - k(n, p_1) \to \infty$ as $n \to \infty$ for the selection of the true underlying model consistently with increasing n. The most common form for $k(n, p)$ found in the literature is $k(n, p) = \alpha p$. Akaike (1973) uses values of α in the interval $1 \leq \alpha \leq 2.5$ whereas Aitkin (1991) suggests $\alpha = \log 2$. However, these approaches produce some inconsistent result. Schwarz et al. (1978)

suggests $k(n,p) = (p/2)\log n$ where k depends on n and eliminates inconsistency. Some other versions are suggested by Nelder and Baker (1972), Hannan and Quinn (1979) and Shibata (1980).

2.3 Bayesian View of Model Selection

We describe the choice of a model among a possible set of models in Bayesian paradigm.

We begin with the Bayes factor, the formal Bayes approach to compare two models. Different versions of the Bayes factor are also available in the literature. One can consider reading Gelfand et al. (1992), Kadane and Lazar (2004), and their attendant discussions. In addition, implementation of the asymptotic and exact methods are described with examples in Gelfand and Dey (1994).

In a Bayesian model, we need to specify prior $p(\boldsymbol{\theta})$ in addition to the likelihood specification. All the inference is made only from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta};\boldsymbol{y}) \times p(\boldsymbol{\theta})$. Here $\boldsymbol{\theta}$ are the parameters of interest, and $L(\boldsymbol{\theta};\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ stands for the likelihood function. In the Bayesian paradigm, different approaches have been proposed for selection of models. Some of these are based on posterior distributions and some others instead focus on predictive distributions.

The two model components might be not fixed in a Bayesian model selection problem. Sometimes the likelihood L is held fixed, and only the prior $p(\boldsymbol{\theta})$ is varied; this is used to check Bayesian robustness (Berger, 2013) by assessing how sensitive the posterior is due to prior variation. Sometimes, the likelihood L is varied. Now, the formal Bayesian model selection procedure follows in the following subsection using Bayes factor.

2.3.1 General Version of Bayes Factor

We use the same notation as discussed in the classical approaches in section 2.2. The sampling density for model M_1 is $f(\boldsymbol{y}|\boldsymbol{\theta}_1, M_1)$ and the competing model is M_2 with sampling density $f(\boldsymbol{y}|\boldsymbol{\theta}_2, M_2)$. There is no need to specify something common between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For example, $f(\boldsymbol{y}|\boldsymbol{\theta}_1, M_1)$ might be the density of a gamma with parameters $(\theta_{11}, \theta_{12})$ and $f(\boldsymbol{y}|\boldsymbol{\theta}_2, M_2)$ might be the density of a log-normal with parameters $(\theta_{21}, \theta_{22})$.

Suppose, the prior distributions $p_1(\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{\theta}_2)$ are specified for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Also, let the prior probability of M_i is w_i, $i = 1,2$, with $w_2 = 1 - w_1$. We consider the comparison of M_1 versus the alternative M_2. Now, a Bernoulli random variable M for the models taking values 0 and 1 is defined and the joint density for this comparison can be written as:

$$p(\boldsymbol{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, M) = f(\boldsymbol{y}|\boldsymbol{\theta}_1, M_1)\, p_1(\boldsymbol{\theta}_1)\, w_1\, I_0(M) + f(\boldsymbol{y}|\boldsymbol{\theta}_2, M_2)\, p_2(\boldsymbol{\theta}_2)\, w_2\, I_1(M).$$

We compute $f(\boldsymbol{y}|M_i)$, the predictive density for model, M_i, as

$$f(\boldsymbol{y}|M_i) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_i, M_i)\, p_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i. \tag{2.2}$$

Suppose $\boldsymbol{y}_{obs}$ denotes the observed data. Using Bayes theorem, the posterior probability for model M_1 can be written as

$$\Pr(M = 0|\boldsymbol{y}_{obs}) = \frac{w_1\, f(\boldsymbol{y}_{obs}|M_1)}{w_1\, f(\boldsymbol{y}_{obs}|M_1) + w_2\, f(\boldsymbol{y}_{obs}|M_2)}.$$

Now, the posterior odds of model M_1 with respect to model M_2 are

$$\begin{aligned}
\frac{\Pr(M = 0|\boldsymbol{y}_{obs})}{\Pr(M = 1|\boldsymbol{y}_{obs})} &= \frac{\frac{w_1\, f(\boldsymbol{y}_{obs}|M_1)}{w_1\, f(\boldsymbol{y}_{obs}|M_1) + w_2\, f(\boldsymbol{y}_{obs}|M_2)}}{\frac{w_2\, f(\boldsymbol{y}_{obs}|M_2)}{w_1\, f(\boldsymbol{y}_{obs}|M_1) + w_2\, f(\boldsymbol{y}_{obs}|M_2)}} \\
&= \frac{w_1\, f(\boldsymbol{y}_{obs}|M_1)}{w_2\, f(\boldsymbol{y}_{obs}|M_2)} \\
&= \frac{w_1}{w_2} \times \frac{f(\boldsymbol{y}_{obs}|M_1)}{f(\boldsymbol{y}_{obs}|M_2)} \\
&= \text{prior odds} \times BF,
\end{aligned} \tag{2.3}$$

where the Bayes factor (of model M_1 with respect to model M_2) denoted by BF, is expressed as

$$BF = \frac{f(\boldsymbol{y}_{obs}|M_1)}{f(\boldsymbol{y}_{obs}|M_2)}. \tag{2.4}$$

Thus, equation (2.3) demonstrates a relationship between the posterior odds, prior odds w_1/w_2 and the Bayes factor. Jeffreys (1961) and Pettit and Young (1990) give a scale for interpretation of the Bayes factor (of model M_1 with respect to model

Table 2.1: Interpretation of the Bayes factor (BF) values

Value of Bayes factor	Strength of evidence
$< 10^0$	Negative (supports M_2)
10^0 to $10^{1/2}$	Barely worth mentioning
$10^{1/2}$ to 10^1	Substantial
10^1 to $10^{3/2}$	Strong
$10^{3/2}$ to 10^2	Very strong
$> 10^2$	Decisive

M_2)[see Table 2.1].

The interpretation of the Bayes factor (BF) is straightforward and easily understandable for choosing between two models. Bayes factor does not require that the two models being compared are nested. Also, model fitting is not required for computing Bayes factor.

2.3.2 Criticisms of Bayes Factor

The Bayes factor has some limitations as a model comparison tool. One limitation is related to the specification of a prior distribution. Depending on the locations of the priors for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the Bayes factor may lead to change the decision on which model is favoured. Even when the priors are proper, the Bayes factors have non-robustness issues for the prior specification. Also, if the prior distribution $p(\boldsymbol{\theta})$ is improper as a non-informative specification, then the density function $f(\boldsymbol{y})$ is improper as well. So, $f(\boldsymbol{y}|M_i)$ cannot be interpreted as the densities of these models which in turn imply the non-interpretability of the Bayes factor ratio.

Another limitation is that the well-known Lindley's paradox (Lindley, 1957) may appear in the presence of an improper prior. Then, according to Lindley's paradox, it is unlikely that Model M_2 will be chosen as the sample size n grows large. Thus, the Bayes factor has a contradiction with the likelihood ratio test which provides way too much support for the model under alternative (i.e. Model M_2). Smith and Spiegelhalter (1980) incorporate the idea of local Bayes factor to overcome the 'Lindley's Paradox' where non-decreasing prior probabilities are assigned to an appropriate local neighbourhood of the parameters $\boldsymbol{\theta}$.

2.3.3 Other Versions of Bayes Factor

The automatic Bayesian methods are suggested by some authors for model selection to cope with the improper prior related problems such as Lindley's paradox, Bayes factor dependency on prior specification and complexities in calculation and interpretation of the Bayes factor. Note that in the automatic Bayesian methods, users are not required to specify the hyperparameters; rather, there is an algorithm to set the hyperparameters. Berger and Pericchi (1996) and Laud and Ibrahim (1995) argue that the automatic methods are essential as in practice, proper (or subjective) prior specification wouldn't be feasible for a wide range of models that are initially considered. On the other hand, Lindley (1997) argues that objective priors (reference or non-informative priors are often improper) are not commonly used in practice; the author also mentions the absence of sensible interpretation for model selection in the presence of improper priors. However, this controversy of prior specification continues. To overcome this critical activity, different methodologies are proposed. In particular, the main reason to do this is to avoid the difficulties of Bayes factor with improper or vague priors.

The intrinsic Bayes factor, a version of the Bayes factor is proposed by Berger and Pericchi (1996). To construct this, the data is needed to divide into two parts. Those parts are regarded as training and test data. Then, to compute the Bayes factor, one can consider the testing data as the data and the posterior distributions using the training data as the prior. With $\boldsymbol{y}(l)$ and $\boldsymbol{y}(-l)$ as a training sample and a test sample respectively, an intrinsic Bayes factor denoted by BF_{int}, for the training sample $\boldsymbol{y}(l)$ is defined as

$$BF_{int}(l) = \frac{f(\boldsymbol{y}_{(-l)}|\boldsymbol{y}_{(l)}, M_1)}{f(\boldsymbol{y}_{(-l)}|\boldsymbol{y}_{(l)}, M_2)}. \tag{2.5}$$

Here $f(\boldsymbol{y}_{(-l)}|\boldsymbol{y}_{(l)}), M_i,\ i = 1,2$ represents the marginal density of the testing sample. As a popular choice, a minimal training sample is used as a training sample. But, a given data set usually has more than one minimal training sample. In that case, one possible option might be to use the arithmetic or geometric averages of the intrinsic Bayes factors that are computed using the available minimal training samples of the data. Berger and Pericchi (1996) also discuss some versions of the

intrinsic Bayes factor.

The average version of the intrinsic Bayes factor is a bad choice to use for a large data set as we might ended up averaging over many minimal training sets available for that data. Sometimes, due to difficulties with minimal training sample, intrinsic Bayes factors fail to discriminate between competing models (O'Hagan, 1997). An alternative approach is the fractional Bayes factor (O'Hagan, 1995). To illustrate this approach, let us denote a fraction b as the ratio of the size of the training sample (u) to the size of the entire data set (v). The fractional Bayes factor, denoted by BF_{frac} in this case, can be defined as

$$BF_{frac} = \frac{M_1(b,\boldsymbol{y})}{M_2(b,\boldsymbol{y})}, \tag{2.6}$$

where

$$M_i(b,\boldsymbol{y}) = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta}_i)\, p_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int f(\boldsymbol{y}|\boldsymbol{\theta}_i)^b\, p_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}.$$

It is important to remember that the motivation for the fractional Bayes factor is simply asymptotic (in u and v). O'Hagan (1997) shows that fractional Bayes factors have many similar properties to ordinary Bayes factors, like adherence to the likelihood principle, and invariance to data transformation, which are not enjoyed by intrinsic Bayes factors.

Berger and Pericchi (1998) introduces median intrinsic Bayes factor with two different versions. In the first version, the authors use the median instead of the mean (arithmetic or geometric) over the training data. Now, the median intrinsic Bayes factor has the form

$$BF_{int}^{M} = \text{median}[BF_{int}(l)], \tag{2.7}$$

where $BF_{int}(l)$ is defined in equation (2.5). The second version of the median intrinsic Bayes factor has the form

$$BF_{int}^{RM} = \frac{\text{median}[f(\boldsymbol{y}_{(-l)}|\boldsymbol{y}_{(l)}, M_1)]}{\text{median}[f(\boldsymbol{y}_{(-l)}|\boldsymbol{y}_{(l)}, M_2)]}. \tag{2.8}$$

Berger and Pericchi (1998) argues that BF_{int}^{M} and BF_{int}^{RM} are stable compared to the

intrinsic BF, they proposed earlier.

The posterior Bayes factor is defined by Aitkin (1991) which uses the the posterior distribution $p_i(\boldsymbol{\theta}_i|\boldsymbol{y})$ as a replacement of the prior distribution $p_i(\boldsymbol{\theta}_i)$ in the Bayes factor formulation (equation 2.4). However, this approach has some criticisms related to double use of the data and use of posterior as prior. These criticisms are discussed by many authors in the discussion of Aitkin (1991).

2.4 Predictive Distribution as Comparative Tool

In model selection, some particular forms of predictive distributions have been used within the Bayesian approach for a long time. Box (1980) argues that conditional on the model adequacy, the posterior distribution is utilized to estimate the model parameters. On the other hand, the criticisms of the model given the existing data can be obtained using the predictive distribution (Box, 1980). In addition, the predictive distributions of two models will be comparable, not the posteriors while examining those two models.

Many approaches to model selection have been suggested by using predictive criteria instead of Bayes factors over the years since the 1970's. The idea of a pseudo Bayes factor arises by using cross-validation ideas (Geisser, 1975; Stone, 1974). Some predictive ideas are already incorporated in the discussed intrinsic Bayes factors and posterior Bayes factors.

A predictive density emerges by averaging a likelihood defined in the sample space with respect to the updated prior based on the data (that is the posterior). Suppose the data $\boldsymbol{y}$ is a collection of conditionally (given $\boldsymbol{\theta}$) independent univariate observations $y_j,\ j = 1,\ldots,n$. Also, suppose under Model M_i, y_j has density $f(y_j|\boldsymbol{\theta}_i, M_i)$, $i = 1,2$, and let J_n denote the set $\{1,\ldots,n\}$, with S as an arbitrary subset of J_n. We define the likelihood as:

$$L(\boldsymbol{\theta}_i; \boldsymbol{y}_s, M_i) = \prod_{j=1}^{n} f(y_j|\boldsymbol{\theta}_i, M_i)^{d_j},$$

where indicator function $d_j = 1$ if $j \in S$ or $d_j = 0$ if $j \notin S$. Similarly as section 2.3.1, let $p_i(\boldsymbol{\theta}_i),\ i = 1,2$, be the prior density under Model M_i. Now, the formal conditional density can be considered as

$$f(\boldsymbol{y}_{s_1}|\boldsymbol{y}_{s_2},M_i) = \int L(\boldsymbol{\theta}_i;\boldsymbol{y}_{s_1},M_i)\, p_i(\boldsymbol{\theta}_i|\boldsymbol{y}_{s_2})\, d\boldsymbol{\theta}_i \tag{2.9}$$
$$= \frac{\int L(\boldsymbol{\theta}_i;\boldsymbol{y}_{s_1},M_i)\, L(\boldsymbol{\theta}_i;\boldsymbol{y}_{s_2},M_i)\, p_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int L(\boldsymbol{\theta}_i;\boldsymbol{y}_{s_2},M_i)\, p_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i},$$

where S_1 and S_2 are arbitrary subsets of J_n. Equation (2.9) represents a predictive density; here the joint density of $\boldsymbol{y}_{s_1}$ is averaged over the prior distribution of $\boldsymbol{\theta}_i$ updated by the portion the data $\boldsymbol{y}_{s_2}$. Equation (2.9) represents a general formulation of predictive approach for Bayesian model selection. Using different specifications for S_1 and S_2, we can obtain different predictive distributions used in Bayesian model selection approaches found in the literature. We discuss some of these approaches in next subsections.

2.4.1 Different Predictive Distribution Based Approaches

Several examples of density (2.9) can be obtained in the literature. All of these examples vary in specification of the subsets S_1 and S_2 of J_n. We list some of these examples here, which are discussed by Gelfand and Dey (1994).

(i) If $S_1 = J_n$ and $S_2 = \phi$, then $f(\boldsymbol{y}_{s_1}|\boldsymbol{y}_{s_2},M_i)$ becomes the standard marginal density of the data. In this case, the denominator integral is not considered. Clearly, this produces the general Bayes factor given in equation (2.4).

(ii) If $S_1 = \{r\}$ and $S_2 = J_n - \{r\}$, then a cross-validation density results. This leads to the conditional predictive ordinate criterion which eventually produces the pseudo Bayes factor. We discuss this approach in detail in the next subsection.

(iii) If S_1 is considered as a subset of J_n, usually some (> 1) elements of J_n and $S_2 = J_n - S_1$, then we have an extended version of (ii) (Pena and Tiao, 1992). For our study, we focus on this extension.

(iv) The choice $S_1 = J_n$ and $S_2 = J_n$ indicate posterior predictive density (Aitkin, 1991). Posterior Bayes factor is produced directly by using (iv).

(v) Another choice is to specify $S_1 = J_n - S_2$ and $S_2 = \{1, \ldots, [\rho n]\}$, where [] represents the greatest integer function. A proportion ρ of the observations n is used for updating prior distribution whereas the observations $(1-\rho) \times n$ are used for model selection.

(vi) If $S_1 = J_n - S_2$ and S_2 is nothing but a minimal subset (Berger and Pericchi, 1996) with a proper density $p_i(\boldsymbol{\theta}_i|\boldsymbol{y}_{s_2})$, then $f(\boldsymbol{y}_{s_1}|\boldsymbol{y}_{s_2}, M_i)$ becomes proper. Several versions of intrinsic Bayes factor can be developed from (vi).

Among the six different specifications of S_1 and S_2, (i) and (vi) are quite different in terms of asymptotic behavior than the rest of the specifications $(ii) - (v)$. The cardinality of S_2 approaches infinity as sample size $n \to \infty$. We discuss conditional predictive ordinate criterion noted in (ii) in detail in the next subsection as this is our primary interest.

2.4.2 Conditional Predictive Ordinate Criterion and Pseudo Bayes Factor

The conditional predictive ordinate criterion is obtained by specifying the subsets of J_n as $S_1 = \{r\}$ and $S_2 = J_n - \{r\}$. This specification yields the cross-validation density $f(y_r|\boldsymbol{y}_{-r}, M_i)$ by putting the values of S_1 and S_2 in density (2.9) where $\boldsymbol{y}_{-r} = (y_1, y_2, \ldots, y_{r-1}, y_{r+1}, \ldots, y_n)$ (Stone, 1974; Geisser, 1975). According to Geisser (1980), this cross-validation density $f(y_r|\boldsymbol{y}_{-r}, M_i)$ is popularly known as the conditional predictive ordinate (CPO) when evaluated at the observed $\boldsymbol{y}_{obs}$. Also, Geisser and Eddy (1979) propose the product of these cross-validation densities (or Bayesian predictive densities) $\prod_{r=1}^{n} f(y_r|\boldsymbol{y}_{-r}, M_i)$ as a proxy for sampling density $f(\boldsymbol{y})$.

It should be noted that $\prod_{r=1}^{n} f(y_r|\boldsymbol{y}_{-r}, M_i)$ is built by treating the y_r's as predictively independent conditional on the parameters. This product is a compromise between Bayesian and non-Bayesian methods in some ways. Firstly, the product of the conditional predictive densities $f(y_r|\boldsymbol{y}_{-r}, M_i),\ r = 1, \ldots, n$ is used instead of joint predictive density for the ease of the computational complexity. Secondly, the conditional predictive density of y_r depends on both the $\boldsymbol{y}_{-r}$ and prior distribution of $\boldsymbol{\theta}_i$ whereas the joint predictive density depends only on prior distribution. In addition, as we discussed earlier, the joint predictive density of $\boldsymbol{y}$ becomes improper if

an improper prior distribution is used. These problems are addressed when instead we use the joint cross-validation densities $\prod_{r=1}^{n} f(y_r|\boldsymbol{y}_{-r}, M_i)$.

Using the idea of CPO, we can construct the Bayesian model selection tool, pseudo Bayes factor of Model M_1 with respect to Model M_2 (denoted by PBF here) as suggested by Geisser and Eddy (1979):

$$PBF = \frac{\prod_{r=1}^{n} f(y_r|\boldsymbol{y}_{-r}, M_1)}{\prod_{r=1}^{n} f(y_r|\boldsymbol{y}_{-r}, M_2)}. \tag{2.10}$$

The log version of the equation (2.10) is used commonly for computational simplicity which follows:

$$\log PBF = \sum_{r=1}^{n} \log f(y_r|\boldsymbol{y}_{-r}, M_1) - \sum_{r=1}^{n} \log f(y_r|\boldsymbol{y}_{-r}, M_2). \tag{2.11}$$

The cross-validation densities used to form the pseudo Bayes factor lead to an interesting asymptotic approximation. Log pseudo Bayes factor can be approximated to a quantity that is the summation of the logarithm of likelihood ratio test statistic and a function of the parameters in two competing models for a large sample size (n). We can write as $n \to \infty$, then

$$\log PBF \approx \log \lambda_n + \frac{p_2 - p_1}{2}.$$

Here λ_n represents the likelihood ratio test statistic and p_1 and p_2 are number of parameters for Model M_1 and M_2. This asymptotic approximation leads to a bridge between the Bayesian (pseudo Bayes factor) and non-Bayesian (likelihood ratio test statistic) methods which strengthen the motivation of using pseudo Bayes factor as a model comparison tool.

CPO criterion can be called as a criterion based on "leave-out one" in the cross-validation sense. Hence, the pseudo Bayes factor is also based on "leave-out one". We tag these "leave-out one" as in the cross-validation (or predictive) density $f(y_r|\boldsymbol{y}_{-r}, M_i)$, y_r is conditional on all elements of $\boldsymbol{y}$ except y_r (implying leaving out y_r). Only one element of $\boldsymbol{y}$ is considered as leave-out in this approach for which the conditional cross-validation density is formed. From examples (i) and (ii) of Bayesian predictive distribution based approaches discussed in subsec-

tion 2.4.1, it is clear that, if we consider $r = J_n$, then $J_n - r = \phi$ which leads (ii) to be turned into (i); this implies that with "leave-out all", pseudo Bayes factor is nothing but the formal Bayes factor. Now what happens in between, i.e., between "leave-out one" and "leave-out all"? To give insight into this matter, we need to extend the idea of CPO criterion for different "leave-out sizes" which is discussed in the next subsection.

2.4.3 Extending the CPO Criterion

CPO criterion is defined for "leave-out one" case in the cross-validation sense which is already discussed in the previous subsection. Now, we discuss the extension of CPO criterion with different leave-out options. That will enable us to examine the model selection between the formal Bayes factor and the pseudo Bayes factor.

The motivation for examining this is to explore that how well the Bayesian model comparison using Bayes factor can be approximated by the extended CPO criterion with different leave-out options. Suppose we write the log-likelihood of the cross-validation densities for CPO for "leave-out one" with data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and Model $M_i,\ i = 1, 2$ as

$$m^{(1)} = \frac{1}{n} \sum_{r=1}^{n} \log f(y_r | \boldsymbol{y}_{-r}, M_i). \tag{2.12}$$

For two leave-out points r and t, we can define

$$m^{(2)} = \frac{1}{2} \binom{n}{2}^{-1} \sum_{r<t}^{n} \log f(\boldsymbol{y}_{r,t} | \boldsymbol{y}_{-(r,t)}, M_i), \tag{2.13}$$

where $\boldsymbol{y}_{r,t} = (y_r, y_t)$ and $\boldsymbol{y}_{-(r,t)} = (y_1, \ldots, y_{r-1}, y_{r+1}, \ldots, y_{t-1}, y_{t+1}, \ldots, y_n)$. The multiplier $\binom{n}{2}^{-1}$ is added to take average over all possible combinations of $\boldsymbol{y}_{r,t}$ and multiplier $\frac{1}{2}$ is used for adjustment due to two leave-out points. Equation 2.13 can be re-written as,

$$m^{(2)} = \binom{n}{2}^{-1} \sum_{r<t}^{n} \frac{1}{2} \left\{ \log f(\boldsymbol{y}_r | \boldsymbol{y}_{-(r,t)}, M_i) + \log f(\boldsymbol{y}_t | \boldsymbol{y}_{-(r)}, M_i) \right\}.$$

In general, with k leave-out, the adjustment factor will be $1/k$. Following the pattern, for three leave-out points r, t and s we can define

$$m^{(3)} = \frac{1}{3}\binom{n}{3}^{-1} \sum_{r<t<s}^{n} \log f(\boldsymbol{y}_{r,t,s}|\boldsymbol{y}_{-(r,t,s)}, M_i), \tag{2.14}$$

and so on for increasing number of leave-out points up to the sample size n. If we want to compare two models (1 and 2), the difference $\gamma_k = m_1^{(k)} - m_2^{(k)}$ has a similar form to the log Bayes factor; in a cross-validation sense of "leave-out k", this difference represents how much better/worse model 1 predicts than model 2.

We know that $k = 1$ corresponds to CPO criterion and $k = n$ corresponds to formal Bayesian model comparison using Bayes factors; both of these are already discussed. The marginal density of the data is obtained from $\exp(n\,m^{(n)})$. Hence, the log Bayes factor of M_1 with respect to M_2 is represented by $n(m_1^{(n)} - m_2^{(n)})$ whereas the log pseudo Bayes factor of M_1 with respect to M_2 is represented by $n(m_1^{(1)} - m_2^{(1)})$. In this study, I wish to examine whether the log of extended pseudo Bayes factor (PBF), $n(m_1^{(k)} - m_2^{(k)})$, $1 \leq k < n$ can be treated as approximation of log Bayes factor when the real Bayes factor comparisons are desired but hard to compute.

2.4.4 Computation of the Extended CPO Criterion

The CPO criterion is well used as it is easy to compute directly from Monte Carlo (MC) or Markov chain Monte Carlo (MCMC) output. Suppose we know that the elements of the data $\boldsymbol{y}$ are conditionally independent given the parameter vector $\boldsymbol{\theta}$. Then using MCMC technique, Monte Carlo samples can be obtained from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. After that, one can express the cross-validation density (for "leave-out one" with $k = 1$) $f(y_r|\boldsymbol{y}_{-r})$, $r = 1,\ldots,n$ as a function of posterior mean that is computed from the posterior with all data points (Newton and Raftery, 1994). Then, we can express $f(y_r|\boldsymbol{y}_{-r})$ as

$$f(y_r|\boldsymbol{y}_{-r}) = E\{f(y_r|\boldsymbol{\theta})^{-1}|\boldsymbol{y}\}^{-1}. \tag{2.15}$$

The right hand side of the equation (2.15) shows the posterior harmonic mean of the likelihood, so Raftery et al. (2006) term it the 'harmonic mean identity'. The

authors suggest to approximate the cross-validation density $f(y_r|\boldsymbol{y}_{-r})$ using the sample harmonic mean of the likelihoods

$$\left[\frac{1}{B}\sum_{q=1}^{B}\frac{1}{f(y_r|\boldsymbol{\theta}^{(q)})}\right]^{-1}, \tag{2.16}$$

where $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)}$ are the B draws from the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$. These sample draws may come directly from the output of a standard MC or MCMC implementation.

Equation (2.15) is very straightforward and easy to compute which leads us to the computation of $m^{(1)}$ discussed in subsection 2.4.3. Also, this approach can be extended for $k > 1$, i.e., for $k = 2$, $k = 3$ and so on. For example, for $k = 2$ we can write:

$$f(y_{r,t}|\boldsymbol{y}_{-(r,t)}) = E\{[f(y_r|\boldsymbol{\theta})\, f(y_t|\boldsymbol{\theta})]^{-1}|\boldsymbol{y}\}^{-1}, \tag{2.17}$$

and this can be approximated as per equation (2.16):

$$\left[\frac{1}{B}\sum_{q=1}^{B}\left(\frac{1}{f(\boldsymbol{y}_r|\boldsymbol{\theta}^{q})}\cdot\frac{1}{f(\boldsymbol{y}_t|\boldsymbol{\theta}^{q})}\right)\right]^{-1}. \tag{2.18}$$

We know that with increasing leave-out size, the model comparison using CPO criterion (hence the pseudo Bayes factor) becomes closer to the model comparison from formal Bayes factor. Moreover, Raftery et al. (2006) argue that finding formal Bayes factor using this approach is hard due to Monte Carlo error. This implies that the CPO criterion with a larger leave-out size is more vulnerable to Monte Carlo error than with a smaller leave-out size.

2.5 Summary

There are many approaches to model comparison in both the frequentist and Bayesian contexts. The Bayes factor is a well-known tool for Bayesian model comparison. But, there are some criticisms for Bayes factor that arise due to prior specification and other issues. Different versions of Bayes factors have been suggested to address these criticisms. Predictive distribution based approaches such as pseudo

Bayes factor based on CPO criterion are also popular in Bayesian model comparison setting. From the cross-validation viewpoint, CPO criterion corresponds to "leave-out one" and the extension of CPO criterion ("leave-out all") leads to formal Bayesian model comparison.

Chapter 3

Bayesian Model Comparison for Linear Regression Models

Linear regression models are very common in studying the effect of one or more variables (explanatory variables) on a variable of interest (response variable). Different combinations of explanatory variables lead to different specifications of the linear regression models. These models can be compared using the model comparison techniques described in the previous chapter. Both the classical and Bayesian approaches of model comparison can be applied. For example, one can use likelihood ratio test which is a classical approach. Also, a Bayesian approach, say Bayes factor can be applied for model comparison after specifying the linear regression model in Bayesian context. In this chapter, we discuss the extended pseudo Bayes factor (EPBF) utilizing the extended CPO criterion, a compromise between the pseudo Bayes factor and the formal Bayes factor, as a model selection tool for linear regression models. Particularly our interest is to examine how the model selection behaves when we change the "leave-out size" in the extended CPO criterion. It is well-known that different model comparison tools, for example, likelihood ratio test, AIC, BIC, Mallow's CP may not select the same model among a set of candidate models. For this reason, we hope to observe whether there is any agreement or disagreement in between the pseudo Bayes factors and the formal Bayes factors as model comparison tool.

3.1 Bayesian Linear Regression Models

The Bayesian approach can be applied to estimate the parameters of the linear regression model. We start our discussion with defining a linear regression model. Suppose our response variable is $\boldsymbol{y}$ and we have a set of explanatory variables: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$. Here, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector of length n representing responses for n observations and each of the $\boldsymbol{x}_k$'s, $k = 1, \ldots, p$ are vectors of length n as well. Hence, the design matrix [of dimension $n \times (p+1)$] in this case is $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$. The mean value of the response for the i^{th} individual y_i can be described as,

$$E(Y_i \,|\, \boldsymbol{\beta}, \boldsymbol{X}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \;\; i = 1, \ldots, n, \tag{3.1}$$

where $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the explanatory values for the i^{th} individual and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\intercal$ are unknown regression parameters. Suppose we denote $\boldsymbol{x}_i^\intercal = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ as the i^{th} individual's row vector of explanatory variables. Then, the mean value in equation (3.1) can be re-expressed as

$$E(Y_i \,|\, \boldsymbol{\beta}, \boldsymbol{X}) = \boldsymbol{x}_i^\intercal \, \boldsymbol{\beta}.$$

In linear regression setting, one assumption made is that the responses $\{Y_i\}$ are independent conditional on the values of the parameters and the explanatory variables. Another assumption of equal variance is also made, that is, $\text{var}(Y_i \,|\, \boldsymbol{\beta}, \boldsymbol{X}) = \sigma^2$. Now, the vector of all unknown parameters in this linear regression setting becomes $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$. Also, we assume that the errors $\varepsilon_i = y_i - E(y_i \,|\, \boldsymbol{\beta}, \boldsymbol{X}), i = 1, \ldots, n$ are independent of one another. The errors ε_i's are distributed as normal with mean 0 and variance σ^2.

For a vector of n observations $\boldsymbol{y}$, we can write (in matrix notation):

$$\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \;\sim\; N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I), \tag{3.2}$$

with I as an $n \times n$ identity matrix. The vector of responses $\boldsymbol{y}$ has the multivariate normal distribution of dimension n with mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 I$.

From equation (3.2), the likelihood (joint density of $\boldsymbol{y}$) as a function of $\boldsymbol{\beta}$ and σ^2 is given by,

$$L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{X}) = L(\boldsymbol{\beta},\sigma^2;\boldsymbol{y},\boldsymbol{X}) \propto \prod_{i=1}^{n} \sigma^{-1} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta})^2\right\}. \tag{3.3}$$

The likelihood $L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{X})$ is used for the estimation of the parameters $\boldsymbol{\theta}$ in classical setting. In Bayesian context, we need to specify prior distribution, say $g(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ and then make inference on the $\boldsymbol{\theta}$ from the posterior distribution $p(\boldsymbol{\theta}\,|\,\boldsymbol{y},\boldsymbol{X})$ which has the following general form

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{y},\boldsymbol{X}) \propto L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{X}) \times g(\boldsymbol{\theta}). \tag{3.4}$$

Different prior specifications, for example, a reference prior or a conjugate prior lead to different posterior distributions. The posterior distribution of the parameters can be broken down into a marginal distribution of σ^2 and a conditional distribution of $\boldsymbol{\beta}$ given σ^2; this intuition is helpful to make inference for $\boldsymbol{\beta}$ and σ^2 respectively.

In the Bayesian linear regression setting, one might be interested in predicting a future observation $\tilde{y}$ corresponding to a vector of values of the explanatory variables say $\tilde{\boldsymbol{x}}$. From the equation (3.2) we can say that conditional on $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, $\tilde{y}$ is distributed as $N(\tilde{\boldsymbol{x}}'\boldsymbol{\theta}, \sigma^2)$. Then, averaging the conditional density of $\tilde{y}$, $p(\tilde{y}\,|\,\boldsymbol{\theta},\tilde{\boldsymbol{x}})$ over the posterior distribution of the parameters $\boldsymbol{\theta}$ we can obtain the posterior predictive density of $\tilde{y}$, $p(\tilde{y}\,|\,\boldsymbol{y})$ follows:

$$p(\tilde{y}\,|\,\boldsymbol{y},\tilde{\boldsymbol{x}},\boldsymbol{X}) = \int_{\boldsymbol{\theta}} p(\tilde{y}\,|\,\boldsymbol{\theta},\tilde{\boldsymbol{x}})\, p(\boldsymbol{\theta}\,|\,\boldsymbol{y},\boldsymbol{X})\, d\boldsymbol{\theta}. \tag{3.5}$$

Having the posterior, and the posterior predictive distribution, the next thing is to compute the extended CPO criterion discussed in the subsection 2.4.3 as a model selection tool. The posterior predictive distribution formulated in equation (3.5) can be used to obtain the cross-validation densities, and hence to compute the CPO criterion.

Two types of unknown parameters to be estimated are included in $\boldsymbol{\theta}$: parameters for the location (i.e., $\boldsymbol{\beta}$ to calculate mean of the responses) and parameter σ^2 to

calculate the variance of the responses. For simplicity, we initially assume that the variance parameter σ^2 is known. Now, $\boldsymbol{\theta}$ reduces to $\boldsymbol{\beta}$. We refer to this situation as "Normal model with unknown mean and known variance". At first, we examine the behavior of model comparisons of such models using the extended CPO criterion with different "leave-out sizes" which is discussed in section 3.2. We discuss this as a building block to understand the general situation with both the $\boldsymbol{\beta}$ and σ^2 unknown; this situation can be termed as "Normal model with unknown mean and variance." Behavior of model comparisons using the extended CPO criterion with both the $\boldsymbol{\beta}$ and σ^2 unknown is discussed later in section 3.3.

3.2 Normal Model with Unknown Mean and Known Variance

In this section, we discuss the extended CPO criterion applied to linear regression models with known variance of the responses. The likelihood of the responses given in equation (3.3) can be re-expressed in matrix notation as

$$L(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X},\sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\intercal\boldsymbol{y} - 2\boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y} + \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta}\right)\right\}. \tag{3.6}$$

Now, a prior distribution for $\boldsymbol{\beta}$ is needed to commence the Bayesian inference through constructing the posterior distribution of $\boldsymbol{\beta}$. The distribution of $\boldsymbol{y}$ is multivariate normal, and from the equation (3.6), we see that the $\boldsymbol{\beta}$ plays the same role in the exponent looks like $\boldsymbol{y}$; this gives an intuition that the multivariate normal prior distribution for $\boldsymbol{\beta}$ is conjugate. Hence, let us consider the prior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$ as multivariate normal (of dimension $p+1$) with mean vector $\boldsymbol{\beta}_0$ and variance-covariance matrix Σ_0. The posterior distribution of $\boldsymbol{\beta}$ has the following expression:

$$\begin{aligned} p(\boldsymbol{\beta}\,|\,\boldsymbol{y},\boldsymbol{X},\sigma^2) \propto & L(\boldsymbol{\beta};\boldsymbol{y},\boldsymbol{X},\sigma^2) \times p(\boldsymbol{\beta}) \\ \propto & \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\intercal\boldsymbol{y} - 2\boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{y} + \boldsymbol{\beta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\beta}\right)\right\} \\ & \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^\intercal\Sigma_0^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0^\intercal\Sigma_0^{-1}\boldsymbol{\beta}_0\right)\right\}. \end{aligned} \tag{3.7}$$

After simplification, we can write

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2) \propto \exp\left\{ -\frac{1}{2}\left[\boldsymbol{\beta}^\intercal \left(\Sigma_0^{-1} + \frac{\boldsymbol{X}^\intercal \boldsymbol{X}}{\sigma^2} \right) \boldsymbol{\beta} - 2\boldsymbol{\beta}^\intercal \left(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{\boldsymbol{X}^\intercal \boldsymbol{y}}{\sigma^2} \right) \right] \right\} \tag{3.8}$$

$$= \exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^\intercal \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \right\},$$

which is proportional to a multivariate normal density, with mean vector

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathrm{E}[\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = \left(\Sigma_0^{-1} + \frac{\boldsymbol{X}^\intercal \boldsymbol{X}}{\sigma^2} \right)^{-1} \left(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{\boldsymbol{X}^\intercal \boldsymbol{y}}{\sigma^2} \right), \tag{3.9}$$

and variance-covariance matrix

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \mathrm{Var}[\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = \left(\Sigma_0^{-1} + \frac{\boldsymbol{X}^\intercal \boldsymbol{X}}{\sigma^2} \right)^{-1}. \tag{3.10}$$

From the formula of posterior mean $\mathrm{E}[\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2]$ in equation (3.9), it is evident that if the prior variance-covariance matrix Σ_0 has elements with large magnitude (that is the precision matrix Σ_0^{-1} has elements with small magnitude), then the posterior mean approximately equals the least square estimate of $\boldsymbol{\beta}$: $(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{y}$. Alternatively, if the variance of responses σ^2 is very large, then the expectation is approximately equals $\boldsymbol{\beta}_0$, the prior mean.

Since the posterior distribution of $\boldsymbol{\beta}$ is multivariate normal, and the sampling distributions of the y_i's are also normal, in this setting the predictive posterior distribution of a future observation, say $\tilde{y}$, will be normal as well. So, to compute the extended CPO criterion, we can directly use the closed form densities of the predictive posterior distribution as the cross-validation densities. Alternatively, if we pretend that the predictive posterior formulation has no closed form solution, one can approximate the cross-validation densities using the Monte Carlo (MC) samples from the posterior distribution of $\boldsymbol{\beta}$. After that, the extended CPO criterion can be computed as discussed in subsection 2.4.4. In general, having closed form results, there is no need to use the MC based results. But, here our purpose is to examine how well the MC based results approximate the closed form results since in more difficult problems we must rely on such MC results. Hence, we use both

the approaches discussed in next two subsections to compute cross-validation densities and hence the extended CPO criterion or the extended pseudo Bayes factor (PBF) as a model comparison tool.

3.2.1 Extended CPO Criterion using Closed Form Posterior Predictive Distributions

The posterior predictive distribution of a new observation $\tilde{y}$, $p(\tilde{y}\,|\,\boldsymbol{y})$ can be readily obtained using the equation (3.5). As discussed in section 3.1, conditional on $\boldsymbol{\beta}$, $\tilde{y}$ is distributed as normal with mean $\tilde{\boldsymbol{x}}'\boldsymbol{\beta}$ and variance σ^2. Also, the posterior distribution of $\boldsymbol{\beta}$ is multivariate normal with mean vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and variance-covariance matrix $\boldsymbol{V}_{\boldsymbol{\beta}}$. Now, $p(\tilde{y}\,|\,\boldsymbol{y})$ has the following expression:

$$p(\tilde{y}\,|\,\boldsymbol{y}) = \int_{\boldsymbol{\beta}} \phi(\tilde{\boldsymbol{y}};\tilde{\boldsymbol{x}}'\boldsymbol{\beta},\sigma^2)\,\phi_{p+1}(\boldsymbol{\beta};\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{V}_{\boldsymbol{\beta}})\,d\boldsymbol{\beta}, \tag{3.11}$$

where $\phi(\tilde{\boldsymbol{y}};\tilde{\boldsymbol{x}}'\boldsymbol{\beta},\sigma^2)$ and $\phi_{p+1}(\boldsymbol{\beta};\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{V}_{\boldsymbol{\beta}})$ are the corresponding normal densities for $N(\tilde{\boldsymbol{x}}'\boldsymbol{\beta},\sigma^2)$ and $N_{p+1}(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{V}_{\boldsymbol{\beta}})$. Simplifying the results, we can show that $\tilde{y}\,|\,\boldsymbol{y}$ is distributed as normal with mean $\tilde{\boldsymbol{x}}'\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and variance $\tilde{\boldsymbol{x}}'\boldsymbol{V}_{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}+\sigma^2$. Such univariate results can be extended to multivariate ones when we want to find the predictive distribution of a vector of new observations.

A cross-validation density with "leave-out one", say $f(y_r\,|\,\boldsymbol{y}_{-r})$, $r=1,\ldots,n$ with $\boldsymbol{y}_{-r}=\{y_1,\ldots,y_{r-1},y_{r+1},\ldots,y_n\}$ can be computed from the posterior predictive formulation (3.11) by setting $\tilde{y}=y_r$ and $\boldsymbol{y}=\boldsymbol{y}_{-r}$; this is the CPO criterion. Then, as described in subsection 2.4.3, the log-likelihood of these cross-validation densities (or CPO's), denoted by $m^{(1)}$, can be computed for two models. The difference of $m^{(1)}$ in two competing models multiplied by the number of observations n is known as the log pseudo Bayes factor. Using the value of log pseudo Bayes factor (PBF), one can decide on one model over the other.

Similarly, cross-validation densities with "leave-out size" greater than one, can be computed by substituting $\tilde{y}$ by the leave-out elements in the posterior predictive formulation (3.11). Accordingly, we can compute the log extended pseudo Bayes factor based on this extended CPO criterion and then compare the competing two models. The procedure is already discussed in subsection 2.4.3.

3.2.2 Extended CPO Criterion using Monte Carlo Samples

In general, how the MC samples from a posterior distribution can be used directly to compute the CPO criterion is discussed in subsection 2.4.4. Suppose we have B draws of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(B)}$ from the posterior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2)$ specified by equation (3.8). The CPO criterion or cross-validation density with "leave-out one" can be approximated using (2.16). Log PBFs can be computed using the approximated cross-validation densities to compare two linear regression models in Bayesian context. Cross-validation densities in extended CPO criterion with "leave-out two" can be approximated by equation (2.18). Similarly, cross-validation densities for different "leave-out sizes" can be approximated; log EPBFs computed using these approximated cross-validation densities can then be used as to compare linear regression models.

We discuss the comparative behavior of the extended CPO criterion or the EPBF in the next subsection for "Normal model with unknown mean and known variance" setting with different "leave-out sizes". Both the cases are considered: when the extended CPO criterion is computed as a closed form solution using the predictive posterior distribution, and when the extended CPO criterion is approximated using the MC samples from the posterior distribution of the linear regression parameters.

3.2.3 Comparative Behavior of Extended CPO Criterion: Closed Form Versus Monte Carlo Samples

The unknown regression parameters $\boldsymbol{\beta}$ correspond to the mean of the responses whereas the variance of the responses σ^2 is known in "Normal model with unknown mean and known variance" which is already discussed in section (3.2). Here, we examine how the extended CPO criterion, a model comparison tool behaves in real life scenario for both the situations: closed form and MC based solution with an illustrating example. We use a data file here named `automobile` that is taken from the UCI Machine Learning Repository (Lichman, 2013). This data is originally extracted from the `1985 Ward's Automotive Yearbook`.
Among the three types of entities contained in the original data file, we consider only the specification of an automobile in terms of various characteristics, say

length of the automobiles, height of the automobiles, etc. Particularly, we consider the following five explanatory variables regarding the characteristics of the automobiles:

- *length*,
- *width*,
- *height*,
- *compression ratio*, and
- *horsepower*

with $\log(price)$ as our response variable. We want to examine whether the $\log(price)$ of the automobiles depend on the listed characteristics of the automobiles. There are in total $n = 195$ non-missing observations for this `automobile` data. We specify the models and priors for the regression parameters $\boldsymbol{\beta}$ below.

Model Specification

Three different models are considered; these models denoted by Model 1, Model 2, and Model 3 have different combinations of the explanatory variables. These three models have the following specifications.

1. Model 1: Full Model (Model with all explanatory variables listed above) with the formulation

$$\begin{aligned} lprice_i =& \beta_0 + \beta_1\, length_i + \beta_2\, width_i + \beta_3\, height_i + \beta_4\, comp.ratio_i \\ &+ \beta_5\, horsepower_i + \varepsilon_i, \end{aligned}$$

   where $i = 1, 2, \ldots, 195$, $lprice$ denotes $\log(price)$ and $comp.ratio$ indicates *compression ratio*.

2. Model 2: Model which leaves only *height* out of the full Model.

3. Model 3: Model which leaves *height* and *compression ratio* out of the full Model.

We consider 2500 MC samples from the posterior distribution of the regression parameters $\boldsymbol{\beta}$ (see equation (3.8)) for the MC based calculation of the extended CPO criterion. We divide these into 10 batches of equal size 250. We consider the log extended CPO and log extended pseudo Bayes factor calculation for each of these batches. Also, we compute the closed form results for CPO. To assess how well the MC based calculation approximates the exact answer we examine the deviation of the results found in 10 separate MC results from the closed form result. All these are done for different "leave-out sizes" (from "leave-out 1" to "leave-out all"). We consider 15 different "leave-out sizes"; these are 1, 5, 10, 30, 50, 80, 100, 120, 150, 170, 175, 180, 185, 190, and 195 (i.e., all). These sizes are taken arbitrarily with more sizes near the "leave-out all". In addition, while taking the combinations of MC samples required for extended CPO calculation with different "leave-out sizes", we take all combinations if the total combinations are less than 2000 and take only 2000 combinations randomly if the total combinations exceed 2000 (as the number of combinations increases drastically/exponentially with "leave-out size").

Prior Specification

For the regression parameters $\boldsymbol{\beta}$, we consider multivariate normal prior with mean vector $\boldsymbol{0}$ and variance-covariance matrix identity matrix $\boldsymbol{I}$. We run a simple linear regression with the response and explanatory variables discussed above and observe that the linear regression coefficients are within ± 2 which makes the considered mean vector for prior reasonable. If the regression coefficients are bigger than ± 2, then this prior mean specification might not be useful.

We use the model and prior setup discussed above to obtain the results described in the next subsection. We describe our findings for the three models considered. We have three possible pairwise comparisons of these three models: Model 1 versus Model 2, Model 1 versus Model 3, and Model 2 versus model 3. At first, we describe the behavior of the extended CPO criterion for the three models separately and then extend to three model comparisons using the EPBF which relies on the extended CPO criterion.

3.2.4 Results

First we visualize the closed form results of log extended CPO values obtained from Model 1, Model 2 and Model 3 (see Figure 3.1). The purpose of this visualization is to display the log extended CPO values for different "leave-out sizes" (smaller to larger) for all three models at the same time. We can observe the pattern of the log extended CPO values over the increasing "leave-out sizes" from the Figure 3.1.

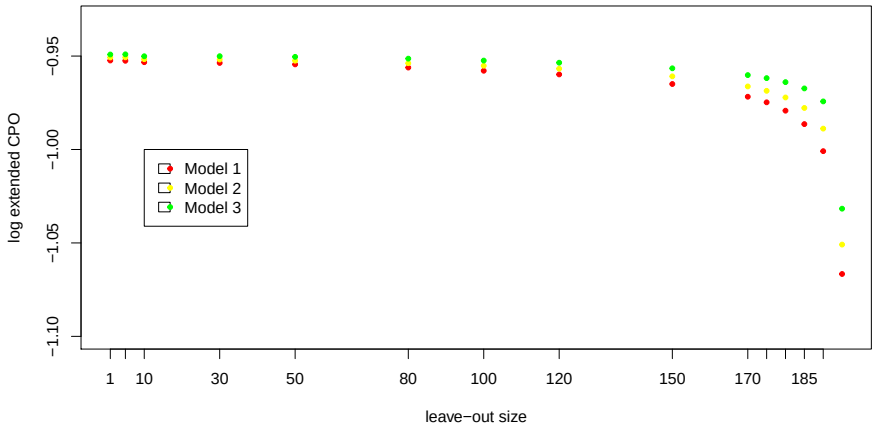

Figure 3.1: Comparison of closed form results for three models (unknown mean but known variance)

All three models show quite a similar pattern with decreasing log extended CPO values over the increasing leave-out size. We observe a smaller decrease in log extended CPO values up to "leave-out 170" but a sharp decrease in an exponential manner after "leave-out 170". All three models have close log extended CPO values, but the distance between the log extended CPO values increases with increasing "leave-out sizes", and the difference is clearly visible at "leave-out all". The relative position of the three models remains the same at each of the "leave-out sizes". This indicates the consistent pattern of these three models over different "leave-out sizes". The similarity of the closed form log extended CPO values for three models can be concluded from Figure 3.1 except the "leave-out all" point.

Such close models are interesting to examine the model comparison behavior over different "leave-out sizes". In closed form result, we have consistent pattern for these three models at all "leave-out sizes"; but this may not be the case when we must compute using MC samples. The bias related to the results obtained from the MC samples increases with the increasing leave-out points. This bias is an important issue to investigate further. We need to compare the closed form results and results based on MC samples to examine whether MC based results can replicate the closed form results or there is a bias that distort the MC based results from the closed form results. This will give us an intuition for the future when we have only MC based results due to unavailability of the closed form results. Now, for the Model 1 we observe the log extended CPO values from the closed form solutions as well as from the MC samples with the setup described above in the Figure 3.2. The green filled circles represent the closed form values at each "leave-out size".

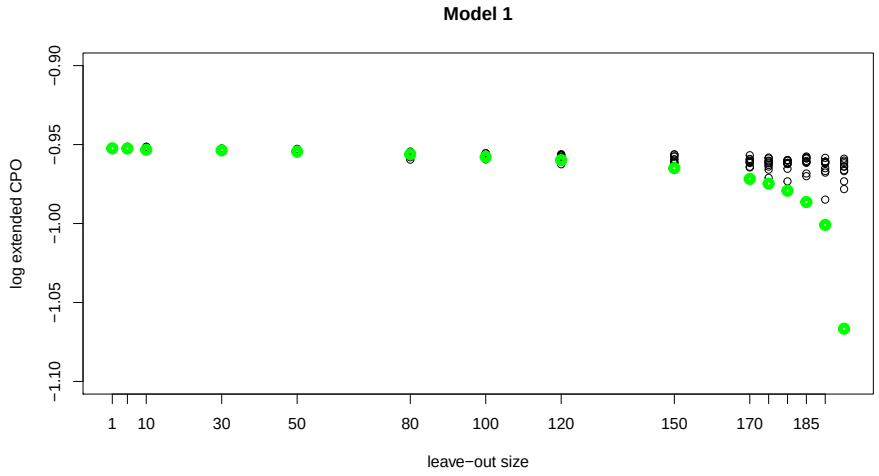

Figure 3.2: Comparison of closed form results and MC based results for Model 1 (unknown mean but known variance)

The MC based results (for 10 batches here) match the corresponding closed form result up to "leave-out 120" (see Figure 3.2). The visible changes between the results from MC samples and closed forms are observed at "leave-out 150" and higher "leave-out sizes". Compared to nearer "leave-out sizes", at "leave-out all" we see a sharp decrease in the log extended CPO value for the closed form result

which is not the case for MC samples. Clearly, MC based results show an upward bias from the closed form result. This is not only the case for Model 1 as we observe the same pattern for the other two models considered here. The closed form result demonstrates the true values, and we observe substantive positive departure of MC based results from these values at higher "leave-out sizes".

Now we compare the three models pairwise. First compare Model 1 with Model 2. For model comparison, we use the extended pseudo Bayes factor which can be computed from the calculated extended CPO for the competing models. Figure 3.3 describes this comparison visually. The red filled circles denote the closed form results whereas the 10 black circles represent MC results from 10 batches at each "leave-out size". We use the same specification for Figure 3.4 and Figure 3.5.

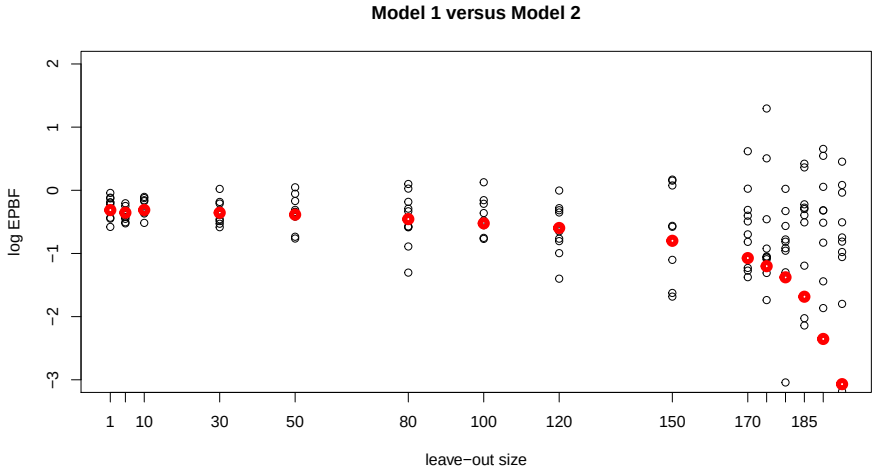

Figure 3.3: Comparison of Model 1 and Model 2 (closed form results and MC based results for "unknown mean and known variance" situation)

We observe some interesting results from Figure 3.3. As we have negative log EPBFs at all "leave-out sizes", the closed form log EPBFs suggest the choice of Model 2 over Model 1 (hence the same choice). The strength of evidence increases with increasing "leave-out sizes". MC based results vary over the direction with increasing "leave-out sizes". All MC based results from 10 batches yield the same choice of model (Model 2 over Model 1) for "leave-out sizes" less than 80 for this particular data. But, from "leave-out size" 80, the choices of the model fluctuate

for the MC based results obtained from 10 batches. This fluctuation is due to the MC error. We need to quantify the MC error to examine how this increases for increasing 'leave-out sizes".

For the smaller "leave-out sizes" (for example, 1, 5, 10) the closed form results are near the center of the MC based results. For the larger "leave-out sizes" say "leave-out 100" or more, the MC based results from most of the 10 batches have a tend to have higher values than the corresponding closed form results, Hence from Figure 3.3, it is clearly observed that the MC based results become positively biased and more variable as the "leave-out size" increases. We have the interest in whether it is possible to find a "leave-out size" where the EPBF is close to the formal Bayesian comparison with smaller MC error which is discussed later.

Moreover, at "leave-out all", the EPBF becomes a version of the formal Bayesian comparison, that is, the Bayes factor. In Figure 3.3, the closed form result at "leave-out all" indicates the value of the formal Bayes factor. MC based results from all 10 batches show positive bias from the formal Bayes factor value which implies that we are very poorly computing real Bayesian model comparison with some positive MC errors for MC based results.

Now we check the two other possible model comparisons: Model 1 versus Model 3 and Model 2 versus Model 3. Closed form and MC based log EPBFs for Model 1 versus Model 3 are displayed in Figure 3.4.

Figure 3.4 displays the similar pattern as Figure 3.3. Here we compare Model 1 with Model 3. As the comparison between Model 1 and Model 2, the closed form results once again indicate the same choice of model over different "leave-out sizes", and the evidence is stronger with increasing "leave-out sizes". Interestingly, the change in log EPBFs between two different "leave-out sizes" is steeper for comparison of Model 1 versus Model 3 than the comparison of Model 1 versus Model 2. Based on the closed form results, Figure 3.4 suggests the choice of Model 3 over Model 1.

As with the comparison between Model 1 and Model 2, the closed form results are near the center of the MC based results from 10 batches for smaller "leave-out sizes" for the comparison between Model 1 and Model 3. Figure 3.4 shows that the MC based results become positively biased and more variable as the "leave-out size" increases which is similar to Figure 3.3. The only distinction with the pre-

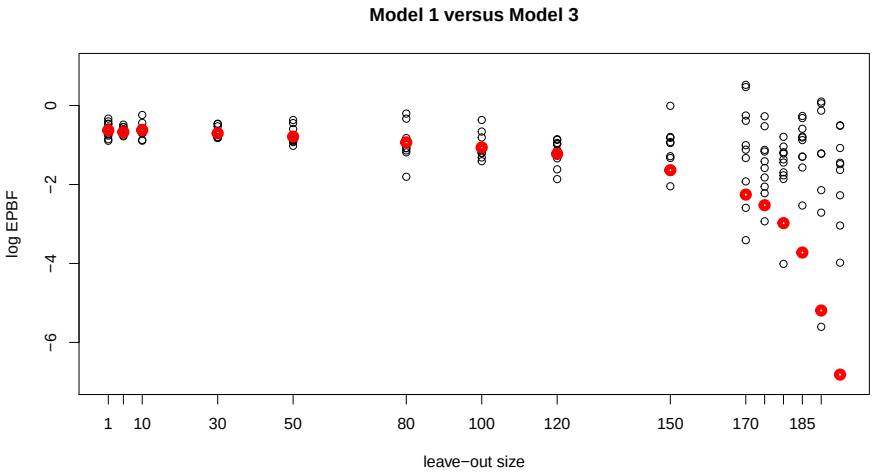

Figure 3.4: Comparison of Model 1 and Model 3 (closed form results and MC based results for "unknown mean and known variance" situation)

vious comparison is that from "leave-out 120", the choices of the model fluctuate for the MC based results obtained from 10 batches, whereas this fluctuation starts from the "leave-out 80" in the comparison between Model 1 and Model 2.

Now, we move to the last comparison: Model 2 versus Model 3. The closed form and MC based results for this comparison are displayed in Figure 3.5. Like the previous two comparisons, the closed form results in this comparison indicate the same choice of the model over different "leave-out sizes". Similarly, the strength of evidence increases with increasing "leave-out sizes". The closed form results displayed in Figure 3.5 suggest the choice of Model 3 over Model 2 as we have negative log EPBF values at all "leave-out sizes" for this comparison. All MC based results from 10 batches yield the same choice of model (Model 3 over Model 2) for "leave-out sizes" less than 100, and start to fluctuate from "leave-out size" 100 and higher. The same behavior of the MC based results at "leave-out sizes" is observed in Figure 3.5 as in Figures 3.3 and 3.4. Once again, MC based results become positively biased and more variable as the "leave-out size" increases might be due to MC error.

In summary, all three model comparisons show us the same behavior of the EPBF (computed from extended CPO criterion) as a model selection tool. The

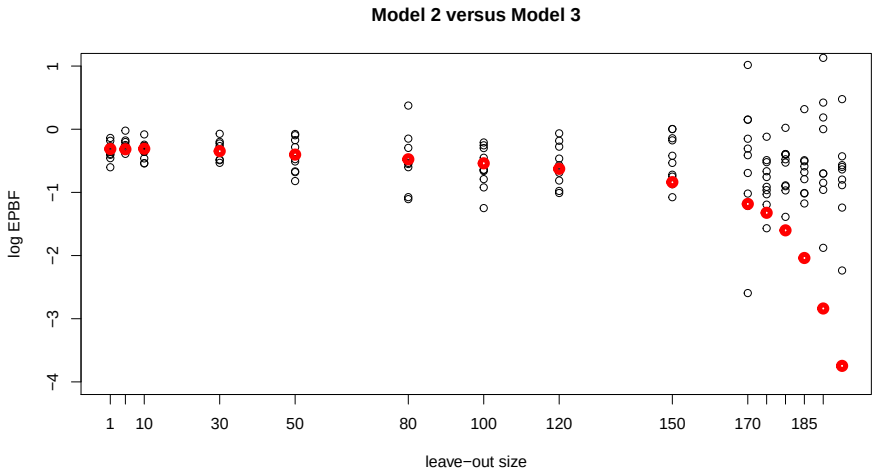

Figure 3.5: Comparison of Model 2 and Model 3 (closed form results and MC based results for "unknown mean and known variance" situation)

close form results choose the model with a small number of parameters in all three comparisons. Also, for closed form solutions, with larger "leave-out sizes", the EPBFs show larger difference among the competing models. MC based results deviate from the closed form results in all three comparisons in an increasing pattern with the increasing "leave-out sizes". The deviation is due to the MC error, and we try to compute the contribution of MC error at different "leave-out sizes" which we discuss in the following subsection.

3.2.5 Summarizing the Computations

From the discussion in the previous subsection, we came to know the variation of the MC based results at different "leave-out sizes". We compute the root mean squared error (RMSE) of the MC based estimates of the log EPBFs at all considered "leave-out sizes" to get a measurement of the variation from the closed form results. Note that, RMSE is a way of measuring how good the MC based estimates of the log EPBFs compared to the closed form log EPBF. The smaller the RMSE, the better way the MC based results are behaving in general. The RMSE of the estimated log EPBFs for all three model comparisons are tabulated in Table 3.1.

Table 3.1: Root mean squared errors of the estimated log EPBFs for all three model comparisons

Leave-out size	Model 1 vs. Model 2	Model 1 vs. Model 3	Model 2 vs. Model 3
1	0.178	0.166	0.153
5	0.190	0.136	0.142
10	0.103	0.143	0.092
30	0.161	0.202	0.121
50	0.203	0.279	0.137
80	0.303	0.351	0.166
100	0.385	0.755	0.558
120	0.396	0.626	0.710
150	0.800	0.862	1.099
170	1.130	1.291	1.126
175	0.879	1.391	0.711
180	1.088	1.689	0.915
185	1.545	2.527	1.615
190	1.374	4.369	3.474
all	2.274	6.292	4.182

Now, we can interpret the RMSE values from Table 3.1. For example, for comparison of Model 1 with Model 2 at "leave-out 100" the RMSE of the estimated log EPBFs is 0.385 which implies that the model selections while using MC based EPBFs instead of the corresponding closed form are erroneous by a factor of $\exp(0.385) = 1.47$ (which implies $(\exp(0.385) - 1) \times 100 = 47$ percent erroneous model comparisons by the MC samples). Similarly, while comparing Model 1 with Model 3, at "leave-out 120" the RMSE of the estimated log EPBFs is 0.626 which leads to 87 percent of the erroneous decision on model comparisons using MC based EPBFs relative to the corresponding closed form. We can interpret all other RMSE values in a similar fashion.

For all three comparisons, the RMSE values of the estimated log EPBFs from MC samples have the same pattern; the RMSE values increase with the increasing "leave-out size" and these increase drastically at higher "leave-out sizes". We can use several cut-off values for the RMSE values to examine the level of error in model comparison decision while using the MC based EPBFs relative to closed form counterparts at different "leave-out sizes". The RMSE values from Table 3.1

are in logarithmic scale, and we consider three cut-off values for these values: log 1.25, log 1.5, and log 2 which correspond to 25, 50 and 100 percent erroneous model selection when using the log EPBFs estimated from MC samples instead of the closed form values. The comparative RMSE values for the MC based EPBFs at the considered "leave-out sizes" with vertical lines through the cut-off values are plotted in Figure 3.6.

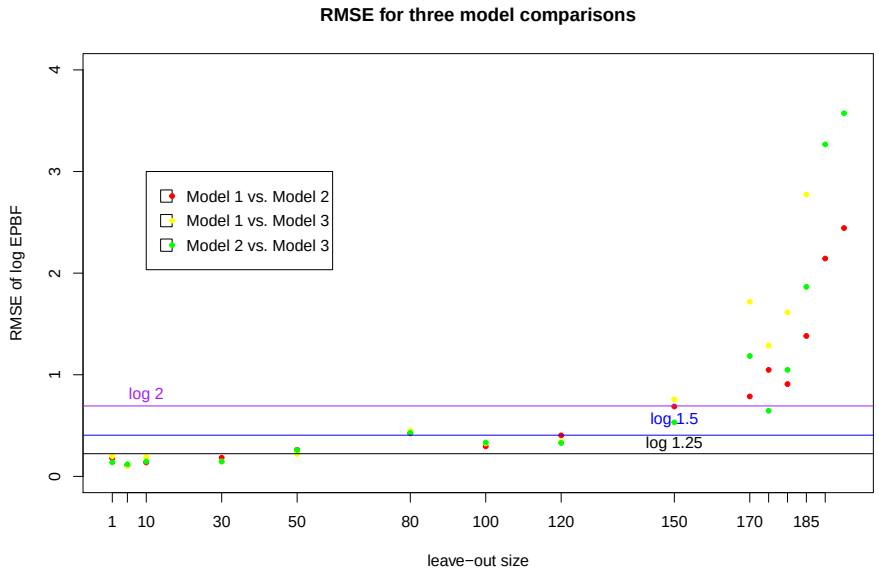

Figure 3.6: Root mean squared error of the log EPBFs from 2500 MC samples with three cut-offs (unknown mean but known variance)

If we consider the log 1.25 as the cut-off value, then from Figure 3.6 we observe that up to "leave-out 30" the RMSE values lie below the cut-off value for all three comparisons. The RMSE values are on the both sides of the cut-off value between "leave-out 50" and "leave-out 80" whereas all the RMSE values exceed the cut-off value at any "leave-out size" greater than or equal to 100. Similarly, up to "leave-out 80" and "leave-out 120", the RMSE values for all comparisons are less than or equal to the cut-off values log 1.5 and log 2 respectively. Also, the RMSE values for all comparisons are greater than the cut-off values log 1.5 and log 2 at "leave-out size" 150 and higher.

To reduce the errors in model selection via model comparisons using MC samples we need to increase either the number of batches or the MC samples as per the MC rule. For our analysis, we have 2500 MC samples and 10 equal sized batches with 250 MC samples. As we are computing RMSE values in logarithmic scale, if we want error reduction by a factor k compared to the current level x, then we need to increase the current MC sample size by a factor of $(\ln(x \times k)/\ln(k))^2$. For example, according to MC rule, at least $2500 \times (\ln(6)/\ln(2))^2 = 16705$ MC samples are needed if we want error reduction in model selection by a factor of 3 from the MC based results at cut-off point $\log(2)$. In other words, the model comparison results within RMSE value $= \log 2$ (or within 100% relative error) for 2500 MC samples, will be within RMSE value $= \log 4/3$ (or within 33% relative error) for 16705 MC samples. Then, the relative error reduces to one third with the increased MC samples. So, more MC samples will provide a correct model comparisons based on the MC samples.

3.3 Normal Model with Unknown Mean and Variance

We discuss the extended CPO criterion applied to linear regression models with unknown mean and variance of the responses in this section. Compared to models discussed in the previous section, we have an additional unknown parameter for the variance of the responses. Accommodating this, now the likelihood of the responses (3.6) have the following expression:

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}, \boldsymbol{X}) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\beta}\right)\right\}. \tag{3.12}$$

Having specified the likelihood function, the next step is to specify the prior distribution for $\boldsymbol{\beta}$ and σ^2. We have already learned from the section 3.2 that if σ^2 is known, the multivariate normal prior distribution for $\boldsymbol{\beta}$ is conjugate. If $\sigma^2 \sim$ inverse-gamma (a, b), that is

$$p(\sigma^2) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{b}{\sigma^2}\right), \ \sigma^2 > 0,$$

then we can write the posterior distribution of σ^2 with $\boldsymbol{\beta}$ known as

$$\begin{aligned}
p(\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) &\propto p(\sigma^2)\, L(\sigma^2; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta},) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{b}{\sigma^2}\right) \sigma^{-n} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}\right)\right\} \\
&= \left(\frac{1}{\sigma^2}\right)^{a+\frac{n}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}\left(\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}\right)\right]\right\},
\end{aligned}$$

which is simply an inverse-gamma density, so that the conjugate prior for σ^2 is:

$$\{\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}\} \sim IG\left(a - \frac{n}{2}, \left[b + \frac{1}{2}\left(\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}\right)\right]\right).$$

Now, we can factorize the joint conjugate prior of $\boldsymbol{\beta}$ and σ^2 as

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta} \,|\, \sigma^2)\, p(\sigma^2) = N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{V}_{\boldsymbol{\beta}}) \times IG(a,b) = NIG(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}, a, b) \tag{3.13}$$

$$\begin{aligned}
&= \frac{b^a}{(2\pi)^{(p+1)/2} |\boldsymbol{V}_{\boldsymbol{\beta}}|^{1/2} \Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+1+\frac{(p+1)}{2}} \\
&\quad \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\}\right] \qquad (3.14) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{(p+1)}{2}+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\}\right],
\end{aligned}$$

where $a, b > 0$, and $\Gamma(\cdot)$ denotes the Gamma function. This prior is called the normal-inverse-gamma prior and can be denoted as $NIG(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}, a, b)$ (Banerjee, 2008).

With the likelihood (3.12) and joint prior distribution (3.13), the joint posterior

distribution of $(\boldsymbol{\beta}, \sigma^2)$ has the following expression

$$p(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta}\right)\right\} \tag{3.15}$$
$$\times \left(\frac{1}{\sigma^2}\right)^{a+\frac{(p+1)}{2}+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\}\right].$$

As suggested by Banerjee (2008), to derive the joint posterior distribution $p(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X})$, we use the *multivariate completion of squares* identity with a symmetric positive definite matrix D:

$$\boldsymbol{u}^\intercal D \boldsymbol{u} - 2\boldsymbol{\alpha}^\intercal \boldsymbol{u} = (\boldsymbol{u} - D^{-1}\boldsymbol{\alpha})^\intercal D (\boldsymbol{u} - D^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^\intercal D^{-1} \boldsymbol{\alpha}. \tag{3.16}$$

An application of the identity (3.16) gives,

$$\begin{aligned}
&\frac{1}{\sigma^2}\left[b + \frac{1}{2}\left\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) + (\boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\beta}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{\beta})\right\}\right] \\
=&\frac{1}{\sigma^2}\left[b^* + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)^\intercal \boldsymbol{V}^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)\right].
\end{aligned}$$

Using this, $p(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X})$ in (3.15) can be re-written as

$$p(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto \left(\frac{1}{\sigma^2}\right)^{a^*+\frac{(p+1)}{2}+1} \times \exp\left[-\frac{1}{\sigma^2}\left\{b^* + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)^\intercal \boldsymbol{V}^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*)\right\}\right], \tag{3.17}$$

which can be identified as a $NIG(\boldsymbol{\mu}^*, \boldsymbol{V}^{*-1}, a^*, b^*)$ with

$$\begin{aligned}
\boldsymbol{\mu}^* &= (\boldsymbol{V}_\beta^{*-1} + \boldsymbol{X}^\intercal \boldsymbol{X})^{-1}(\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \boldsymbol{X}^\intercal \boldsymbol{y}) \\
\boldsymbol{V}^* &= (\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} + \boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \\
a^* &= a + n/2 \\
b^* &= b + \frac{1}{2}\left[\boldsymbol{\mu}_{\boldsymbol{\beta}}^\intercal \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{\mu}^{*\intercal} \boldsymbol{V}^{*-1}\boldsymbol{\mu}^*\right].
\end{aligned}$$

Having the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 as *NIG*, and the sampling distributions of the y_i's as normal, the predictive posterior distribution of any future observations can be obtained as a closed form solution after some calculations.

Hence, as the section 3.2, to compute the extended CPO criterion, we can directly use the closed form densities of the predictive posterior distribution as the cross-validation densities. Similarly, pretending the predictive posterior formulation has no closed form solution, one can approximate the cross-validation densities using the Monte Carlo (MC) samples from the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 and then extended CPO criterion can be computed as discussed in the previous section. Once again, we use both the closed form and MC samples based approaches, and these are discussed in next two subsequent subsections to compute the extended pseudo Bayes factor (PBF) as a model comparison tool.

3.3.1 Extended CPO Criterion using Closed Form Posterior Predictive Distributions: Unknown Variance Situation

Suppose we want to predict the outcome $\tilde{\boldsymbol{y}}$ for a future $t \times (p+1)$ matrix of regressors $\tilde{\boldsymbol{X}}$. These $\tilde{\boldsymbol{y}}$ are independent of $\boldsymbol{y}$, and given $\boldsymbol{\beta}$ and σ^2 known, we can write the sampling distribution of $\tilde{\boldsymbol{y}}$: $\tilde{\boldsymbol{y}} \sim N(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2 I_t)$. Now, the predictive posterior distribution of $\tilde{\boldsymbol{y}}$, that is $p(\tilde{\boldsymbol{y}}, |\, \boldsymbol{y}, \tilde{\boldsymbol{X}})$ can be obtained using the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 as

$$
\begin{aligned}
p(\tilde{\boldsymbol{y}} \,|\, \boldsymbol{y}, \tilde{\boldsymbol{X}}) &= \int_{\boldsymbol{\beta}, \sigma^2} \phi_t(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2 I_t), \times \phi^{NIG}_{(p+1)}(\boldsymbol{\beta}; \boldsymbol{\mu}^*, \boldsymbol{V}^{*-1}, a^*, b^*)\, d\boldsymbol{\beta}\, d\sigma^2 \quad (3.18) \\
&= MVSt_{2a^*}\left(\tilde{\boldsymbol{X}}\,\boldsymbol{\mu}^*, \frac{b^*}{a^*}(I + \tilde{\boldsymbol{X}}\boldsymbol{V}^*\tilde{\boldsymbol{X}}^{\intercal})\right),
\end{aligned}
$$

where $\phi_t(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2 I_t)$ and $\phi^{NIG}_{(p+1)}(\boldsymbol{\beta}; \boldsymbol{\mu}^*, \boldsymbol{V}^{*-1}, a^*, b^*)$ are the corresponding normal and normal-inverse-gamma densities for $N(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2 I_t)$ and $NIG(\boldsymbol{\mu}^*, \boldsymbol{V}^{*-1}, a^*, b^*)$ respectively. The final expression follows from the marginal distribution of $\boldsymbol{y}$ with NIG prior distribution for $(\boldsymbol{\beta}, \sigma^2)$ which utilizes the well-known *Sherman-Woodbury-Morrison* identity and some matrix identities.

As discussed for the "Normal model with unknown mean and known variance" situation in section 3.2, the cross-validation density with "leave-out one", say $f(y_r|\boldsymbol{y}_{-r})$, $r = 1, \ldots, n$ can be computed directly from the posterior predictive formulation (3.18) by setting $\tilde{\boldsymbol{y}} = y_r$ and $\boldsymbol{y} = \boldsymbol{y}_{-r}$. Then, the log PBF can be computed for a comparison of two available models using the CPO criterion. Similarly, cross-validation densities with "leave-out size" greater than one, can be computed

directly. Hence, we can compute the log EPBF based on this extended CPO criterion and then compare the competing model pairs as described in section 3.2.

3.3.2 Extended CPO Criterion using Monte Carlo Samples: Unknown Variance Situation

The purpose of calculating the extended CPO criterion using MC samples for the "Normal model with unknown mean and known variance" situation is already discussed in section 3.2. However, instead of taking MC samples of $\boldsymbol{\beta}$ and σ^2 from their joint posterior distribution, it is preferable to factorize the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 into a marginal posterior of σ^2 and a conditional posterior of $\boldsymbol{\beta}$ (given σ^2) that is $p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) \times p(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$. Then, we can take samples of $\sigma^2|\boldsymbol{y}, \boldsymbol{X}$ and $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}$ from $p(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$ and $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X})$ respectively. Since, both the prior and posterior belongs to the same family of distribution, using equation (3.13), it can be shown that $\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X} \sim N(\boldsymbol{\mu}^*, \sigma^2 \boldsymbol{V}^{*-1})$ and $\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim IG(a^*, b^*)$. Thus B MC samples for $(\boldsymbol{\beta}, \sigma^2)$: $(\boldsymbol{\beta}, \sigma^2)^{(1)}$, $(\boldsymbol{\beta}, \sigma^2)^{(2)}$, $\ldots, (\boldsymbol{\beta}, \sigma^2)^{(B)}$ can be drawn from the posterior.

As discussed in subsection 3.2.2, the CPO criterion or cross-validation density with "leave-out one" and "leave-out two" can be approximated using (2.16) and (2.18), and in general, cross-validation densities for different "leave-out sizes" can be approximated. Accordingly, the log EPBFs can be computed using the approximated cross-validation densities to compare two linear regression models in Bayesian context. Cross-validation densities in extended CPO criterion with "leave-out two" can be approximated by equation (2.18). Similarly, cross-validation densities for different "leave-out sizes" can be approximated; extended log pseudo Bayes factors computed using these approximated cross-validation densities can then be used as to compare linear regression models.

The comparative behavior of the extended CPO criterion or the EPBF for "Normal model with unknown mean and variance" setting with different "leave-out sizes" is discussed in the next subsection for both the closed form and MC samples based solution.

3.3.3 Comparative Behavior of Extended CPO Criterion for Unknown Variance Situation: Closed Form Versus Monte Carlo Samples

In this subsection, we examine how the extended CPO criterion, a model comparison tool behaves in real life scenario for both the cases: closed form and MC based solution with an illustrating example for "Normal model with unknown mean and variance" setting. As the "Normal model with unknown mean and known variance" situation discussed in subsection (3.2.3), we use the same data set, response variable, explanatory variables and the same model setup. Since, we have additional unknown parameter σ^2 compared to subsection (3.2.3), only the prior specification for the parameters needs some work.

Prior Specification: Unit Informative and g Priors

We need to specify the prior parameters $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, $\boldsymbol{V}_{\boldsymbol{\beta}}$, a, and b. But, it is hard to find representable values of these parameters for actual prior information, specifically for $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$. Again, with increasing regressors, the construction of an informative prior distribution gets harder. For example, with $p+1$ regressors, the number of prior correlation parameters is $\binom{p+1}{2}$, and it increases quadratically in p.

Sometimes one can use the least squares estimates $\hat{\boldsymbol{\beta}}_{OLS}$ as the prior mean for $\boldsymbol{\beta}$ in the absence of precise prior information; then, no probability statements about $\boldsymbol{\beta}$ can be made. Another idea is to use minimally informative prior as possible when the prior distribution doesn't represent the real prior information about the parameters. To some extent, using this compared to an informative prior distribution, more "objective" result can be obtained from the posterior distribution. Kass and Wasserman (1995) describes unit information prior as a weakly informative prior. The amount of information in a unit information prior is just the information contained in only a single observation. For example, $(\boldsymbol{X}^{\intercal}\boldsymbol{X})/\sigma^2$ denotes the precision of $\hat{\boldsymbol{\beta}}_{OLS}$ that can be thought of as the amount of information from n observations. Then, the unit information prior will set $\frac{1}{\sigma^2}\boldsymbol{V}_{\boldsymbol{\beta}} = (\boldsymbol{X}^{\intercal}\boldsymbol{X})/n\,\sigma^2$. Also, using $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$ is suggested by Kass and Wasserman (1995). This specification requires knowledge of $\boldsymbol{y}$ and hence cannot be a real prior distribution. However, for this unit information prior, only a small amount of information in $\boldsymbol{y}$ is used.

The idea of invariant parameter estimation to the scale change of the regressors can be implemented for choosing priors of $\boldsymbol{\beta}$. Suppose $\tilde{\boldsymbol{X}} = \boldsymbol{X}H$ where $\boldsymbol{X}$ represents a set of regressors and H is a $(p+1)\times(p+1)$ matrix. Also, suppose the posterior distributions of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ are obtained using $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ with same $\boldsymbol{y}$. Then, according to the invariance principle, both the posterior distributions of $\boldsymbol{\beta}$ and $H\tilde{\boldsymbol{\beta}}$ should be the same. We need to specify $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ and $\sigma^2\boldsymbol{V}_{\boldsymbol{\beta}} = k(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$, $k > 0$ to fulfill this condition. The choice of prior parameters with $k = g\sigma^2$ reveals a version of the well-known "g-prior" (Zenllner, 1986). Note that, with $g = n$, we get the unit information prior discussed in the previous paragraph.

For the regression parameters $\boldsymbol{\beta}$, we consider "g-prior" that is a multivariate normal prior with mean vector $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{0}$ and variance-covariance matrix $\sigma^2\boldsymbol{V}_{\boldsymbol{\beta}} = n\sigma^2(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$.

Using the prior specification discussed above and the model specification discussed in subsection 3.2.3, we obtain results that are described in the next subsection. As subsection 3.2.4, we describe our findings for the three models considered with three possible pairwise comparisons of these three models. Once again, we describe the behavior of the extended CPO criterion for the three models separately and then extend to three model comparisons using the EPBF for "Normal model with unknown mean and variance" situation.

3.3.4 Results

For "Normal model with unknown mean and variance" situation, first we visualize the closed form results of log extended CPO values obtained from Model 1, Model 2 and Model 3 (see Figure 3.7) as discussed for "Normal model with unknown mean and known variance" in subsection 3.2.4.

All three models exhibit the same pattern with decreasing log extended CPO values over the increasing leave-out size with a sharp decrease in an exponential manner after "leave-out 50". Also, all three models have close log extended CPO values. The relative position of the three models remains more or less the same at each "leave-out size" though compared to other two models, Model 2 and Model 3 have a slightly higher log extended CPO values at smaller and higher "leave-out sizes". So, compared to the known variance case, from Figure 3.7 we observe that

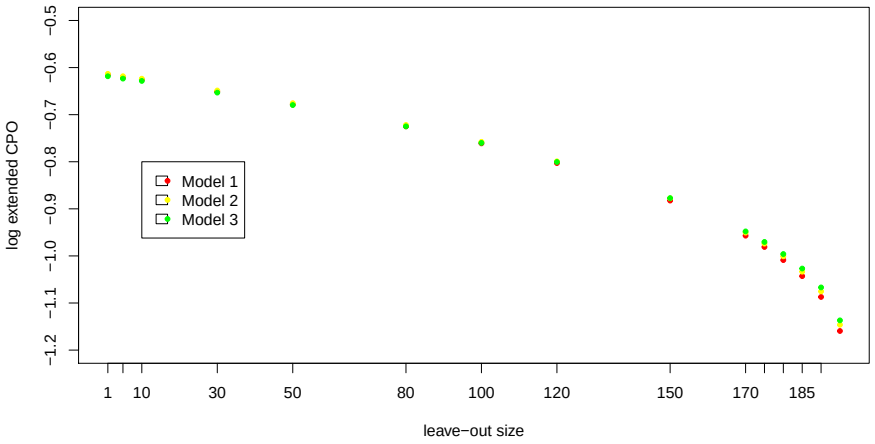

Figure 3.7: Comparison of closed form results for three models (unknown mean and variance)

the strength of evidence changes for the considered models at different "leave-out sizes". We hope that this pattern will be more clearly observed when we compare the models pairwise through log EPBFs as a model selection criterion.

Along the closed form results, we want to examine the strength of evidence for these models at all "leave-out sizes" when we must compute using MC samples. As we observed in the "known variance" situation, the strength of evidence might be affected due to the bias related to the results obtained from the MC samples that increase with the increasing leave-out points. Once again, we need to examine further this bias issue. Also, as the "known variance" situation, we want to compare the closed form results and results based on MC samples to check whether MC based results can replicate the closed form results in "unknown variance" situation. For the Model 1 we observe the log extended CPO values from the closed form solutions as well as from the MC samples in the Figure 3.8. The red filled circles represent the closed form values at each "leave-out size".

Compared to "known variance" situation (up to "leave-out 120"), the MC based results match the corresponding closed form result up to a smaller "leave-out sizes" ("leave-out 50") that can be observed from Figure 3.8. After "leave-out 50", the

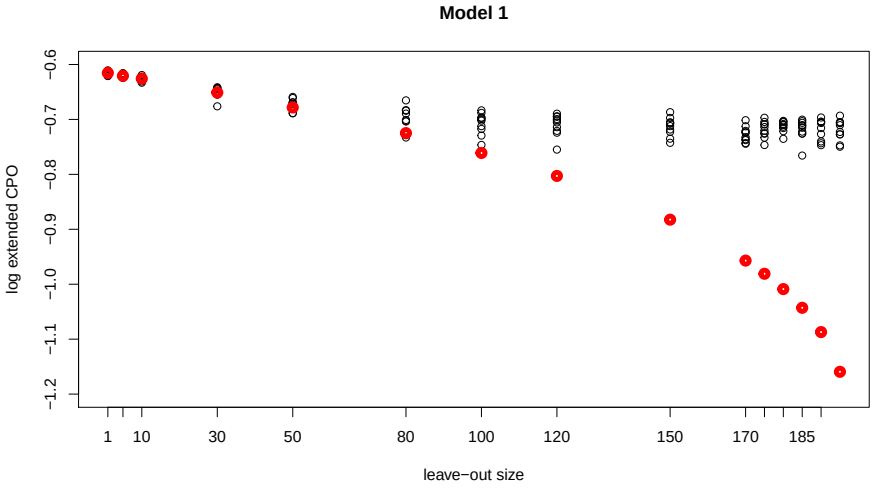

Figure 3.8: Comparison of closed form results and MC based results for Model 1 (unknown mean and variance)

changes between the results from MC samples and closed forms become visible and at higher "leave-out sizes", MC based results show an upward (exponentially increasing) bias from the closed form result. This pattern is also observed for the "known variance" situation discussed in subsection 3.2.4. Also, for this "unknown variance" situation, the same pattern is observed for Model 2 and Model 3 as we observe the same pattern for the other two models considered here. For three models, a substantive positive departure of MC based results from the closed form values is observed at higher "leave-out sizes".

We compare Model 1 with Model 2 now in the similar manner as discussed in in subsection 3.2.4. The EPBF, computed from the extended CPO is used as a model comparison tool. This comparison can be described from Figure 3.9 where the red filled circles and black circles denote the closed form results and MC results from 10 batches at each "leave-out size" respectively. Same specifications are used for Figure 3.10 and Figure 3.11.

Figure 3.9 exhibits the same pattern in comparing Model 1 with Model 2 as Figure 3.3 except the wider values of the MC based results. We observe negative closed form log EPBFs at all "leave-out sizes" suggesting the choice of Model 2

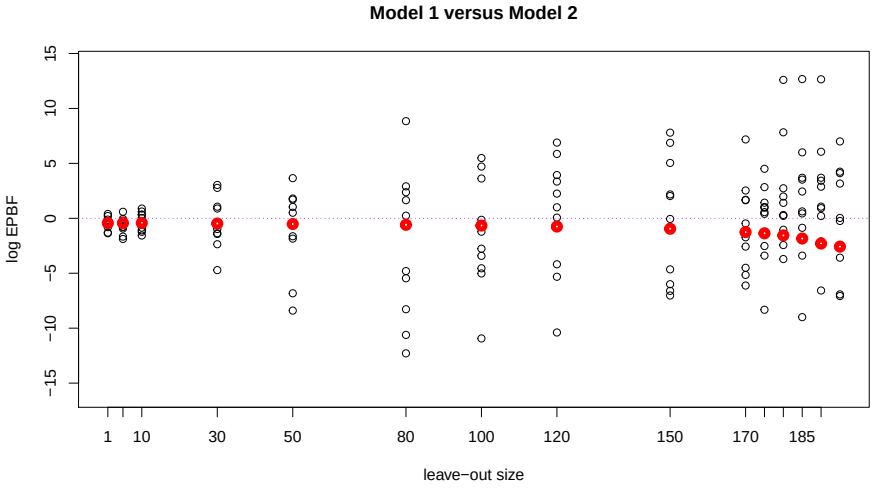

Figure 3.9: Comparison of Model 1 and Model 2 (closed form results and MC based results for "unknown mean and variance" situation)

over Model 1 (hence the same choice) with increasing strength of evidence for increasing "leave-out sizes". However, MC based results vary over the direction starting from early "leave-out sizes", and have values with wider range at higher "leave-out sizes". As discussed before, the fluctuation from the closed form results is due to the MC error. We will quantify the MC error to examine how this increases with increasing 'leave-out sizes".

The closed form results are near the center of the MC based results at smaller "leave-out sizes" (1, 5, and 10). For other "leave-out sizes", the MC based results from most of the 10 batches have a tend to have higher values than the corresponding closed form results. As the "known variance" situation, Figure 3.9 demonstrates that MC based results become positively biased and more variable as the "leave-out size" increases for "unknown variance" situation. MC based results from most of the 10 batches show positive bias from the formal Bayes factor value (closed form value at "leave-out all") implying that using MC samples we are very poorly computing real Bayesian model comparison with some positive MC errors. Our goal is to examine the possibility of finding a "leave-out size" where the EPBF is close to the formal Bayesian comparison with smaller MC error which is dis-

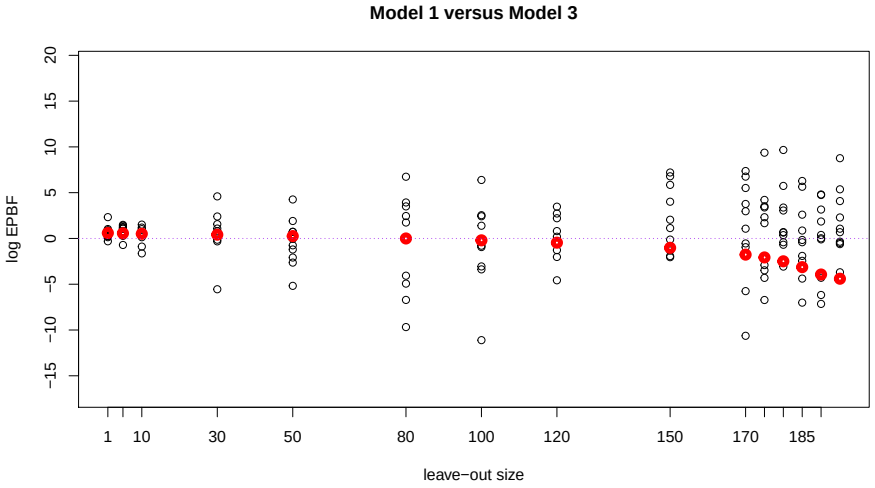

Figure 3.10: Comparison of Model 1 and Model 3 (closed form results and MC based results for "unknown mean and variance" situation)

cussed later.

Next we examine two other possible model comparisons: Model 1 versus Model 3 and Model 2 versus Model 3. Figure 3.10 displays the closed form and MC samples based log EPBFs for Model 1 versus Model 3.

Figure 3.10 displays similar pattern of choosing the smaller model as Figure 3.9 with increasing "leave-out sizes". Compared to the comparison between Model 1 and Model 2, for the comparison between Model 1 and Model 3, the closed form results indicate a change pattern in the choice of the model (changes direction at "leave-out 80") over different "leave-out sizes". Up to "leave-out 50", the closed form log EPBF values are positive indicating the evidence of choosing Model 1 over Model 3 though this evidence decreases with increasing "leave-out sizes". Negative closed form log EPBF values at "leave-out 80" and higher "leave-out sizes" suggest the choice of Model 3 over Model 1 and the strength of this evidence increases with increasing "leave-out sizes" from "leave-out 80" and onwards. The choice of model alters here for "unknown variance" situation which doesn't alter in the "known variance" situation though the pattern of stronger evidence of smaller model with increasing "leave-out sizes" is prevalent in both situations.

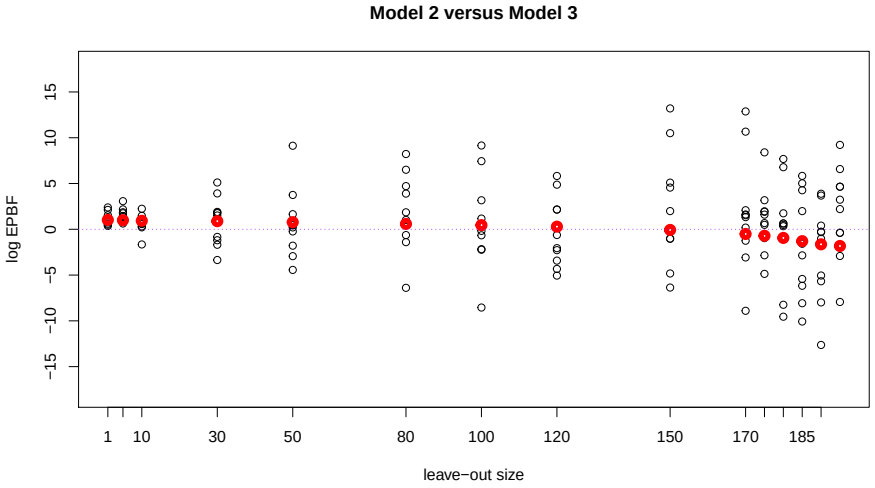

Figure 3.11: Comparison of Model 2 and Model 3 (closed form results and MC based results for "unknown mean and variance" situation)

For "unknown variance" situation, the closed form results are near the center of the MC based results from 10 batches at early "leave-out sizes" for the comparison between Model 1 and Model 3 as the comparison between Model 1 and Model 2 (see Figure 3.10). Again, the MC based results become positively biased and more variable as the "leave-out size" increases which is similar as previous comparison (Figure 3.9) as well as the same comparison for "known variance" situation shown in Figure 3.4. Compared to the "known variance" situation, the MC samples based results have a wider range with large MC errors.

At last, we focus on the last comparison: Model 2 versus Model 3 in "unknown variance" situation. The closed form and MC based results for this comparison are displayed in Figure 3.11. Like the previous comparison displayed in Figure 3.10, the closed form results in this comparison indicate the same pattern of choosing bigger model (Model 2) at early "leave-out sizes" and smaller model (Model 3) at higher "leave-out sizes". The strength of evidence for bigger model decreases up to "leave-out 120", and changes of direction (evidence of smaller model) is observed at "leave-out 120" and onwards. Hence, Figure 3.10 suggests the choice of Model 2 up to "leave-out 120" and Model 3 from "leave-out 150". Compared to "known

variance" counterpart, similar decreasing pattern in log EPBF values is observed though direction of model choice changes in "unknown variance" situation. This implies that possibly the magnitude of change in log EPBF values for increasing "leave-out sizes" is greater for the "unknown variance" than the "known variance" situation. The same behavior of the MC based results at all "leave-out sizes" is observed in Figure 3.11 as Figure 3.9 and Figure 3.10 with wide range of values. As previous, MC based results are positively biased and more variable as the "leave-out size" increases that might be due to MC error.

In short, all three model comparisons show us the same behavior of the EPBF (computed from extended CPO criterion) as a model selection tool. The close form results have a tendency to choose the model with a small number of parameters in all three comparisons with increasing "leave-out sizes". For comparisons between Models 1 and 3 and Models 2 and 3, the values of log EPBF become negative from positive at some "leave-out sizes" that indicates the change in the evidence for model selection. Moreover, the MC based results have a wider range of the "unknown variance" situation compared to the "known variance" situation, and deviate from the closed form results in all three comparisons in an increasing pattern with the increasing "leave-out sizes". MC error might be responsible for this deviation so that our intention is to compute the contribution of MC error at different "leave-out sizes" which we discuss in the following subsection.

3.3.5 Summarizing the Computations for Unknown Variance Situation

As "known variance" situation, we compute the root mean squared error (RMSE) of the MC based estimates of the log EPBFs at all considered "leave-out sizes" to examine the variation from the closed form results for "unknown variance" situation. The RMSE of the estimated log EPBFs for all three model comparisons are tabulated in Table 3.2.

From Table 3.2 for comparison of Model 1 with Model 2 at "leave-out 10", the RMSE of the estimated log EPBFs is 0.863 which implies that model comparisons using MC based EPBFs relative to the corresponding closed form are erroneous by a factor of $\exp(0.863) = 2.37$. Similarly, while comparing Model 1 with Model 3, at "leave-out 120", the RMSE of the estimated log EPBFs is 2.460 which leads to a

Table 3.2: Root mean squared errors of the estimated log EPBFs for all three model comparisons in unknown variance situation

Leave-out size	Model 1 vs. Model 2	Model 1 vs. Model 3	Model 2 vs. Model 3
1	0.575	0.715	0.690
5	0.771	0.716	0.900
10	0.863	0.968	1.043
30	2.370	2.575	2.626
50	3.884	2.694	3.777
80	7.069	5.388	4.419
100	5.128	4.693	5.044
120	5.558	2.460	3.871
150	5.775	4.807	6.659
170	4.140	6.318	6.633
175	3.810	5.639	3.942
180	6.034	5.748	5.542
185	6.697	5.164	5.759
190	6.785	5.645	5.326
all	5.400	7.072	6.210

factor of 11.7 erroneous model comparisons while using MC based EPBFs relative to the corresponding closed form. We can interpret all other RMSE values in a similar fashion.

However, the RMSE values of the estimated log EPBFs increase with the increasing "leave-out size" and these increase drastically at higher "leave-out sizes" for three model comparisons. As the "known variance" situation, we use several cut-off values (two here) for the RMSE values to examine the level of error in the model selection that occur due to using the MC based EPBFs at different "leave-out sizes". The RMSE values from Table 3.1 are in logarithmic scale, and we consider two cut-off values for these values: log 1.25, and log 2 which correspond to 25, and 100 percent erroneous model selection while using the log EPBFs estimated from MC samples instead of the closed form values. The RMSE values for the MC based EPBFs at the considered "leave-out sizes" with vertical lines through the cut-off values are plotted in Figure 3.12.

We are observing a lot of error in model selection for "unknown variance" situation from Figure 3.6. If we consider the log 1.25 as the cut-off value, then we

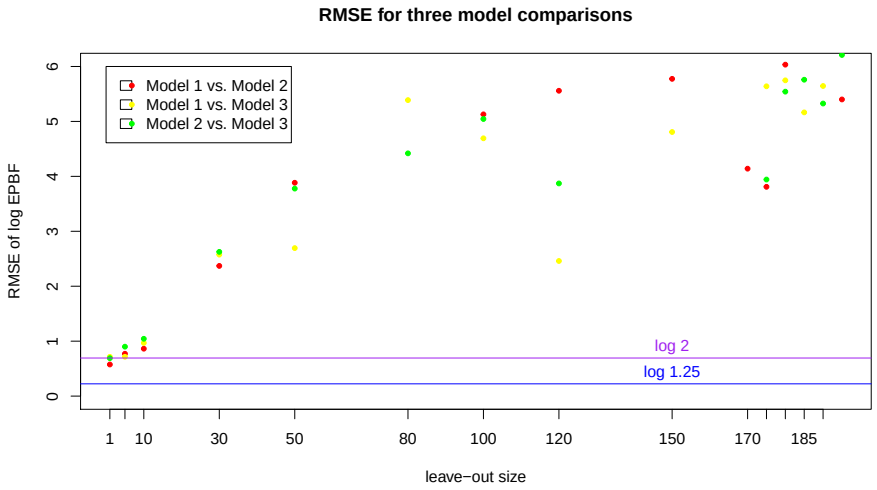

Figure 3.12: Root mean squared error of the log EPBFs from 2500 MC samples with three cut-offs (unknown mean and variance)

observe that no RMSE values lie below this cut-off value for all three comparisons. The RMSE values for three model comparisons at "leave-out 1" lie under the log 2 cut-off value only and exceed the cut-off value at any "leave-out size" greater than 1. So, compared to "known variance" situation, we require larger MC samples or more batches to get accurate model comparisons from the MC based results.

Also, according to the discussion of increasing MC sample sizes to get more correct model comparisons based on the MC sample in subsection 3.2.4, we increase the MC samples from 2500 to 25000 that is now we have 10 batches with 2500 MC samples. According to MC rule, we expect an error reduction from the MC based model comparisons by a factor of 8.95 (based on the formulation in subsection 3.2.4) with this 25000 MC samples compared to the 2500 MC samples. This implies an expectation to have all the RMSE values less than $\log(8.95) = 2.19$ in this Figure 3.12 that is up to "leave-out 30" below the cult-off value $\log(2)$ with the 25000 MC samples. We plot the revised RMSE values considering 25000 MC samples for three model comparisons in Figure 3.13.

As per our expectation, Figure 3.13 shows that all the RMSE values for three model comparisons up to "leave-out 30" lie below the cut-off point $\log(2)$. In

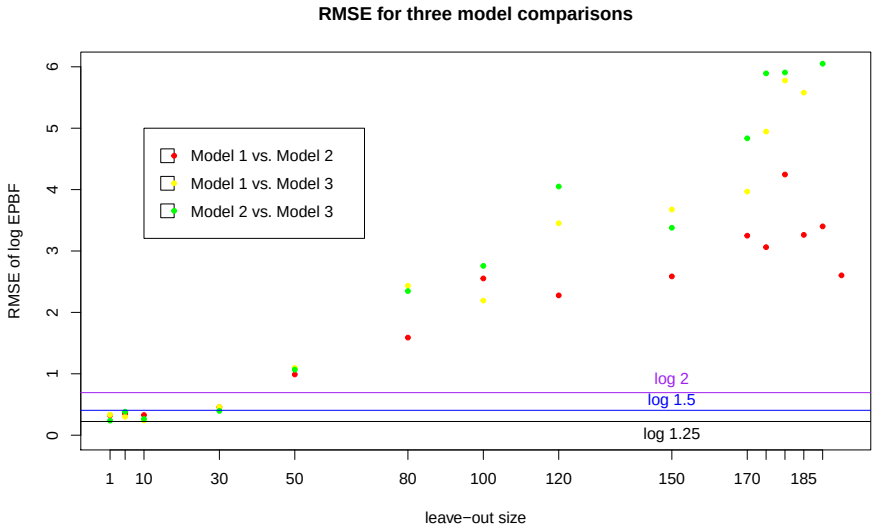

Figure 3.13: Root mean squared error of the log EPBFs from 25000 MC samples with three cut-offs (unknown mean and variance)

addition, we observe some RMSE values (up to "leave-out 10") below the cut-off point $\log(1.5)$. Hence, the MC results show a lot of improvement in error reduction with the 25000 samples compared to the 2500 samples.

3.4 Summary

In this chapter, we discussed the Bayesian model comparison for linear regression model with the extended CPO criterion as the model selection tool. Two linear regression models are compared using extended pseudo Bayes factor (EPBF) which is a compromise between the pseudo Bayes factor and the formal Bayes factor. We consider 15 different "leave-out sizes" where the model comparisons at "leave-out 1" and "leave-out all" indicate pseudo Bayes factor and formal Bayes factor respectively. Two different situations namely "normal model with unknown mean and known variance" and "normal model with unknown mean and variance" are considered for the Bayesian linear regression model and discussed in detail with an example in sections 3.2 and 3.3. Since the closed form results are available, we compare the MC based results with these to examine how much the model comparison results deviate from the corresponding closed form ones in both the

situations. We take a single MC sample and then create 10 equal-sized batches from that sample. Model comparison results are computed for 10 batches. Three linear regression models: Model 1, Model 2, and Model 3 are considered here that contain five, four and three explanatory variables respectively. This implies that we have three pairwise model comparisons.

There are several take-away messages from the closed form results obtained in both the situations. One of these is the change in the evidence of the model at different "leave-out sizes". For "normal model with unknown mean and known variance" situation, closed form results of the three model comparisons exhibit same choice of models over different "leave-out sizes". But, for the "normal model with unknown mean and variance" situation, the choice of model changes at some "leave-out size" for the comparisons between Model 1 and Model 3 and between Model 2 and Model 3. This implies that the direction of evidence under the pseudo Bayes factor and the formal Bayes factor is different in these comparisons though pseudo Bayes factor is used as a proxy of the formal Bayes factor in the literature (Gelfand and Dey, 1994). In addition, using the EPBF values at different "leave-out sizes", we can exactly specify the "leave-out size" up to which the evidence of a specific model remains the same and changes hereafter. Another take-away message from the closed form solutions is that for both the situations, the strength of the evidence increases for the comparatively smaller model (regarding the number of explanatory variables) among the two models compared in all the comparisons.

The MC samples based results show a rapid departure from the closed form results with increasing "leave-out sizes" for all the model comparisons in both the situations. The ranges of the MC results are much higher in the "normal model with unknown mean and variance" situation compared to the other situation. We compute RMSE of the MC samples based estimates of the log EPBFs for three model comparisons to measure how well the MC samples based estimates approximate the closed form log EPBF. With 2500 MC samples, MC based estimates exhibit less than 25% error in model selection compared to the closed form results for "leave-out size" 50 or smaller in "normal model with unknown mean and variance" situation. However, RMSE values of the MC samples based estimates of the log EPBFs are too high in "normal model with unknown mean and variance" situation. Error in the decision of model selection from the MC based results can

be reduced by increasing the number of MC samples.

This chapter focuses on Bayesian model comparisons for linear regression model using extended CPO or EPBF as a model comparison tool. Both the closed form and MC samples based results are available here. For the generalized linear models, for example, logistic regression models, no closed form solutions are available, and model comparisons are examined using some form MC samples only. We will discuss the Bayesian model comparisons for logistic regression models in the next chapter.

Chapter 4

Bayesian Model Comparison for the Generalized Linear Models

We discussed the Bayesian model comparison for the linear regression models in the previous Chapter. In this Chapter, we focus on the Bayesian model comparison for the generalized linear models, for example, logistic regression models. Logistic regression models are used in studying the effect of explanatory variables on a nominal response variable (response is continuous for linear regression models). As with linear regression models, both the classical and Bayesian model selection methods can be applied to logistic regression models. For example, one can use either a classical approach say AIC or any Bayesian approach. In this chapter, we discuss the extended pseudo Bayes factor (EPBF) as a model selection method for logistic regression models in Bayesian context. Our interest is to examine how the model selection behaves when we change the "leave-out size" in the EPBF. From the EPBF, we can get the pseudo Bayes factors (PBFs) as well as the formal Bayes factors (BFs) depending on the "leave-out sizes". We want to examine whether there is any agreement or disagreement in between the PBFs and the formal BFs as model selection methods.

4.1 Bayesian Logistic Regression Models

We can apply the Bayesian approach to estimate the parameters of the logistic regression models. In the classical approach, we only need the likelihood function for estimation purposes. But, in Bayesian approach, in addition, we do require a prior specification for parameters of the logistic regression model. The logistic regression model constructs a model to predict the probability of the presence of an indicator using the available explanatory variables. The log transformation of the ratio of probabilities, known as log odds or logit, linearizes the relationship between the response and the explanatory variables. As the linear regression case, we start our discussion with defining a logistic regression model. Our response variable, say $\boldsymbol{y}$ has binary responses. Suppose the binary response variable represents an indicator (0 =Absence, 1 =Presence) of an event. Also, suppose we have a set of explanatory variables: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$. Here, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector of length n representing binary responses of interest for n observations.

Suppose $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\intercal$ are unknown regression parameters and $\boldsymbol{x}_i^\intercal = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ represents the i^{th} individual's row vector of explanatory variables. The design matrix [of dimension $n \times (p+1)$] in this case is $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$. The response variable for the i^{th} individual indicates the presence or absence of the event for that subject by setting $y_i = 1$ and $y_i = 1$ respectively. If $p(\boldsymbol{x}_i)$ represents the probability that the event is present for subject i, and the i^{th} individual's set of the explanatory values are contained in $\boldsymbol{x}_i^\intercal$, then the logistic regression model can be written as:

$$\operatorname{logit} p(\boldsymbol{x}_i) = \log\left[\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}, \;\; i = 1, \ldots, n. \tag{4.1}$$

Now, we can express the probability $p(\boldsymbol{x}_i)$ of the presence of the event as

$$p(\boldsymbol{x}_i) = \Pr(y_i = 1 | \boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\intercal \boldsymbol{\beta}}}.$$

As the response is binary, the likelihood contribution from the i^{th} observation can

be written using a Bernoulli likelihood expression:

$$\begin{aligned} L_i(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) &= [p(\boldsymbol{x}_i)]^{y_i} [1 - p(\boldsymbol{x}_i)]^{(1-y_i)} \\ &= \left[\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right]^{y_i} \left[1 - \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right]^{(1-y_i)}. \end{aligned}$$

The individuals are assumed to be independent from each other, and hence the likelihood function over the n individuals has the following expression:

$$L(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right]^{y_i} \left[1 - \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right]^{(1-y_i)}. \quad (4.2)$$

We can utilize the likelihood function (4.2) to estimate the unknown parameters $\boldsymbol{\beta}$ in classical inference. Also, we need to specify the prior distribution, say $g(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ for the unknown parameters $\boldsymbol{\beta}$ so that we can compute the posterior distribution $p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X})$ to obtain the Bayesian inference of those parameters. A general expression of the posterior follows:

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}) \propto L(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) \times g(\boldsymbol{\beta}). \quad (4.3)$$

The posterior in the equation (4.3) has no closed form expression. But, we can take realizations from the expression in equation (4.3) to obtain a valid empirical guess about the posterior distribution in a Bayesian manner, and hence make inference for the parameters. There are many proposed approaches available in the literature which includes the Metropolis−Hastings (MH) method, many latent-variable schemes that facilitate Gibbs sampling, etc. and so on to obtain Markov chain Monte Carlo (MCMC) samples from the posterior.

We use one of these approaches in this study which is efficient enough and popularly known as the 'Pólya−Gamma approach' (Polson et al., 2013). If we parameterize the binomial likelihoods (say 4.2) by the log odds, then the assumption in the 'Pólya−Gamma approach' is that the likelihood is expressed as a mixture of normals. It can be noted that the Pólya−Gamma distribution is a subset of the class of infinite convolutions of gamma distributions (Polson et al., 2013), and generalization of the Pólya distributions (Barndorff-Nielsen et al., 1982). The

'Pólya−Gamma approach' is implemented in a `R` package called `BayesLogit` which utilizes an accept/reject sampler based on the alternating-series method that is proposed by Devroye (1986). This sampler is very efficient and requires exponential and inverse-Gaussian draws only; also, the bound of the probability of accepting a proposed draw is uniformly bounded below at 0.99919 (Polson et al., 2013). Also, no tuning is needed which makes this a reliable black box sampling routine in all situations with the logit link, even in complex hierarchical models. As per the claim from Polson et al. (2013), 'Pólya−Gamma approach' is nearly efficient as the independence MH sampler for simple logistic models with no hierarchical structure, and most efficient in all other cases.

We can directly use the MCMC samples from the posterior distribution of $\boldsymbol{\beta}$ to approximate the cross-validation densities as no closed form densities of the predictive posterior distribution are available for Bayesian logistic regression models. Then, as discussed in Chapter 3 for the linear regression models, we can compute the extended conditional predictive ordinate (CPO) criterion. However, in the absence of the closed form results, we can't make a comparison between the closed form and posterior samples based results for the Bayesian logistic regression models. But, motivated from examining how well the closed form results can be approximated by the posterior samples for the linear regression models in the Chapter 3, we can rely on such MCMC samples based results. Then the estimated extended CPO criterion can be used to compute the extended pseudo Bayes factor (PBF) to compare the Bayesian logistic regression models for model selection purpose.

4.2 Extended CPO Criterion using MCMC Samples for Logistic Regression Models

We describe the computation of the extended CPO criterion directly from the posterior MC/MCMC samples in the subsection 2.4.4 of the Chapter 2. According to that, at first we draw B realizations of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(B)}$ from the expression of the posterior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X})$ in equation (4.3). Then, having B posterior samples of $\boldsymbol{\beta}$, as the linear regression models, the CPO criterion or cross-validation density for logistic regression models with "leave-out one" can be

computed approximately using (2.16). After that, we can compute log PBFs using the approximated cross-validation densities to compare two logistic regression models in Bayesian context. Again, cross-validation densities in extended CPO criterion with "leave-out two" can be approximated by using the formulation in equation (2.18). Similarly, cross-validation densities for different "leave-out sizes" can be approximated; log EPBFs computed using these approximated cross-validation densities can then be used as to compare logistic regression models in Bayesian context as the linear regression models discussed in Chapter 3.
We discuss the behavior of the extended CPO criterion or the EPBF in the next section for Bayesian logistic regression models with different "leave-out sizes" in a real life scenario.

4.3 Examining the Behavior of Extended CPO criterion: A Practical Example

Here, we attempt to examine how the extended CPO criterion, a Bayesian model comparison tool, behaves with an illustrative example. We use a well-known data here named `birthwt` which is available in the `R` package `MASS`. The location of the data collection is Baystate Medical Center, Springfield, Mass (Venables and Ripley, 2002). The aim of collecting the data was to investigate whether some factors related to mother are responsible for low child birth weight. From this specific data we consider the following five explanatory variables:

- *smoke*: mothers' smoking status during pregnancy (1 = Yes, 0 = No),
- *ui*: presence of uterine irritability for mothers (1 = Yes, 0 = No),
- *ht*: hypertension status of mothers (1 = Yes, 0 = No),
- *lwt*: mothers' weight at last menstrual period (in pounds), and
- *ptl*: mothers' previous premature labours (numbers).

We have three binary, one continuous (*lwt*), and one count (*ptl*) explanatory variables. Our binary response variable is *low* which is simply an indicator of birth weight less than 2.5 kg (1 = Yes, 0 = No). We want to fit a logistic regression

to model the probability that a child is born with low birth weight with respect to factors related to the mother of that child. There are 189 non-missing observations for this `birthwt` data. We specify the models considered for comparisons and priors for the regression parameters $\boldsymbol{\beta}$ below.

4.3.1 Model Specification

Two models are considered here, and these are denoted by Model 1 and Model 2 with different combinations of the explanatory variables. If p denotes the probability that a child is born with low birth weight (low = 1), then these two models have the following specifications.

1. Model 1: Big Model or Full Model (Model with all explanatory variables listed above) with the formulation

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\, smoke_i + \beta_2\, ui_i + \beta_3\, ht_i + \beta_4\, lwt_i + \beta_5\, ptl_i,$$

   where p_i denotes the probability that i^{th} child is born with low birth weight, $i = 1, 2, \ldots, 189$. All explanatory variables represent the binary status of different factors related to the mother of the i^{th} child.

2. Model 2: Small Model (Model which leaves only ptl out of the full Model).

Previously, for linear regression models discussed in Chapter 3, we consider 2500 MC samples from the posterior distribution of the regression parameters $\boldsymbol{\beta}$. MC samples are taken independently whereas MCMC scheme provides subsequent dependent samples. Hence, we consider a large number of MCMC samples (100000) from the posterior distribution of the regression parameters $\boldsymbol{\beta}$ for the computation of the extended CPO criterion in the Bayesian logistic regression setting. We divide those into 10 batches of equal size 10000. As the linear regression models, we consider the log extended CPO and log extended pseudo Bayes factor calculation for each of these batches at different "leave-out sizes" (from "leave-out 1" to "leave-out all"). For logistic regression models, we consider 11 different "leave-out sizes"; these are 1, 5, 10, 30, 50, 80, 100, 120, 150, 170, and 189 (i.e., all). Also, to calculate extended CPO calculation with different "leave-out sizes",

we take all combinations if the total combinations are less than 2000 and take only 2000 combinations randomly if the total combinations exceed 2000.

4.3.2 Prior Specification: Weakly Informative Prior

We use the efficient 'Pólya−Gamma approach' for generating MCMC samples from the posterior distribution (4.3) for Bayesian logistic regression models. However, the 'Pólya−Gamma approach' incorporates the normal prior for the parameters only. Different prior specifications for the parameters of the logistic regression model are also available in the literature. For example, a weakly informative prior is suggested by Gelman et al. (2008).

A weakly informative prior is a minimally informative prior as possible, and mostly used when there is a doubt that the prior distribution might not represent the real prior information about the parameters. Gelman et al. (2008) propose a weakly informative prior on the scaled explanatory variables. Scaling is an important issue for the logistic regression models as different scaling can render the exponentiation of the parameter coefficients (the odds ratios) very different. Hence, the explanatory variables need standardization, and one such application can be found in Raftery (1996) for Bayesian generalized linear models.

Two proposals are made by Gelman et al. (2008). For binary explanatory variables, a shift is suggested so that their means are 0s and ranges are 1s. Shifting and scaling for the non-binary explanatory variables are suggested in a way so that their mean and standard deviation become 0 and 0.5 respectively. With this scaling, the continuous variables have the same scale as the symmetric binary variables in the interval $[-0.5, 0.5]$ with standard deviation 0.5. In our 'Pólya−Gamma approach' to Bayesian logistic regression models, we use this scaling on the explanatory variables. Also, we have normal prior distribution: $\beta_j \sim N(\mu_j, \sigma_j^2), \quad j = 0, 1, \ldots, p$ for the regression parameters $\boldsymbol{\beta}$. With this prior, the posterior distribution of the parameters of the logistic regression model 4.3 can be rewritten as

$$p(\boldsymbol{\beta}\,|\,\boldsymbol{y},\boldsymbol{X}) \propto \prod_{i=1}^{n}\left[\frac{e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}\right]^{y_i}\left[1-\frac{e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}}\right]^{(1-y_i)} \times \prod_{j=0}^{p}\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left[-\frac{1}{2}\left(\frac{\beta_j-\mu_j}{\sigma_j}\right)^2\right]. \tag{4.4}$$

Now, we need to specify the hyperparameters for the inference of the parameters $\boldsymbol{\beta}$ using equation (4.4). We use independent normal prior distributions with mean 0 and variance $\log(5)$ for all the parameters in the logistic regression model except the intercept term. This implication comes from two suggestions. Firstly, we consider the epidemiological idea of prior construction using the odds ratio (Greenland, 2006). The value of odds ratio = 5 is considered as a meaningful upper bound of the possible odds ratio. We consider this as the variance of the individual parameters. The second suggestion comes from Gelman et al. (2008) that recommends 5 as the upper bound for the absolute difference in the logit probability. If we consider the variance of the individual normal priors as $\log(5)$, then both the suggestions can be taken into account.

For the intercept, we consider a normal prior with mean 0 and a higher variance than the other variances (hyperparameters) taken into account for the other parameters. Gelman et al. (2008) documents that a change of 5 in the input implies the change in the probability either from 0.01 to 0.5, or from 0.5 to 0.99 for logistic regression which indicates that a change of 10 leads to the change in the probability from 0.01 to 0.99. According to this concept, for the intercept we use a normal prior with mean 0 and variance $\log(10)$. Such specification allows that on an average, the expected probability of success is within the bound $[0.01, 0.99]$.

We discuss the results in the next subsection using the prior specification above and the model specification in subsection 4.3.1. As the results for model comparisons for the linear regression models in Chapter 3, we describe our findings for model comparison for the two models considered for different "leave-out sizes" using the EPBF.

4.3.3 Results

We do not have closed form model comparison results for Bayesian logistic regression models which we examined for Bayesian linear regression models. Model comparison results based on the MCMC samples between Model 1 and Model 2 for different "leave-out sizes" is displayed in Figure 4.1 as discussed in subsection 4.3.1. Here, the black triangles represent the log EPBF values from 10 different batches of size $10,000$ and the red dots represent the mean log EPBF values obtained from those 10 batches at each "leave-out size".

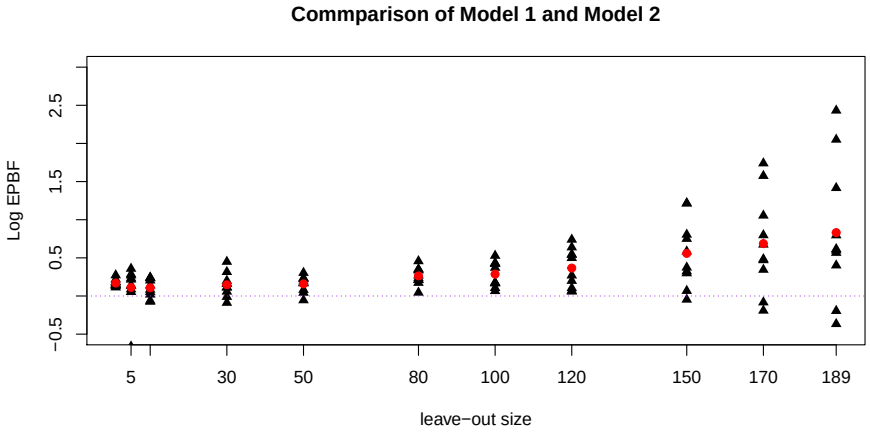

Figure 4.1: Comparison of big Model vs. small Model in Bayesian logistic regression

Figure 4.1 represents how the model comparisons behave while comparing Model 1 and Model 2 using the log EPBFs. As we have positive log EPBFs at all "leave-out sizes" except some batches at some small and large "leave-out sizes", the computed log EPBFs from the MCMC samples suggest the choice of the big Model (Model 1) over the small Model (Model 2) at all "leave-out sizes". From the estimated mean log EPBFs at each "leave-out size", this pattern is observable. As the results found for the linear regression models in the Chapter 3, the strength of evidence increases with increasing "leave-out sizes" for the logistic regression models. So, a pattern of the strength of evidence for a specific model over the increasing "leave-out sizes" is observed in both the linear and logistic re-

gression model comparisons. However, the pattern suggests a stronger evidence for the small Model over the big Model in linear regression cases but the reverse (big Model over the small Model) for logistic regression cases though these are data-driven decisions.
From Figure 4.1 it is clear that, MCMC based results vary over the direction at some "leave-out sizes" (for "leave-out sizes" 10, 30, 50 and 150 and above). Monte Carlo (MC) error might be responsible for this fluctuation. If possible, quantifying the MC error will allow us to examine how this increases for increasing 'leave-out sizes".

For the smaller "leave-out sizes" (for example, 1, 5, 10), the MCMC based model comparison results (log EPBF values) from 10 batches are close to the estimated mean of the model comparison results of those batches. For the larger "leave-out sizes", say "leave-out 100" or more, the MCMC based results from the 10 batches tend to be more spread (more in the right direction) from their mean value. Hence from Figure 4.1, we can say that as the "leave-out size" increases, the MCMC samples might produce more and more variable and probably positively biased model comparison results for the Bayesian logistic regression models. In Figure 4.1, the EPBF at "leave-out all" indicates the computed formal Bayes factor using MCMC samples. From the theoretical perspective, this is our optimal model selection criterion. However, the Figure 4.1 suggests that we are very poorly computing real Bayesian model comparison with some positive MCMC errors for logistic regression model comparison.

As the linear regression models, we have the interest in whether it is possible to find a "leave-out size" where the EPBF is close to the formal Bayesian comparison with smaller MCMC error which is discussed in the next subsection.

4.3.4 Summarizing the Computations

We can examine how well the posterior samples (MC or MCMC) can approximate the closed form model comparison results given that the closed form results are available. We have already done this for the linear regression models in Chapter 3. But, we do not have the same setup for the logistic regression models due to unavailability of the closed form results. Hence, to examine how well the posterior

MCMC samples compute the model selection criterion EPBF, we focus on the estimated mean and standard error of the log EPBFs obtained from 10 batches of MCMC samples with size 10000. We plot the estimated error bars for the 10 log EPBFs at each of the "leave-out sizes" in Figure 4.2.

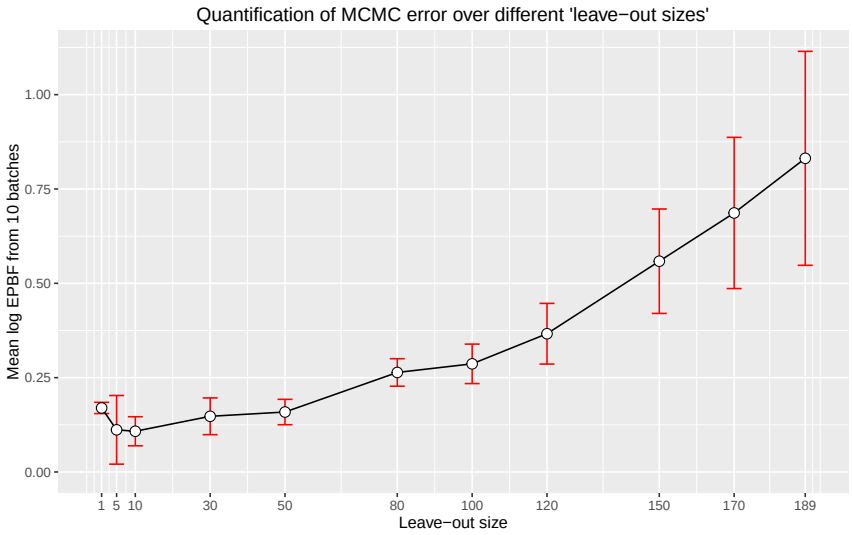

Figure 4.2: Error bar plot of the estimated log EPBFs for Bayesian logistic regression

Figure 4.2 has a very straightforward interpretation about the behavior of the MCMC samples. It is observed that the width of the error bars for the estimated log EPBFs increases with the "leave-out sizes" except "leave-out 5". So, the standard error of the estimated log EPBFs increases with "leave-out sizes" except that "leave-out size" implying higher variability among the estimated log EPBFs from 10 batches at larger "leave-out sizes". For example, at "leave-out 1", the standard error of the estimated log EPBFs is 0.0148 with estimated mean 0.1697 whereas at "leave-out 150", the standard error is 0.1384 with estimated mean 0.5586, and at "leave-out all", the standard error is 0.2835 with estimated mean 0.8313. The standard error of the estimated log EPBFs increases rapidly for the larger "leave-out sizes" that are close to the "leave-out all" than the smaller "leave-out sizes". The higher variability in the estimated log EPBFs for "leave-out 5" might be due to random chance.

We are not surprised by this overall findings as we have already observed higher variability among the estimated log EPBFs obtained from MC samples for linear regression models discussed in Chapter 3. The same thing happens here. Also, being independently drawn samples, MC samples are less vulnerable to the MC errors than the dependent MCMC samples. However, we have had a direct formulation of improving the model selection results by increasing a specific number of MC samples for the linear regression models discussed in Chapter 3. We can not do it here as no closed form solution is available for model comparison using EPBF for logistic regression models. Still, we can generalize the findings of the Chapter 3 in MC error reduction.

For our analysis, we have 100000 MCMC samples and 10 equal sized batches with 10000 MCMC samples. Either increasing the total MCMC samples from 100000 or the number of batches with 10000 MCMC samples each will improve the result by reducing the error in model selection for sure. As a compromise between the pseudo Bayes factor and the formal Bayes factor, the EPBF allows us to choose "leave-out sizes" greater than one with some errors. From the Figure 4.2, we can think of "leave-out sizes" up to 80 with small variability (and hence small error) among the estimated log EPBFs. Then, we can get close to the estimated optimal formal Bayesian comparisons ("leave-out all") more than the estimated pseudo Bayes factors ("leave-out 1"). Also, with increasing posterior MCMC samples, we can go further closer to the estimated optimal formal Bayesian comparisons i.e., the Bayes factors by using larger "leave-out sizes".

4.4 Summary

We discussed the Bayesian model comparison for the logistic regression models with the extended CPO criterion or the extended pseudo Bayes factor (EPBF) as the model comparison method in Chapter 4. It is an extension of the application of the extended CPO criterion or EPBF for the linear regression models discussed in Chapter 3 to the generalized linear models. Two logistic regression models are compared at 11 different "leave-out sizes". Without any closed form results, we only rely on the MCMC based results to compute the EPBF for model comparison. Model comparison results are computed for each 10 equal-sized batch of MCMC

samples. Two logistic regression models: Model 1 and Model 2 are considered with five and four explanatory variables respectively.

The MCMC samples based results obtained from 10 batches show a greater variability from their mean result with increasing "leave-out sizes". The standard error of the estimated log EPBFs gets bigger with increasing "leave-out sizes", and this poses the requirement of more MCMC samples to obtain a less variable result. Error in the decision of model selection from the available only MCMC based results can be reduced by increasing the number of MCMC samples.

This chapter focuses on Bayesian model comparisons for generalized linear regression models, particularly logistic regression models using extended CPO or EPBF as a model selection method. Motivated by the findings in Chapter 3 for linear regression models, we generalize the findings here to select generalized linear models using EPBF as a model selection method. We hope that the EPBF can be applied for comparing other types of models where no closed form results are available.

Chapter 5

Discussions and Conclusions

We try to document our overall findings comparatively in this Chapter. Also, we state some possible further investigations.

5.1 Discussions

We have examined how a model selection method, the extended conditional predictive ordinate (CPO) criterion or the extended pseudo Bayes factor (EPBF), behaves for linear regression models and generalized linear regression models, with illustrative examples in Chapter 3 and Chapter 4 respectively. Two real life data sets were used in the illustrative examples. We discuss the overall findings here.

We have closed form expressions for the posterior distributions and the predictive distributions for the linear regression models in the Bayesian setting with a conjugate prior setup. This empowered us to compare the model selection results obtained from the closed form with the results obtained using the Monte Carlo (MC) samples from the posterior distribution of the linear regression model parameters (as if no closed form expressions are available). Two different situations were considered for linear regression models: variance of the error is (i) known and (ii) unknown. In the second situation, there was an unknown scale parameter in addition to the unknown location parameters (regression coefficients).

We considered three linear regression models to be compared that produce three pairwise model comparisons. We computed the extended CPO criterion

for the individual models using the closed form and the MC samples setting for both the situations. The posterior MC samples were divided into 10 equal-sized batches to examine the deviation of the results from the closed form results for those batches. For known variance situation, the MC based results matched the corresponding closed form results up to a large "leave-out size" ('leave-out 120" for all three models with 195 data points), and then started to deviate from the closed form results at larger "leave-out sizes", especially at close to "leave-out all". For unknown variance situation, the MC based results matched the corresponding closed form results up to "leave-out 50" for three models. Hence, it can be said that the MC error is getting bigger at the smaller "leave-out sizes" for the unknown variance situation than the known variance situation. Empirical results from the considered data suggest that we can use the "leave-out size" considerably larger than one without too much concern for either situation while using the cross-validation approach, even when we do not have the closed form results.
However, this result is not general as we have evidence from one data set only. If we observe the similar pattern from other data, then the generalization can be established. Also when working with MC samples, based on the evidence, it is recommended to avoid the "leave-out sizes" close to "leave-out all" for high deviation from the closed form results due to MC error.

We have observed a decrease in the log extended CPO value for the closed form results at larger "leave-out sizes". The result obtained for "leave-out all" is different than the "leave-out 1" with a decreasing pattern for increasing "leave-out sizes". Also, the direction of the MC error can be examined from the results. The MC based results posed a substantive positive departure from the corresponding closed form results for all three models at larger "leave-out sizes" indicating a trend of upward bias with increasing "leave-out sizes" close to "leave-out all". Both the situations exhibited this pattern, and the MC error was a bit worse for the unknown variance situation than the known variance situation.

For logistic regression models, we relied on the MCMC samples from the posterior to compute the extended CPO criterion for the individual models as we did not have the closed form expressions for the posterior distribution of the unknown parameters. We have used the mean result of the 10 batches as a proxy of the closed form results for both the models considered. As expected, the extended CPO crite-

rion computed from 10 batches of MCMC samples increasingly deviate from their mean result for individual models with increasing "leave-out sizes". The deviations are pretty small at smaller "leave-out sizes". So, we have the similar overall findings for the simple linear regression models (simple models) and the generalized linear models (non-simple models) regarding the cross-validation approach with different "leave-out sizes". In both cases, we can use "leave-out size" greater than one with a small increase in the MC errors.

Next, we focus on the winning models that come through model comparison using the extended pseudo Bayes factor (EPBF) as a model selection method for both the linear and logistic regression models. For linear regression models, we had three pairwise model comparisons for the three considered models. For known variance situation, the computed log EPBs in close form setting confirmed the comparatively smaller models regarding the parameters as the winning models in those three pairwise comparisons. However, this was not the case for the unknown variance situation where the winning model changed for two pairwise comparisons. The comparatively smaller model was the winning model in the other model comparison. For the model comparisons with the differing winning model, the comparatively large model win up to a "leave-out size", and after that small model starts to win and keep the trend in the same direction up to "leave-out all".

We can conclude two important findings from these results. At first, a general trend is observed in both the situations: the strength of the evidence for the smaller model is increasing with increasing "leave-out sizes". Also, in the presence of the closed form results, we have the computed optimal Bayesian model comparison or the Bayes factor value which is nothing but the EPBF at "leave-out all". Now, we can safely say that the winning models are different for the EPBF at "leave-out 1" and "leave-out all", at least for our model comparisons. So, the so-called approximation of the formal Bayes factor, the PBF has a different winning model here compared to the one found through the computed formal Bayes factor (EPBF at "leave-out all"). Hence, the importance of using "leave-out sizes" other than one is observed to get a closer optimal Bayesian model comparison.

For logistic regression models, the strength of the evidence for the big model increased with the increasing "leave-out sizes". As the closed form results of the linear regression models, the mean log EPBF value of the 10 MCMC batches

showed a particular pattern with the increasing "leave-out sizes". We can report the general finding from both the cases as a trend of selecting a model over increasing "leave-out sizes" which sometimes lead to change in the winning model for the comparison of two close models.

Finally, we want to summarize the computations from the posterior samples when the closed form expressions of the posteriors are unavailable. The summaries will focus on the possible error reduction and guide us to find a "leave-out size" that produces the closest possible formal Bayesian model comparison with a smaller increase in the MC error.

For the linear regression model comparisons, we formulated a mathematical expression to find the sample size required to obtain a specific amount of improvement by reducing MC errors. We used root mean squared error (RMSE) to measure the MC errors for the linear regression models in the Bayesian setting. For example, for the known variance situation, 2500 MC samples ensured that the MC samples based result would have less than 25% error in selecting the winning model up to "leave-out 30". Increasing the MC sample size with the exact formula given in Chapter 3 would allow us to lower the error percentage in model selection as well as the consideration of the larger "leave-out sizes".

For the unknown variance situation, we can use the same formulation but keeping in mind that a reasonably large MC sample is required in the presence of the unknown variance parameter. For the logistic regression models, no such exact formulation is available as we do not have the closed form posteriors. We computed the standard error of the estimated log EPBFs at different "leave-out sizes" to measure how much MC error prone the MCMC samples based model selections were. High variability (due to high MC errors) was observed at larger "leave-out sizes" implying the use of more MCMC samples to obtain a less variable model selection result for the larger "leave-out sizes". Also, being a dependent sample, an MCMC sample of a large size is required for the logistic regression models compared to the size of the independent MC samples used in the linear regression model comparisons. A general take-away message from both the linear and logistic regression models is to use a reasonably large sample from the posterior for the cross-validation approach with a desired 'leave-out size" so that one can have the model comparison result close to the formal Bayesian model comparison which is

mathematically optimal.

We also attempt to find the sources of variation for the MC errors. We considered 10 different splits of 2000 combinations of MC or MCMC samples for the computation of the extended CPO or the EPBF at different "leave-out sizes". Then, we have two sources of variation for the MC error. Some error might come due to the batches, and the random splits might contribute to some error. For both the linear and logistic regression models, we computed the contribution of these two sources in the overall MC error at different "leave-out sizes". We found that both the sources have more or less similar contribution to the overall MC error at all "leave-out sizes" without no clear pattern. So, neither the batches nor the splits are the dominating source of the MC errors with the current setup.

5.2 Further Scope

In this study, we consider 10 equal-sized batches for all MC and MCMC based results. Also, 2000 combinations of these posterior samples are considered in the calculation of the extended CPO or the EPBF for different "leave-out sizes" when the total number of combinations exceed 2000. One further extension of this study might be to examine the effect of the number of the batches with equal and unequal (decreasing/increasing) sizes as well as the number of combinations on the model comparisons. A set of some combinations can be examined for this purpose. Examining the decomposition of the overall MC error for the variable number of batches and number of combinations might be a good idea to explain the sources of the MC error in detail.

We use two different data sets for constructing linear and logistic regression models. The model comparison results utilize those data and consider different "leave-out sizes" accordingly. However, it might be a good idea to look for the relative "leave-out sizes" of a specific data and try to generalize findings for the relative "leave-out sizes" with an application on data with the number of data points from very small to very big. For example, "leave-out 5" for a data set with data points 100 is very different than the "leave-out 5" with data points 1000 regarding the percentage of "leave-out sizes". We expect the MC/MCMC based model comparison results for these will be different too. Thus, the relative "leave-out sizes"

can be examined for different types of simple and non-simple models with varying data points as a potential further scope of this study.

Only logistic regression models are considered as an example of non-simple models. One can use this methodology for comparing other different non-simple models such as nonlinear models, mixture models, hierarchical models, etc. Also, model selection is very sensitive and important for the causal inference problems. That might be a good application to utilize the extended CPO or the EPBF as a model selection method in causal inference problems to select a useful model.

5.3 Conclusions

The Bayes factor is the desired model selection method in the Bayesian setting because it is optimal. Due to computational issue for the non-simple models, in particular, with no closed form posterior, some other predictive approaches say cross-validation approach with "leave-out 1", popularly known as pseudo Bayes factor is used as an approximation. Mathematically, the Bayes factor can be computed from the cross-validation approach with "leave-out all". However, then the Bayes factor is computed with more MC error at "leave-out all" than the "leave-out 1". In this study, we set our objective to find a "leave-out size" that produces a closer Bayesian model comparison with a small increase in the MC error for the non-simple models where the closed form results are unavailable.

We start with the comparison of simple models, say the linear regression models. Since the closed form results are available, we compare those with the MC samples based results (hypothetically assuming unavailability of the closed forms). MC samples based results produce a good agreement with the corresponding closed form results implying the usability of the MC samples in the absence of the closed forms. We observe that the winning model changes for different "leave-out sizes" with a pattern for some pairwise linear regression model comparisons; this puts a contradiction of using cross-validation approach with "leave-out 1" as it has a different winning model compared to "leave-out all". However, some "leave-out sizes" that are greater than one can produce a closer Bayesian model comparison possibly with some MC error. We can reduce this MC error by increasing the size of the MC samples.

For the logistic regression models, the MCMC samples based results are reliable as per the behavior of the MC samples as a proxy of the closed forms in the linear regression models. Increasing sample size will decrease the MC error for the logistic regression models too. The use of the cross-validation approach with "leave-out sizes" more than one produces closer Bayesian model comparison which is evident from both the linear and logistic regression models. Different other non-simple models can be examined to generalize our findings from this study.

Bibliography

Aitkin, M. (1991). Posterior bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–142. → pages 5, 11, 12

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akademinai Kiado. → pages 5

Banerjee, P. N. B. S. (2008). Bayesian linear model: Gory details. *Dowloaded from http://www. biostat. umn. edu/˜ ph7440*. → pages 37, 38

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 145–159. → pages 56

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media. → pages 6

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122. → pages 9, 13

Berger, J. O. and Pericchi, L. R. (1998). Accurate and stable bayesian model selection: The median intrinsic bayes factor. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 1–18. → pages 10

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430. → pages 11

Devroye, L. (1986). Introduction. In *Non-Uniform Random Variate Generation*, pages 1–26. Springer. → pages 57

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328. → pages 11, 13

Geisser, S. (1980). Discussion on sampling and bayes' inference in scientific modeling and robustness (by gep box). *Journal of the Royal Statistical Society A*, 143:416–417. → pages 13

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160. → pages 1, 13, 14

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514. → pages 1, 6, 12, 52

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document. → pages 6

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383. → pages 60, 61

Greenland, S. (2006). Bayesian perspectives for epidemiological research: I. foundations and basic methods. *International journal of Epidemiology*, 35(3):765–775. → pages 61

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195. → pages 6

Jeffreys, H. (1961). Theory of probability. → pages 7

Kadane, J. B. and Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290. → pages 6

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(431):928–934. → pages 41

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 247–262. → pages 9

Lichman, M. (2013). UCI machine learning repository. → pages 25

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):187–192. → pages 8

Lindley, D. V. (1997). Some comments on bayes factors. *Journal of Statistical Planning and Inference*, 61(1):181–189. → pages 9

Nelder, J. A. and Baker, R. J. (1972). Generalized linear models. *Encyclopedia of statistical sciences*. → pages 6

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society Series B. v56*, pages 1–48. → pages 16

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–138. → pages 10

O'Hagan, A. (1997). Properties of intrinsic and fractional bayes factors. *Test*, 6(1):101–118. → pages 10

Pena, D. and Tiao, G. (1992). Bayesian robustness functions for linear models. jm bernardo, jo berger, ap dawid and afm smith. → pages 12

Pettit, L. and Young, K. (1990). Measuring the effect of observations on bayes factors. *Biometrika*, 77(3):455–466. → pages 7

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349. → pages 56, 57

Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266. → pages 60

Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. (2006). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. → pages 16, 17

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. → pages 5

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, pages 147–164. → pages 6

Smith, A. F. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 213–220. → pages 8

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society. Series B (Methodological)*, pages 111–147. → pages 11, 13

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. → pages 58

Zenllner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essaya in Honor of Bruno de Finetti. Amsterdam: North-Hollar*. → pages 42