Model Selection via Bayesian Evidence

Sheldon Fivaz

April 2024

Plagiarism Declaration

This piece of work is a result of my own work and I have complied with the Department's guidance on multiple submission and on the use of AI tools. Material from the work of others not involved in the project has been acknowledged, quotations and paraphrases suitably indicated, and all uses of AI tools have been declared.

Sheldon Fivaz
Date: April 26, 2024

Acknowledgments

I would like to express my gratitude to my project supervisor, Dr K.Perrakis, whose guidance and support throughout the process made this endeavour an immensely rewarding experience.

Contents

Chapter 1

Introduction

The field of Bayesian statistics holds significant importance as it allows for the incorporation of prior beliefs into the statistical framework. This makes it possible to include existing knowledge and expertise into the analysis, leading to more accurate and informed decision-making. Essentially, Bayesian statistics treats any unknown quantity as a random variable, where we can assign prior probabilities based on the degree of belief or previous knowledge about certain characteristics of the quantities. Naturally, this field of statistics extends to a variety of problems. One of its applications is model selection, where we may have a collection of models and must discern which one is the most adequate for explaining the generating process of some data. Since we reflect our prior knowledge about each model and treat the unknown model parameters probabilistically, the Bayesian statistical framework integrates seamlessly with model selection.

At the heart of Bayesian model selection is a quantity called the model evidence which enables direct comparison between competing models through Bayes factors. However, it is common for the model evidence to be analytically intractable which poses a computational problem in and of itself. The challenge of estimating the model evidence has led to the introduction of a diverse pool of methodologies and approaches. The foundation of these methods often requires samples from either the prior distribution, the posterior distribution, or in some cases, an introduced proposal distribution. We must resort to Markov chain Monte Carlo (MCMC) methods when sampling from the posterior distribution in most scenarios.

In Chapter 2, we will formally introduce Bayesian statistics and its applications to model selection, as well as how the model evidence fits into the problem. In Chapter 3, we discuss the strengths and weaknesses of a broad selection of model evidence estimation methods. We initially illustrate the implementation of the selected model evidence estimators using a Gamma-Poisson conjugate model in Chapter 4, where the model evidence is analytically tractable. Therefore, we can compare the results of the model evidence estimators against the true model evidence values and accurately assess their performance. Subsequently, in Chapter 5, we delve into a logistic regression example

that comprises a multivariate parameter space. We will create a set of models for this example, with one true model being used to simulate the data. We can then use the model evidence estimators to approximate the model evidence for each model and then proceed with model selection, assessing whether the estimators will correctly select the true model out of the set. In Chapter 6, we will touch on some potential issues that may arise when dealing with high-dimensional data, where it is essential to perform dimensionality reduction before proceeding with model selection. In Chapter 7, we will draw some final conclusions, highlighting the key takeaways and outlining the future outlook of Bayesian model selection.

Chapter 2

Bayesian Model Selection

2.1 An Introduction to Bayesian Model Selection

In statistics, we often use a range of techniques and methods to learn more about the data we are analysing. In contrast to Frequentist statistics, the Bayesian interpretation of statistics treats all unknown variables as random, allowing for the use of probability to quantify the degree of belief one has for an event to occur. This field of statistics stems from a result known as Bayes's Theorem Bayes (1763). The theorem is given as follows:

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

where $p(A|B)$ is the probability of event A occurring conditional on event B, $p(B|A)$ is the probability of event B occurring conditional on event A, $p(A)$ is the probability of event A occurring, and $p(B)$ is the probability of event B occurring.

To apply Bayes's Theorem in a statistical setting, we fist suppose that we have data $\mathbf{y} = (y_1, \ldots, y_n)^T$ which we assume follows a given model $\mathcal{M}$, with corresponding parameter vector $\boldsymbol{\theta}$. We then define the prior, $p(\boldsymbol{\theta})$, as the probability distribution summarising our initial beliefs about $\boldsymbol{\theta}$; the likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, as the probability of observing the data $\mathbf{y}$ under model $\mathcal{M}$ and vector parameter $\boldsymbol{\theta}$; and finally, the posterior, $p(\boldsymbol{\theta}|\mathbf{y})$, as the probability distribution that expresses the updated beliefs about the parameter vector $\boldsymbol{\theta}$ after having seen the data $\mathbf{y}$. Bayes's Theorem then provides the platform to link these quantities together as follows:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

Noting that the denominator of the right hand side is constant, we simply have:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Moreover, by sampling from the posterior, we can make inferences about the vector parameter $\boldsymbol{\theta}$ within the framework of the model $\mathcal{M}$, such as giving point estimations or

credible intervals, which quantify any uncertainty beliefs about $\boldsymbol{\theta}$.

However, although model $\mathcal{M}$ may be a viable candidate for describing the data, assuming that it is the sole adequate model is a large assumption and might lead to some unsatisfactory results. Instead, we could adopt a more nuanced approach, by exploring a variety of potentially better-fitting models that may provide a more accurate representation of the underlying process generating the data $\mathbf{y}$. This leads to a common problem that arises in statistics known as model selection - where we have a collection of potentially adequate models $\boldsymbol{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$, each with their own parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$, and we need to select which model is the most likely to have produced the data, $\mathbf{y} = (y_1, \ldots, y_n)^T$. Resorting to Bayes's Theorem again, we obtain the result:

$$p(\mathcal{M}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_i)p(\mathcal{M}_i)}{\sum_{\ell=1}^{k} p(\mathbf{y}|\mathcal{M}_\ell)\pi(\mathcal{M}_\ell)} \qquad (\mathcal{M}_i \in \boldsymbol{M}) \tag{1}$$

where $p(\mathcal{M}_i)$ is the prior probability of model $\mathcal{M}_i$, $p(\mathcal{M}_i|\mathbf{y})$ is the posterior probability of model $\mathcal{M}_i$, and $p(\mathbf{y}|\mathcal{M}_i)$ is the evidence of model $\mathcal{M}_i$.

The posterior probabilities are particularly useful for comparing the feasibility of two models given the data. Using (1), we can obtain the ratio of posterior model probabilities of any two models $\mathcal{M}_i,\ \mathcal{M}_j \in \boldsymbol{M}$, with the sums of the denominator cancelling out with each other:

$$\frac{p(\mathcal{M}_i|\mathbf{y})}{p(\mathcal{M}_j|\mathbf{y})} = \frac{\pi(\mathbf{y}|\mathcal{M}_i)}{p(\mathbf{y}|\mathcal{M}_j)} \times \frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}$$

The first factor of the right hand side is known as the Bayes factor, BF_{ij}, and is crucial for calculating the posterior model probabilities and discern which model is more likely to have generated the data. Particularly, if we believe the prior model probabilities are equal, then the Bayes factor will become equal to the ratio of posterior model probabilities.

2.2 The Bayesian Evidence Computation Problem

Finding the Bayes factor involves some computational challenges. Specifically, to obtain the Bayes factor we need the model evidence $p(\mathbf{y}|\mathcal{M})$ - a term which goes under various names in the literature, such as the prior predictive distribution, the marginal likelihood, and more recently the Bayesian evidence. The Bayesian evidence or model evidence seems to be more relevant for the model selection discussion, so we will mainly use this throughout. If $\boldsymbol{\theta}_i = (\theta_1, \ldots, \theta_d)$ are the parameters of model $\mathcal{M}_i$, then the model evidence is given by the following integral,

$$p(\mathbf{y}|\mathcal{M}_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)p(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i$$

In some situations, we have that the posterior and prior distributions belong to the same family of distributions. In such a case, the prior is known as a conjugate prior and we will be able to derive a closed form for our posterior and hence reach the model evidence. Similarly, in non-conjugate cases where the dimensionality of the parameter vector is small i.e 1 or 2 dimensions, we may be able to approximate the defining integral of the model evidence through numerical methods alone. However, outside of these specific cases, the model evidence is often analytically intractable and difficult to approximate, which presents a computational problem. In Chapter 3 we introduce a selection of methods which can be used to overcome this computational problem.

2.3 Interpreting the Bayes Factor

We have seen in the previous Section 2.2, that the Bayes factor is central to the model selection problem. However, we need to be able to interpret it to use it in a practical sense. Jeffreys (1961) gives a valid interpretation that has been widely accepted since its first introduction. We can summarise these interpretations in Table 2.1, which also provides the interpretations on the logarithmic scale. Here, the Bayes factor BF_{ij} corresponds to two models $\mathcal{M}_i$, $\mathcal{M}_j \in \boldsymbol{M}$ which we are comparing against one another.

$\mathbf{BF_{ij}}$	$\mathbf{\log(BF_{ij})}$	**Interpretation**
$\mathbf{> 100}$	$\mathbf{> 4.61}$	**Decisive evidence for $\mathcal{M}_i$**
$\mathbf{30 - 100}$	$\mathbf{3.40 - 0.61}$	**Very strong evidence for $\mathcal{M}_i$**
$\mathbf{10 - 30}$	$\mathbf{2.30 - 3.40}$	**Strong evidence for $\mathcal{M}_i$**
$\mathbf{3 - 10}$	$\mathbf{1.10 - 2.30}$	**Substantial evidence for $\mathcal{M}_i$**
$\mathbf{1 - 3}$	$\mathbf{0 - 1.10}$	**Not worth more than a bare mention**
$\mathbf{1/3 - 1}$	$\mathbf{(-1.10) - 0}$	**Not worth more than a bare mention**
$\mathbf{1/10 - 1/3}$	$\mathbf{(-2.30) - (-1.10)}$	**Substantial evidence for $\mathcal{M}_j$**
$\mathbf{1/30 - 1/10}$	$\mathbf{(-3.40) - (-2.30)}$	**Strong evidence for $\mathcal{M}_j$**
$\mathbf{1/100 - 1/30}$	$\mathbf{(-4.61) - (-3.40)}$	**Very strong evidence for $\mathcal{M}_j$**
$\mathbf{< 1/100}$	$\mathbf{< (-4.61)}$	**Decisive evidence for $\mathcal{M}_j$**

Table 2.1: Table representing the interpretations of the Bayes factor given in Jeffreys (1961), taken and edited from Dittrich et al. (2019)

Alternatively, there is an interpretation of the Bayes factor proposed by Kass & Raftery (1995) that is more commonly used today. In this interpretation, we work with the values of twice the natural logarithm of the Bayes factor which is on the same scale as deviance and likelihood ratio test statistics. These interpretations can be found in Table 2.2

$2\log(\mathbf{BF_{ij}})$	$\mathbf{BF_{ij}}$	Evidence against $\mathcal{M}_j$
$\mathbf{0-2}$	$\mathbf{1-3}$	**Not worth more than a bare mention**
$\mathbf{2-6}$	$\mathbf{3-20}$	**Positive**
$\mathbf{6-10}$	$\mathbf{20-150}$	**Strong**
$\mathbf{>10}$	$\mathbf{>150}$	**Very strong**

Table 2.2: Table representing the interpretations of twice the logarithm of the Bayes factor, taken and edited from Kass & Raftery (1995)

For the purpose of our model selection examples, the interpretations from Table 2.1 will suffice, and will be used in later chapters.

Chapter 3

Model Evidence Estimation Methods

In this section, we will be introducing the model evidence estimation methods which we proceed to investigate in the remaining chapters. For the mathematical descriptions, we will use the simpler notation, $p(\mathbf{y})$, in place of $p(\mathbf{y}|\mathcal{M}_i)$.

3.1 Basic Monte Carlo Estimator: Sampling from the Prior

Arguably the simplest method of estimating the Bayesian Evidence is the basic Monte Carlo estimator Hammersley & Handscomb (1964), which uses samples from the prior distribution to estimate the model evidence. Although this method is very simple to use, as it does not require a posterior sample, its accuracy can be quite sensitive to the form of the prior distribution.

The method uses the basic integral form of the model evidence that we saw when introducing it in Section 2.2:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Note that the marginal likelihood here is essentially the expectation with respect to the prior distribution, which gives the following form:

$$p(\mathbf{y}) = \mathbb{E}_{P(\boldsymbol{\theta})}\left[p(\mathbf{y}|\boldsymbol{\theta})\right]$$

Using the standard Monte Carlo technique, we then approximate this expectation as a sum over samples from the prior distribution. This leads to the basic Monte Carlo estimator:

$$\hat{p}_{MC} = \frac{1}{N}\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\theta}^{(t)})$$

where $\boldsymbol{\theta}^{(t)}$ are draws from the prior distribution.

Although this is a very simple estimator, it is highlighted by Kass & Raftery (1995) that the estimator does not perform well when the posterior is concentrated relative to the prior. This occurs, for example, when the prior is diffuse and non-informative. However, if we have very weak prior beliefs about the parameters of interest, then it is natural to opt for a prior of this form. The problem that arises in these cases is that most of the prior samples, $\boldsymbol{\theta}^{(t)}$, will carry small likelihood values, $p(\mathbf{y}|\boldsymbol{\theta}^{(t)})$, causing the final estimate to be dominated by the small number of large likelihood values.

3.2 Harmonic Mean Estimator

The harmonic mean estimator Newton & Raftery (1994) is a technique which works on the reciprocal of the model evidence. Unlike the basic Monte Carlo estimator, the harmonic mean estimator uses draws from the posterior distribution to arrive at the estimate. We start with the following identity, which is trivially true due to the normalisation condition of probability density functions:

$$1 = \int p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Now, by dividing each side by the model evidence and introducing the factor $\frac{p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})}$ inside the integral of the right hand side, we obtain the following two lines:

$$\begin{aligned}\frac{1}{p(\mathbf{y})} &= \int \frac{p(\boldsymbol{\theta})}{p(\mathbf{y})}\, d\boldsymbol{\theta} \\ &= \int \frac{1}{p(\mathbf{y}|\boldsymbol{\theta})} \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}\, d\boldsymbol{\theta}\end{aligned}$$

By replacing $\frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}$ with the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$, we can simplify the integral. Crucially, this makes its possible to write the reciprocal of the model evidence as an expectation with respect to the posterior distribution:

$$\begin{aligned}\frac{1}{p(\mathbf{y})} &= \int \frac{1}{p(\mathbf{y}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} \\ &= \mathbb{E}_{P(\cdot|\mathbf{y})}\left[\frac{1}{p(\mathbf{y}|\boldsymbol{\theta})}\right]\end{aligned}$$

By again using the standard Monte Carlo technique, we approximate this expectation as a sum, this time over samples from the posterior distribution. This leads to the harmonic mean estimator:

$$\hat{p}(\mathbf{y}) = \left(\frac{1}{N}\sum_{t=1}^{N} \frac{1}{p(\mathbf{y}|\boldsymbol{\theta}^{(t)})}\right)^{-1}$$

where $\boldsymbol{\theta}^{(t)}$ are draws from the prior distribution.

We can see here that the final estimate gives the harmonic mean of the likelihoods, based on the posterior sample. A problem with the harmonic mean estimator is demonstrated in Kass & Raftery (1995), which is almost the opposite of the issue with the basic Monte Carlo estimator. With this estimator, we can see that the likelihood values appear on the denominator of the summed terms. Therefore, a small number of posterior samples which carry small likelihood values will dominate the sum and disproportionately affect the final estimate. However, despite its instability, the harmonic mean estimator does have the property that $\hat{p}(\mathbf{y})$ converges almost surely to $p(\mathbf{y})$ as $N \to \infty$.

3.3 Generalised Harmonic Mean Estimator

The generalised harmonic mean estimator provides a similar solution to the harmonic mean estimator, except now we use a proposal density, $g(\boldsymbol{\theta})$, to obtain the estimate. The first two lines use the same tricks as in the derivation of the harmonic mean estimator, this time replacing $p(\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})$ with $p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ on the denominator:

$$\begin{aligned}\frac{1}{p(\mathbf{y})} &= \int \frac{1}{p(\mathbf{y})} g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int \frac{g(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}\end{aligned}$$

We again write the reciprocal of the model evidence as an expectation, but with respect to the proposal distribution:

$$\frac{1}{p(\mathbf{y})} = \mathbb{E}_{P(\cdot|\mathbf{y})}\left[\frac{g(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\right]$$

Applying standard Monte Carlo technique leads to the generalised harmonic mean estimator:

$$\hat{p}(\mathbf{y}) = \left(\frac{1}{N}\sum_{t=1}^{N} \frac{g(\boldsymbol{\theta}^{(t)})}{p(\mathbf{y}|\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})}\right)^{-1}$$

where $\boldsymbol{\theta}^{(t)}$ are draws from the posterior distribution.

Note that here we are still using the posterior sample for the estimate and no draws are needed from the introduced proposal density. It is suggested in Robert & Wraith (2009) that the proposal density, $g(\boldsymbol{\theta})$, must have thinner tails than the posterior distribution. The purpose of this is to avoid the lower likelihood values dominating the sum, as is the problem in the case of the harmonic mean estimator. Specifically, it is stated in Newton & Raftery (1994) that $\int \frac{g(\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\, d\boldsymbol{\theta} < \infty$ will ensure that the estimator is well behaved. One approach to this is to extract the mean and variance from the posterior sample and use the method of moments to construct a Gaussian proposal density from the posterior sample.

3.4 Newton-Raftery Estimator

The identity used in the generalised harmonic mean estimator has also been referred to as the "reciprocal importance sampling" estimator Frühwirth-Schnatter (2004) due to its fundamental identity. We now introduce the standard self-normalised importance sampling estimator of the model evidence, which begins with the basic integral form of the model evidence:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

which can be written as an expectation with respect to the likelihood:

$$\mathbb{E}_{P(\boldsymbol{\theta})}\left[p(\mathbf{y}|\boldsymbol{\theta})\right]$$

as we saw in Section 3.1 for the basic Monte Carlo estimator. Now, given a possibly un-normalised probability density function $g(\boldsymbol{\theta})$, self-normmalised importance sampling provides an approximation of this expectation of the form:

$$\hat{p}(\mathbf{y}) = \frac{\sum_{t=1}^{N} w_t \; p(\mathbf{y}|\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^{N} w_t} \tag{1}$$

where $w_t = \frac{p(\boldsymbol{\theta}^{(t)})}{g(\boldsymbol{\theta}^{(t)})}$ are the weights of self-normalised importance sampling and $\boldsymbol{\theta}^{(t)}$ are draws from the proposal distribution.

We discussed that the basic Monte Carlo estimator was susceptible to large likelihood values influencing the estimate, whereas the Harmonic Mean estimator struggled with small likelihood values disproportionately affecting the estimate. The Newton-Raftery estimator Newton & Raftery (1994) aims to fix these instability issues through the self-normalised importance sampling estimator, with a proposal density that uses a mixture of the prior and posterior density functions:

$$g(\boldsymbol{\theta}) = \delta p(\boldsymbol{\theta}) + (1-\delta)p(\boldsymbol{\theta}|\mathbf{y}) \qquad\qquad (0 < \delta < 1)$$

This eventually leads to an iterative equation:

$$\hat{p}(\mathbf{y}) = \frac{\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}}{\sum_{t=1}^{N}\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}}$$

where $\boldsymbol{\theta}^{(t)}$ are draws from the proposal distribution.

We can derive this result through (1). First, we must calculate the weights of the self-normalised importance sampling estimator, starting with:

$$w_t = \frac{p(\boldsymbol{\theta}^{(t)})}{g(\boldsymbol{\theta}^{(t)})} = \frac{p(\boldsymbol{\theta}^{(t)})}{\delta p(\boldsymbol{\theta}^{(t)}) + (1-\delta)p(\boldsymbol{\theta}^{(t)}|\mathbf{y})}$$

Replacing $p(\boldsymbol{\theta}^{(t)})$ with $\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})\,p(\boldsymbol{\theta}^{(t)})}{p(\mathbf{y})}$ and then multiplying through by $p(\mathbf{y})$ leads to:

$$\begin{aligned} w_t &= \frac{p(\boldsymbol{\theta}^{(t)})}{\delta p(\boldsymbol{\theta}^{(t)}) + (1-\delta)\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})\,p(\boldsymbol{\theta}^{(t)})}{p(\mathbf{y})}} \\ &= \frac{1}{\delta + (1-\delta)\frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})}{p(\mathbf{y})}} \\ &= \frac{p(\mathbf{y})}{\delta\, p(\mathbf{y}) + (1-\delta)p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})} \end{aligned}$$

Applying these weights to (1) then gives:

$$\begin{aligned} \hat{p}(\mathbf{y}) &= \frac{\sum_{t=1}^{N} \frac{p(\mathbf{y})}{\delta\, p(\mathbf{y}) + (1-\delta)p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})}\; p(\mathbf{y}|\boldsymbol{\theta}^{(t)})}{\sum_{t=1}^{N} \frac{p(\mathbf{y})}{\delta\, p(\mathbf{y}) + (1-\delta)p(\boldsymbol{y}|\boldsymbol{\theta}^{(t)})}} \\ &= \frac{\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\{\delta p(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}}{\sum_{t=1}^{N} \{\delta p(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}} \end{aligned}$$

We obtain the final iterative equation by replacing $p(\mathbf{y})$ with $\hat{p}(\mathbf{y})$ on the right hand side

In practice, we can use the estimator by initialising $\hat{p}(\mathbf{y})$ at $\hat{p}_0$ and then iterating the equation n times to obtain $\hat{p}_1, \ldots, \hat{p}_n$ where $\hat{p}_n$ is the final estimate.

3.5 Laplace-Metropolis Estimator

The Laplace method Tierney & Kadane (1986) is an approach which can be used to estimate the model evidence if we assume that the posterior distribution can be approximated by a Gaussian distribution. The resulting approximation is:

$$\hat{p}(\mathbf{y}) = (2\pi)^{d/2}|\tilde{\boldsymbol{\Sigma}}|^{1/2} p(\mathbf{y}|\tilde{\boldsymbol{\theta}})p(\tilde{\boldsymbol{\theta}})$$

where d is the dimension of $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$ is the posterior mode and $\tilde{\boldsymbol{\Sigma}}$ is the negative inverse Hessian matrix at $\tilde{\boldsymbol{\theta}}$.

However, in many practical cases, both the inverse Hessian matrix $\tilde{\boldsymbol{\Sigma}}$ and the posterior mode $\tilde{\boldsymbol{\theta}}$ are not analytically tractable. The Laplace-Metropolis estimator Lewis & Raftery (1997) aims to bridge this gap by using a posterior sample to provide a variation of the Laplace estimate:

$$\hat{p}(\mathbf{y}) = (2\pi)^{d/2}|\mathbf{S}|^{1/2}p(\mathbf{y}|\boldsymbol{\theta}^{max})p(\boldsymbol{\theta}^{max})$$

$$\boldsymbol{\theta}^{max} = \max_{t=1,\cdots,N}\{p(y|\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})\}$$

where $\mathbf{S}$ is the sample covariance matrix of the posterior sample, used in place of $\tilde{\boldsymbol{\Sigma}}$, and $\boldsymbol{\theta}^{(t)}$ are draws from the posterior distribution.

The justification for using the sample covariance matrix comes from the fact that the posterior variance matrix is asymptotically equal to the inverse Hessian matrix $\tilde{\boldsymbol{\Sigma}}$. We can then provide an unbiased estimate of the posterior variance matrix through the sample covariance matrix. A more advanced technique here could be to use weighted variance matrices from the sample, which is considered more robust by Lewis & Raftery (1997). Similarly, there are other methods discussed for estimating the posterior mode, $\tilde{\boldsymbol{\theta}}$, which include estimating the component-wise entries of $\tilde{\boldsymbol{\theta}}$ through posterior sample means or medians.

3.6 Bridge Sampling Estimator

The bridge sampling technique was originally developed by Meng & Wong (1996) in order to estimate the ratio of the normalising constants, c_1 and c_2, of two densities. We will proceed with a version of bridge sampling from Gronau et al. (2017) and follow the steps laid out in the paper to estimate the model evidence. We will summarise the steps found in the paper in the rest of this section. Here, we start with the identity:

$$1 = \frac{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

where $g(\boldsymbol{\theta})$ is the proposal density and $h(\boldsymbol{\theta})$ is known as the bridge function.

We then multiply both sides of the identity by the model evidence, $p(\mathbf{y})$, to obtain:

$$p(\mathbf{y}) = \frac{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$$

Replacing $\frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$ with $p(\boldsymbol{\theta}|\mathbf{y})$ leads to the following expectations:

$$\begin{aligned} p(\mathbf{y}) &= \frac{\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}|\mathbf{y})h(\boldsymbol{\theta})g(\boldsymbol{\theta})\,d\boldsymbol{\theta}} \\ &= \frac{\mathbb{E}_{G(\cdot)}\left[p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})h(\boldsymbol{\theta})\right]}{\mathbb{E}_{P(\cdot|\mathbf{y})}\left[h(\boldsymbol{\theta})g(\boldsymbol{\theta})\right]} \end{aligned}$$

We can then obtain the estimate by using the standard Monte Carlo technique to approximate the expectations:

$$\hat{p}(\mathbf{y}) = \frac{\frac{1}{N_2}\sum_{t=1}^{N_2} p(\mathbf{y}|\boldsymbol{\theta}_2^{(t)})p(\boldsymbol{\theta}_2^{(t)})h(\boldsymbol{\theta}_2^{(t)})}{\frac{1}{N_1}\sum_{t=1}^{N_1} h(\boldsymbol{\theta}_1^{(t)})g(\boldsymbol{\theta}_1^{(t)})} \tag{2}$$

where N_1 is the size of the proposal sample, N_2 is the size of the posterior sample, $\boldsymbol{\theta}_1^{(t)}$ are draws from the posterior distribution, and $\boldsymbol{\theta}_2^{(t)}$ are draws from the proposal distribution

Note that here we need samples from both the proposal distribution and the posterior distribution for this estimator. For the proposal distribution, we can again use the method of moments to create a Gaussian density which will ideally resemble the posterior density. Aside from that, it is clear that we need to decide on a suitable bridge function to use this estimator in practice. It is suggested in Meng & Wong (1996) that the following bridge function is optimal:

$$h(\boldsymbol{\theta}) = \frac{C}{s_1 p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) + s_2 p(\boldsymbol{y})g(\boldsymbol{\theta})}$$

where $s_1 = \frac{N_1}{N_1+N_2}, s_2 = \frac{N_2}{N_1+N_2}$, and C is an arbitrary constant. Note that the constant C will cancel out when we plug the bridge function into (2), which leads to an iterative equation for the bridge sampling estimate:

$$\hat{p}(\mathbf{y}) = \frac{\frac{1}{N_2}\sum_{t=1}^{N_2} \frac{p(\mathbf{y}|\boldsymbol{\theta}_2^{(t)})p(\boldsymbol{\theta}_2^{(t)})}{s_1 p(\mathbf{y}|\boldsymbol{\theta}_2^{(t)})p(\boldsymbol{\theta}_2^{(t)}) + s_2\hat{p}(\mathbf{y})g(\boldsymbol{\theta}_2^{(t)})}}{\frac{1}{N_1}\sum_{t=1}^{N_1} \frac{g(\boldsymbol{\theta}_1^{(t)})}{s_1 p(\mathbf{y}|\boldsymbol{\theta}_1^{(t)})p(\boldsymbol{\theta}_1^{(t)}) + s_2\hat{p}(\mathbf{y})g(\boldsymbol{\theta}_1^{(t)})}}$$

where again we must initalise $\hat{p}(\mathbf{y})$ at $\hat{p}_0$ and then iterate the equation n times to obtain the final estimate.

3.7 Fourier Integral Estimator: a Newer Method in the Field

In contrary to the previously discussed methods, Rotiroti & Walker (2022) use a unique approach stemming from the central Bayesian equality:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

The idea here is that if we are able to estimate the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ at a single point $\boldsymbol{\theta}^*$ in the parameter space, then we can evaluate the model evidence,

$$p(\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^*|\mathbf{y})}$$

On the logarithmic scale, we can reach the estimate using the adapted formula:

$$\log(p(\mathbf{y})) = \log(p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)) - \log(p(\boldsymbol{\theta}^*|\mathbf{y}))$$

Note here that we only need to estimate the posterior density at a single point $\boldsymbol{\theta}^*$ in order to obtain the estimate. This approach was actually first presented by Besag (1989), where the "candidate's estimator" is demonstrated as a method to estimate the marginal likelihood. The same idea was also used by Chib (1995) and Chib & Jeliazkov (2001), leading to the Chib estimator and Chib and Jeliazkov estimators respectively. These methods are also very effective for estimating the model evidence, however will not be discussed in this report. A comprehensive analysis of the Chib estimator, as well as other methods not discussed in this report such as annealed importance sampling Neal (2001), nested sampling Skilling (2006), and power posteriors Friel & Pettitt (2008) can be found in Friel & Wyse (2012).

We will now discuss how we will obtain the estimate $p(\boldsymbol{\theta}^*|\mathbf{y})$ using a sample from the posterior distribution. We begin by introducing a method called the Fourier transform. Essentially, the Fourier transform works by decomposing a function into sinusoidal waves of various frequencies. We won't delve into specifics here as it is not the focus of the report, but for a function $f(x)$, the Fourier transform is given by:

$$\tilde{f}(s) = \int_{-\infty}^{\infty} f(x)\, \mathrm{e}^{-isx}\, dx \tag{3}$$

By inverting the formula, we obtain the result:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(s)\, \mathrm{e}^{isx}\, dx \tag{4}$$

We can then combine (3) and (4) to derive a double integral solely in terms of $f(x)$:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathrm{e}^{isx}\, \mathrm{e}^{-isy}\, f(y)\, dsdy$$

When $f(x)$ is real, we can replace the exponential factors with trig functions:

$$f(x) = \frac{1}{2\pi} \lim_{R\to\infty} \int_{-\infty}^{\infty} \int_{-R}^{R} \cos(s(x-y))\, ds\, f(y)\, dy$$
$$f(x) = \frac{1}{\pi} \lim_{R\to\infty} \int_{-\infty}^{\infty} \sin(R(x-y))\, f(y)\, dy$$

We can apply this formula to the posterior density, evaluated at a single point θ^* in the parameter space. In the univariate case, this gives the form:

$$p(\theta^*|\mathbf{y}) = \frac{1}{\pi} \lim_{R\to\infty} \int \sin(R(\theta^* - \theta))\, p(\theta|\mathbf{y})\, d\theta$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ now represents the data in our statistical problem. Since we are left with a single integral, we can represent this as an expectation with respect to the posterior distribution and use the standard Monte Carlo trick to approximate this integral, leading to the estimate:

$$p(\theta^*|\mathbf{y}) = \frac{1}{\pi} \lim_{R\to\infty} \mathbb{E}_{P(\cdot|\mathbf{y})} \left[\sin(R(\theta^* - \theta))\right]$$
$$\hat{p}(\theta^*|\mathbf{y}) = \frac{1}{N\pi} \sum_{t=1}^{N} \frac{\sin(R(\theta^* - \theta^{(t)}))}{\theta^* - \theta^{(t)}}$$

for some fixed large R, where $\theta^{(t)}$ are draws from the posterior distribution.

The previous results come from Fourier transforms of univariate functions. We can also apply the Fourier methods to multivariate functions in order to estimate the model evidence to parameter spaces of higher dimensionality. We will omit the details here, but it is possible to arrive at the closed form:

$$f(x) = \frac{1}{\pi^d} \lim_{R\to\infty} \int_{\mathbb{R}^d} \prod_{j=1}^{d} \frac{\sin(R(x_j - x_{ji}))}{x_j - x_{ji}}\, f(y)\, dy$$

Again, we can apply this formula to the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$, evaluated at a single point $\boldsymbol{\theta}^*$ in the now multi-dimensional parameter space:

$$p(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{1}{\pi^d} \lim_{R\to\infty} \int \prod_{j=1}^{d} \frac{\sin(R(\theta_j^* - \theta_j))}{\theta_j^* - \theta_j}\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}$$

By representing this as an expectation and using the Monte-Carlo technique in the multivariate setting, we reach:

$$p(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{1}{\pi} \lim_{R\to\infty} \mathbb{E}_{P(\cdot|\mathbf{y})} \left[\prod_{j=1}^{d} \frac{\sin(R(\theta_j^* - \theta_j))}{\theta_j^* - \theta_j} \right]$$
$$\hat{p}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{1}{N\pi^d} \sum_{t=1}^{N} \prod_{j=1}^{d} \frac{\sin(R(\theta_j^* - \theta_j^{(t)}))}{\theta_j^* - \theta_j^{(t)}}$$

for some fixed large R, where $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_d^{(t)})^T$ are draws from the posterior distribution.

Note that we will refer to this estimation method as the Fourier integral estimator in the remaining chapters.

Chapter 4

Toy Example: The Gamma-Poisson Conjugate Model

4.1 The Gamma-Poisson Conjugate Model

We will first introduce a Gamma-Poisson conjugate model which will provide a situation where we can use the 7 methods to estimate the model evidence. We label this example as the toy example as we are able to directly obtain both the posterior distribution and the marginal likelihood, which enables calculating the error of the estimates. The goal with this model is to test the estimators against a set of known model evidences. We will simulate 50 data sets $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(50)}$ of size n directly from the likelihood. For each data set $\mathbf{y}_{(i)} = (y_{(i)}^{(1)}, \ldots, y_{(i)}^{(n)})$, we can then estimate the model evidence using the 7 methods and compare the estimates to the exact model evidence. The likelihood and prior of the Gamma-Poisson model are given below:

$$\begin{aligned} y^{(j)}|\lambda &\sim \text{Poisson}(\lambda) \qquad (\lambda \in \mathbb{Z}^+) \\ \lambda &\sim \text{Gamma}(\alpha, \beta) \qquad (\alpha, \beta > 0) \end{aligned}$$

After simulating the data, we will immediately be able to derive the posterior distribution due to conjugacy. The posterior will then take the following form:

$$\lambda|\mathbf{y} \sim \text{Gamma}(\alpha + \sum_j y^{(j)}, \beta + n)$$

Moreover, with this simple model we can extract the model evidence by direct integration:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\lambda)\, p(\lambda)\, d\lambda = \frac{\Gamma\left(n\,\overline{\mathbf{y}} + \alpha\right) \beta^{\,\alpha}}{\prod_j y^{(j)}! \, \Gamma(\alpha) (n+\beta)^{n\,\overline{\mathbf{y}} + \alpha}} \tag{1}$$

The model specifications we will use for the remainder of the chapter can be summarised in the following table:

size of data sets:	$n = 50$
posterior/ prior sample size:	$N = 10000$
prior hyper-parameters:	$a = b = 0.1$

We specifically choose the Gamma prior to have shape and rate parameter equal to 0.1 as this gives what can be considered a vague prior. Choosing α and β to have very small values creates a distribution that is heavily skewed towards zero with a long right tail. Therefore, the probability density will be concentrated at zero, with a range of non-zero probabilities for the higher values. Moreover, the mean of the distribution will sit at $\frac{\alpha}{\beta} = \frac{0.1}{0.1} = 1$, whilst being very uncertain about higher values, having a variance of $\frac{\alpha}{\beta^2} = \frac{0.1}{0.1^2} = 10$. Hence, we will be expressing minimal prior beliefs so that the posterior will be largely influenced by the likelihood of the data, reflecting real applications where we may not have much information about the shape of the posterior. A visual comparison of the prior and posterior density can be found in Figure 4.1.

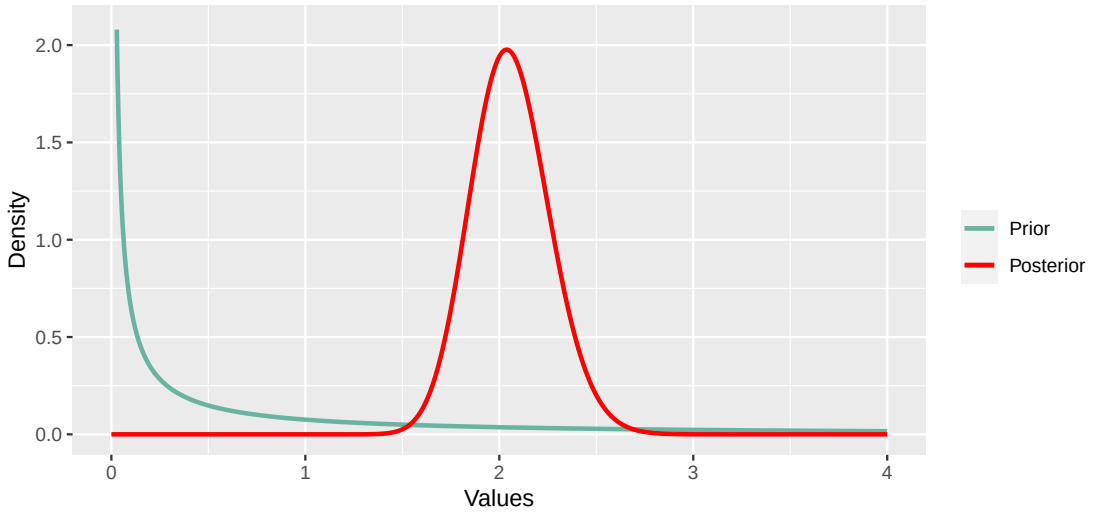

Figure 4.1: Density plot of the posterior and vague prior

4.2 Data Simulation

First we must simulate the data sets. As discussed in Section 4.1, we will directly simulate the data sets $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(50)}$ from Poisson(λ). We will select the value of λ to be 2 for data generation purposes. The data is then summarised in a $50 \times n$ matrix, where

each row corresponds to one of the 50 data sets.

$$Y = \begin{bmatrix} \mathbf{y}_{(1)}^{\mathrm{T}} \\ \mathbf{y}_{(2)}^{\mathrm{T}} \\ \vdots \\ \mathbf{y}_{(50)}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} y_{(1)}^{(1)} & y_{(1)}^{(2)} & \cdots & y_{(1)}^{(n)} \\ y_{(2)}^{(1)} & y_{(2)}^{(2)} & \cdots & y_{(2)}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{(50)}^{(1)} & y_{(50)}^{(2)} & \cdots & y_{(50)}^{(n)} \end{bmatrix}$$

Now that we have determined our 50 data sets, we can calculate the 50 true model evidence values $m_1, \ldots, m_{50}$ on the logarithmic scale using (1). This gives the 50 model evidence values as:

$$p(\mathbf{y}_{(i)}) = \frac{\Gamma\left(n\,\overline{\mathbf{y}}_{(i)} + \alpha\right)\beta^{\,\alpha}}{\prod\limits_{j} y_{(i)}^{(j)}!\,\Gamma(\alpha)(n+\beta)^{n\,\overline{\mathbf{y}}_{(i)}+\alpha}} \qquad (i = 1, \ldots, 50)$$

We will store these values in the vector $\mathbf{m}$:

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{50} \end{bmatrix}$$

4.3 Implementing the Methods

In this section, we will describe how we can implement the estimator methods within the framework of the Gamma-Poisson Conjugate example. For ease of comparison against the true model evidence values, we will be working on the logarithmic scale. Please note that there are likely many ways in which the code used to implement these methods can be optimised for greater efficiency. The code used for this chapter can be found through this link, which allows the reader to change the settings and experiment with the results.

Before proceeding with the methods, we must first generate the prior and posterior samples. For this, we will denote the prior samples as $\boldsymbol{\lambda}_{prior(1)}, \ldots, \boldsymbol{\lambda}_{prior(50)}$ where $\boldsymbol{\lambda}_{prior(i)} = (\lambda_{prior(i)}^{(1)}, \ldots, \lambda_{prior(i)}^{(N)})$ contains N different λ samples. In similar notation, we will denote the posterior samples by $\boldsymbol{\lambda}_{post(1)}, \ldots, \boldsymbol{\lambda}_{post(50)}$ and the proposal samples by $\boldsymbol{\lambda}_{prop(1)}, \ldots, \boldsymbol{\lambda}_{prop(50)}$. We will also summarise the prior, posterior and proposal samples in 3 50 x N matrices, with each row corresponding to a unique sample. The procedures are described below.

4.3.1 Sampling from the Prior

For the prior sample, we will generate this by sampling directly from the prior distribution:

$$\lambda \sim \text{Gamma}(0.1, 0.1)$$

The results of this are given in the matrix Λ_{prior}, taking the form:

$$\Lambda_{prior} = \begin{bmatrix} \boldsymbol{\lambda}^{\mathrm{T}}_{prior\,(1)} \\ \boldsymbol{\lambda}^{\mathrm{T}}_{prior\,(2)} \\ \vdots \\ \boldsymbol{\lambda}^{\mathrm{T}}_{prior\,(50)} \end{bmatrix} = \begin{bmatrix} \lambda^{(1)}_{prior(1)} & \lambda^{(2)}_{prior(1)} & \cdots & \lambda^{(N)}_{prior(1)} \\ \lambda^{(1)}_{prior(2)} & \lambda^{(2)}_{prior(2)} & \cdots & \lambda^{(N)}_{prior(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^{(1)}_{prior(50)} & \lambda^{(2)}_{prior(50)} & \cdots & \lambda^{(N)}_{prior(50)} \end{bmatrix}$$

4.3.2 Sampling from the Posterior

In a similar fashion, we generate the posterior samples by sampling directly from the known posterior distribution. The ith posterior sample will be sampled from the distribution:

$$\lambda^{(j)}_{post(i)} | \mathbf{y}_{(i)} \sim \text{Gamma}(0.01 + \sum_{j} y^{(j)}_{(i)}\,,\; 0.01 + n)$$

The results of this are given in the matrix Λ_{post}, taking the form:

$$\Lambda_{post} = \begin{bmatrix} \boldsymbol{\lambda}^{\mathrm{T}}_{post\,(1)} \\ \boldsymbol{\lambda}^{\mathrm{T}}_{post\,(2)} \\ \vdots \\ \boldsymbol{\lambda}^{\mathrm{T}}_{post\,(50)} \end{bmatrix} = \begin{bmatrix} \lambda^{(1)}_{post(1)} & \lambda^{(2)}_{post(1)} & \cdots & \lambda^{(N)}_{post(1)} \\ \lambda^{(1)}_{post(2)} & \lambda^{(2)}_{post(2)} & \cdots & \lambda^{(N)}_{post(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^{(1)}_{post(50)} & \lambda^{(2)}_{post(50)} & \cdots & \lambda^{(N)}_{post(50)} \end{bmatrix}$$

We discuss the collection of the proposal samples in Section 4.3.8

Now we can proceed with the implementation of our model evidence estimators.

4.3.3 Basic Monte Carlo Estimator

As discussed in Section 3.1, the basic Monte Carlo estimator only relies on a prior sample. We must calculate the likelihood for each of the 50 prior samples generated in Section 4.2 with the corresponding data values $\mathbf{y}$. For example:

$$p(\mathbf{y}_i | \lambda^{(j)}_{prior(i)}) = \prod_{\ell=1}^{n} \text{Poisson}(y_{i,\ell}\,; \lambda^{(j)}_{prior(i)})$$

When utilising a vague prior, it's quite possible to encounter likelihood values so small that they round to zero in R, even when computed on the logarithmic scale. In such cases, we have the decision to round these zero values to the lowest available number in R, $2.225e-308$. However, in practice, we observe that performing this substitution is so insignificant that it will not affect the final estimates. Hence, we will proceed with any zero values for our estimates.

Now, the 50 Monte Carlo estimates are given by:

$$\log \hat{p}_{MC(1)} = \log \left(\frac{1}{N} \sum_{t=1}^{N} p(\mathbf{y}_1 | \lambda^{(t)}_{prior(1)}) \right)$$
$$\vdots$$
$$\log \hat{p}_{MC(50)} = \log \left(\frac{1}{N} \sum_{t=1}^{N} p(\mathbf{y}_{50} | \lambda^{(t)}_{prior(50)}) \right)$$

4.3.4 Harmonic Mean Estimator

The harmonic mean estimator relies on a posterior sample. Similar to the Monte Carlo estimate, we must calculate the likelihood values of the data $\mathbf{y}_1, \ldots, \mathbf{y}_{50}$, except this time with the 50 posterior samples. For example:

$$p(\mathbf{y}_i | \lambda^{(j)}_{post(i)}) = \prod_{\ell=1}^{n} \text{Poisson}(y_{i,\ell}\,; \lambda^{(j)}_{post(i)})$$

Now, the Harmonic Mean estimates can be calculated by taking the log of the formula in Section 3.2:

$$\log \hat{p}(\mathbf{y}) = \log \left(\frac{1}{N} \sum_{t=1}^{N} \frac{1}{p(\mathbf{y} | \lambda^{(t)}_{post})} \right)^{-1} = -\log \left(\frac{1}{N} \sum_{t=1}^{N} \frac{1}{p(\mathbf{y} | \lambda^{(t)}_{post})} \right)$$

The 50 Harmonic Mean estimates are then given by:

$$\log \hat{p}_{HM(1)} = -\log \left(\frac{1}{N} \sum_{t=1}^{N} \frac{1}{p(\mathbf{y}_1 | \lambda^{(t)}_{post(1)})} \right)$$
$$\vdots$$
$$\log \hat{p}_{HM(50)} = -\log \left(\frac{1}{N} \sum_{t=1}^{N} \frac{1}{p(\mathbf{y}_{50} | \lambda^{(t)}_{post(50)})} \right)$$

4.3.5 Generalised Harmonic Mean Estimator

As discussed in Section 3.3, the generalised harmonic mean estimator requires the introduction of a proposal density, ideally having thin tails. For this example, we will use the method of moments to construct 50 Gaussian proposal densities of the form $g(\lambda)$, each corresponding to a posterior sample $\boldsymbol{\lambda}_{post}$. We will first calculate the sample means and sample variances of each posterior sample:

$$\overline{\boldsymbol{\lambda}}_{post(i)} = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{\lambda}^{(t)}_{post(i)} \qquad\qquad s^2_{post(i)} = \frac{1}{N}\sum_{t=1}^{N}\left(\boldsymbol{\lambda}^{(t)}_{post(i)} - \overline{\boldsymbol{\lambda}}_{post(i)}\right)^2$$

Now, the densities will take the form:

$$g(\lambda) = \text{Normal}(\lambda;\, \overline{\boldsymbol{\lambda}}_{post(i)}, s^2_{post(i)})$$

Since we are still using the posterior sample for the Generalised Harmonic Mean estimate, we will be using the same likelihood values as calculated for the Harmonic Mean estimate in Section 4.3.4. However, now we must also calculate the prior density values and proposal density values of the 50 posterior samples. For the prior density values, we have:

$$p(\lambda^{(j)}_{post(i)}) = \text{Gamma}(\lambda^{(j)}_{post(i)}; 0.1, 0.1)$$

Similarly, for the proposal density values, we have:

$$g(\lambda^{(j)}_{post(i)}) = \text{Normal}(\lambda^{(j)}_{post(i)};\, \overline{\boldsymbol{\lambda}}_{post(i)}, s^2_{post(i)})$$

Again, we receive the generalised harmonic mean estimates by taking the log of the formula derived in Section 3.3:

$$\log \hat{p}(\mathbf{y}) = \log\left(\frac{1}{N}\sum_{t=1}^{N}\frac{g(\lambda^{(t)}_{post})}{p(\mathbf{y}|\lambda^{(t)}_{post})\,p(\lambda^{(t)}_{post})}\right)^{-1} = -\log\left(\frac{1}{N}\sum_{t=1}^{N}\frac{g(\lambda^{(t)}_{post})}{p(\mathbf{y}|\lambda^{(t)}_{post})\,p(\lambda^{(t)}_{post})}\right)$$

The 50 Generalised Harmonic Mean estimates are then given by:

$$\log \hat{p}_{GHM(1)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N}\frac{g(\lambda^{(t)}_{post(1)})}{p(\mathbf{y}_1|\lambda^{(t)}_{post(1)})\,p(\lambda^{(t)}_{post(1)})}\right)$$

$$\vdots$$

$$\log \hat{p}_{GHM(50)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N}\frac{g(\lambda^{(t)}_{post(50)})}{p(\mathbf{y}_{50}|\lambda^{(t)}_{post(50)})\,p(\lambda^{(t)}_{post(50)})}\right)$$

4.3.6 Newton-Raftery Estimator

We will be running the Newton-Raftery estimator with 1000 iterations and the δ set to 0.2. For this estimator, we make use of both the posterior sample and the prior sample through the proposal mixture density:

$$g(\lambda) = \delta p(\lambda) + (1-\delta)p(\lambda|\mathbf{y}) \qquad\qquad (0 < \delta < 1)$$

We need to draw a sample of size N from this mixture density, which we will call the NR sample. Obtaining this sample is equivalent to randomly replacing $\delta \times N$ elements of the posterior sample with elements from the prior sample. Doing this with the 50 posterior samples and corresponding prior samples results in the 50 NR samples $\boldsymbol{\lambda}_{NR(1)}, \ldots, \boldsymbol{\lambda}_{NR(1)}$.

We must now calculate the likelihood for each of the 50 NR samples as we did in Section 4.3.3 for the prior samples and in Section 4.3.4 for the posterior samples, where we have:

$$p(\mathbf{y}_i|\lambda^{(j)}_{NR(i)}) = \prod_{\ell=1}^{n} \text{Poisson}(y_{i,\ell}\,; \lambda^{(j)}_{NR(i)})$$

As we have now incorporated the prior samples into the posterior samples to form the NR samples, we face the same decision as with the basic Monte Carlo estimator in Section 4.3.3 with zero likelihoods. We will again choose not to round up any zero values that may occur.

To reach the final estimates, we must now use the calculated likelihoods with the iterative formula stated in Section 3.4:

$$\hat{p}(\mathbf{y}) = \frac{\sum_{t=1}^{N} p(\mathbf{y}|\lambda^{(t)}_{NR})\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\lambda^{(t)}_{NR})\}^{-1}}{\sum_{t=1}^{N}\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}}$$

We must also make a decision on how we initialise $\hat{p}(\mathbf{y})$ with p_0. From trialling different options for p_0, including setting it equal to the Monte Carlo estimate $\hat{p}_{MC}$ and Harmonic Mean estimate, we can conclude that the starting point of the iteration will not significantly affect the final estimate. Therefore, it seems most sensible to initialise at a constant that is unrelated to the other methods. For this, we will simply choose $\hat{p}_0 = 1$ as the initial point for all 50 estimates. For each of the 50 sets of likelihoods, we perform 1,000 iterations of the iterative formula using $\hat{p}_0 = 1$. Figure 4.2 shows how the iterative sequence from the first set of likelihoods converges. After taking the logs of the final iterations, this leaves the 50 Newton-Raftery estimates of the form:

$$\log \hat{p}_{NR(1)}, \ldots, \log \hat{p}_{NR(50)}$$

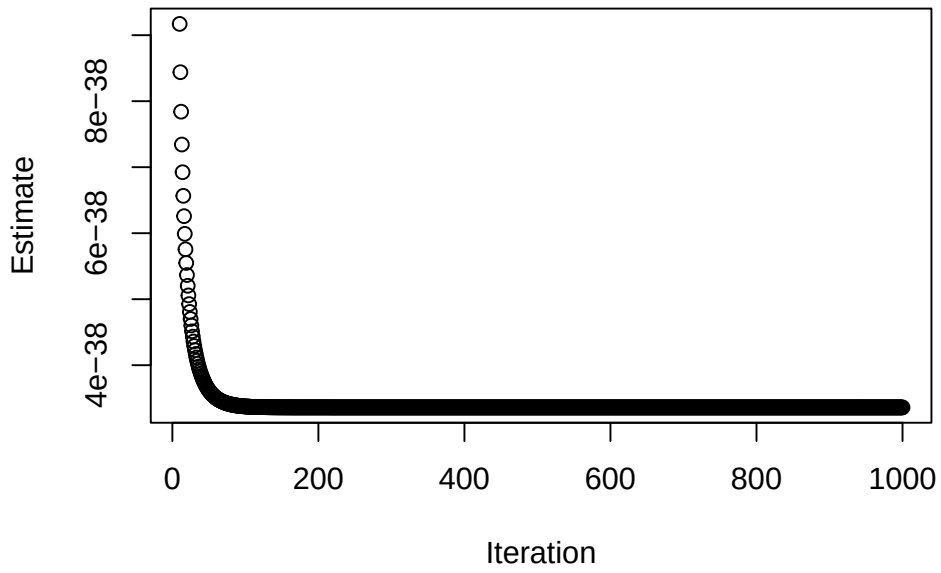

Figure 4.2: Series plot of the 10th iteration to the 1000th iteration for the first Newton-Raftery estimate

4.3.7 Laplace-Metropolis Estimator

In the Gamma-Poisson setting, the Laplace-Metropolis estimator from Section 3.5 is given by:

$$\hat{p}(\mathbf{y}) = (2\pi)^{d/2}|\mathbf{S}|^{1/2}p(\mathbf{y}|\lambda^{max})p(\lambda^{max})$$

$$\lambda^{max} = \max_{t=1,\cdots,N}\{p(y|\lambda^{(t)}_{post})p(\lambda^{(t)}_{post})\}$$

Therefore, we must first calculate the λ^{max} for each of the 50 posterior samples $\boldsymbol{\lambda}_{post(1)}, \ldots, \boldsymbol{\lambda}_{post(50)}$. We will label these values $\lambda^{max}_1, \ldots, \lambda^{max}_{50}$. Following this, we calculate the product of the likelihood and prior densities corresponding to these λ values, which will be of the form $p(\mathbf{y}|\lambda^{max}_{(i)})p(\lambda^{max}_{(i)})$. In order to complete the estimate, we must also calculate $\mathbf{S}$ values, which similar to the λ^{max} values, we will denote $\mathbf{S}_{(1)}, \ldots, \mathbf{S}_{(50)}$ For the Gamma-Poisson model, these values are simply the sample variances of the posterior samples, as we are only working with univariate parameters:

$$\mathbf{S}_{(i)} = \frac{1}{N-1}\sum_{t=1}^{N}\left(\lambda^{(t)}_{post(i)} - \overline{\boldsymbol{\lambda}}_{post(i)}\right)^2$$

Now, the 50 Laplace-Metropolis estimates are given by:

$$\log \hat{p}_{LM(1)} = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{S}_{(1)}| + \log\left(p(\mathbf{y}|\lambda_{(1)}^{max})p(\lambda_{(1)}^{max})\right)$$
$$\vdots$$
$$\log \hat{p}_{LM(50)} = \frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{S}_{(50)}| + \log\left(p(\mathbf{y}|\lambda_{(50)}^{max})p(\lambda_{(50)}^{max})\right)$$

4.3.8 Bridge Sampling Estimator

As with the Newton-Raftery Estimator, we will be running the Bridge Sampling estimator with 1000 iterations. Also, we set the values $N1$ and $N2$ equal to N, which gives:

$$s_1 = \frac{N}{N+N} = \frac{1}{2} \quad \textit{and} \quad s_2 = \frac{N}{N+N} = \frac{1}{2}$$

The bridge sampling method is unique in that it requires a sample from the proposal distribution alongside the posterior sample. For this, we construct 50 proposal densities in the same way as we did for the generalised harmonic mean estimator in Section 4.3.5, resulting in:

$$g(\lambda) = \text{Normal}(\lambda; \overline{\boldsymbol{\lambda}}_{post(i)}, s^2_{post(i)})$$

We will generate the 50 proposal samples of size N by sampling directly from the 50 proposal distributions $G(\lambda)$. We summarise the samples in a matrix labelled Λ_{prop}, which takes the form:

$$\Lambda_{prop} = \begin{bmatrix} \boldsymbol{\lambda}^{\mathrm{T}}_{prop\,(1)} \\ \boldsymbol{\lambda}^{\mathrm{T}}_{prop\,(2)} \\ \vdots \\ \boldsymbol{\lambda}^{\mathrm{T}}_{prop\,(50)} \end{bmatrix} = \begin{bmatrix} \lambda^{(1)}_{prop(1)} & \lambda^{(2)}_{prop(1)} & \cdots & \lambda^{(N)}_{prop(1)} \\ \lambda^{(1)}_{prop(2)} & \lambda^{(2)}_{prop(2)} & \cdots & \lambda^{(N)}_{prop(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^{(1)}_{prop(50)} & \lambda^{(2)}_{prop(50)} & \cdots & \lambda^{(N)}_{prop(50)} \end{bmatrix}$$

Now that we have generated the proposal samples, the formula derived in Section 3.6 can be written:

$$\hat{p}(\mathbf{y}) = \frac{\frac{1}{N_2}\sum_{t=1}^{N_2} \dfrac{p(\mathbf{y}|\lambda^{(t)}_{prop})p(\lambda^{(t)}_{prop})}{s_1 p(\mathbf{y}|\lambda^{(t)}_{prop})p(\lambda^{(t)}_{prop}) + s_2\hat{p}(\mathbf{y})g(\lambda^{(t)}_{prop})}}{\frac{1}{N_1}\sum_{t=1}^{N_1} \dfrac{g(\lambda^{(t)}_{post})}{s_1 p(\mathbf{y}|\lambda^{(t)}_{post})p(\lambda^{(t)}_{post}) + s_1\hat{p}(\mathbf{y})g(\lambda^{(t)}_{post})}}$$

After substituting in the chosen values for N_1 and N_2 as well as the corresponding s_1

and s_2 this becomes:

$$\hat{p}(\mathbf{y}) = \frac{\frac{2}{N}\sum_{t=1}^{N} \frac{p(\mathbf{y}|\lambda_{prop}^{(t)})p(\lambda_{prop}^{(t)})}{p(\mathbf{y}|\lambda_{prop}^{(t)})p(\lambda_{prop}^{(t)}) + \hat{p}(\mathbf{y})g(\lambda_{prop}^{(t)})}}{\frac{2}{N}\sum_{t=1}^{N} \frac{g(\lambda_{post}^{(t)})}{p(\mathbf{y}|\lambda_{post}^{(t)})p(\lambda_{post}^{(t)}) + \hat{p}(\mathbf{y})g(\lambda_{post}^{(t)})}} \tag{2}$$

We notice here that all of the quantities required to compute the denominator of the formula have already been calculated in previous methods. We can use the likelihoods of the posterior sample calculated for the harmonic mean estimator in Section 4.3.4 and the proposal density values and priors of the posterior samples can be taken from the Generalised Harmonic Mean implementation in Section 4.3.5. However, we still need to obtain the proposal, likelihood and prior density values for the new proposal samples. Firstly, the proposal density values are calculated from $g(\lambda)$, which will take the form:

$$g(\lambda_{prop(i)}^{(j)}) = \text{Normal}(\lambda_{prop(i)}^{(j)}; \overline{\boldsymbol{\lambda}}_{post(i)}, s_{post(1)}^2)$$

We similarly calculate the prior density values of the proposal samples:

$$p(\lambda_{prop(i)}^{(j)}) = \text{Gamma}(\lambda_{prop(i)}^{(j)}; 0.1, 0.1)$$

as well as the likelihood density values:

$$p(\mathbf{y}_i|\lambda_{prop(i)}^{(j)}) = \prod_{\ell=1}^{n} \text{Poisson}(y_{i,\ell}\,; \lambda_{prop(i)}^{(j)})$$

We can summarise these values in 3 $50 \times N$ matrices as we have done previously, to ease the computation of the final estimates. Now that we have obtained the 6 matrices of density values from the posterior and proposal samples that we require, we can use the iterative formula (2). As with the Newton-Raftery estimator, we will initialise at $\hat{p}_0 = 1$ and then perform 1000 iterations for each of the 50 sets of posterior and proposal samples. After taking the logs of the final iterations, we have the 50 bridge sampling estimates of the form:

$$\log \hat{p}_{BS(1)}, \ldots, \log \hat{p}_{BS(50)}$$

4.3.9 Fourier Integral Estimator

In order to implement the Fourier integral estimator, we must first decide at which point in the parameter space to centre the posterior density estimate. For this, we take the means of the posterior samples. For example, we set:

$$\lambda_{(i)}^{*} = \overline{\boldsymbol{\lambda}}_{post(i)}$$

In the Gamma-Poisson setting, we must use the posterior density estimate formula derived in Section 3.7 for the univariate case, which can be applied as follows:

$$\hat{p}(\lambda^*_{(i)}|\mathbf{y}_i) = \frac{1}{N\pi}\sum_{t=1}^{N}\frac{\sin(R(\lambda^*_{(i)} - \lambda^{(t)}_{post}))}{\lambda^*_{(i)} - \lambda^{(t)}_{post}}$$

Here, we must decide on a value of R in order to progress with the 50 density estimates. We can assess convergence by producing a series plot of the first 5 estimates, varying R from 1 to 100. Due to the very simple form of the density estimates, we are able to produce 100 estimates for each set quickly and with little computational effort. The results of the series plots are shown in Figure 4.3 which show that we reach convergence around R=10. We will proceed with R=20 for the final steps of the method. The

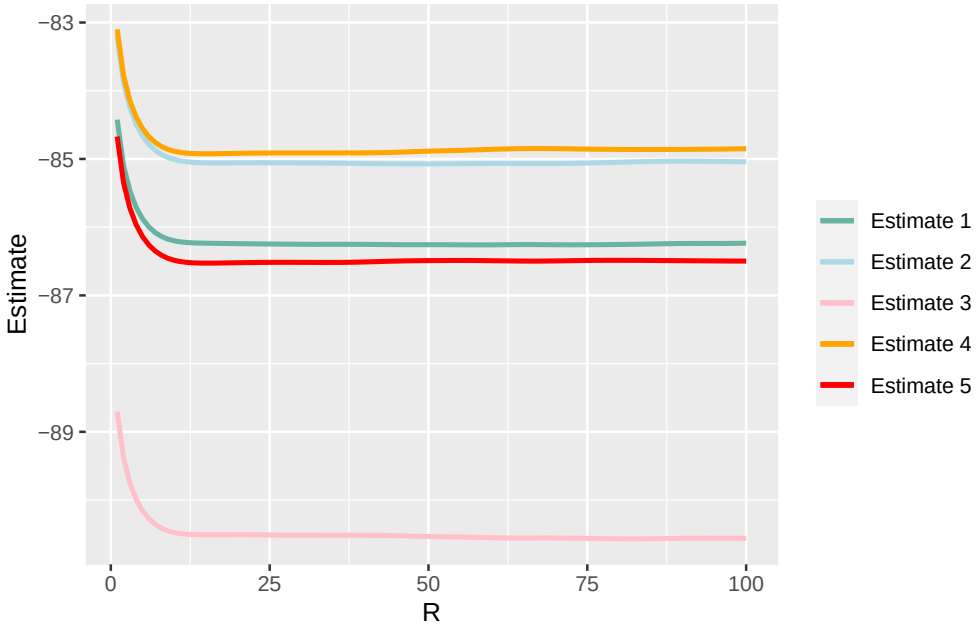

Figure 4.3: Series plot of the first 5 estimates when increasing R from 1 to 100

likelihood and prior density values need to be calculated at the 50 centres $\lambda^*_{(1)}, \ldots, \lambda^*_{(50)}$ before proceeding with the model evidence estimates. These density values will be of the form $p(\mathbf{y}_i|\lambda^*_i)$ and $p(\lambda^*_i)$ respectively. We can now use Bayes' Theorem to arrive at the final estimates:

$$\begin{aligned}
\log \hat{p}_{FI(1)} &= \log \hat{p}(\lambda^*_{(1)}|\mathbf{y}_1) + \log p(\lambda^*_1) - \log p(\mathbf{y}_1|\lambda^*_1) \\
&\vdots \\
\log \hat{p}_{FI(1)} &= \log \hat{p}(\lambda^*_{(1)}|\mathbf{y}_{50}) + \log p(\lambda^*_1) - \log p(\mathbf{y}_{50}|\lambda^*_1)
\end{aligned}$$

4.4 Results (Gamma-Poisson Conjugate Model)

Now that we have obtained the estimates for each method, we can compare these against the true model evidence values stored in vector **m**. First, we will consider the absolute errors of each set of estimates. The constructed box plots of these errors, for each method, can be found in Figure 4.4. We can see that most of the methods have performed well

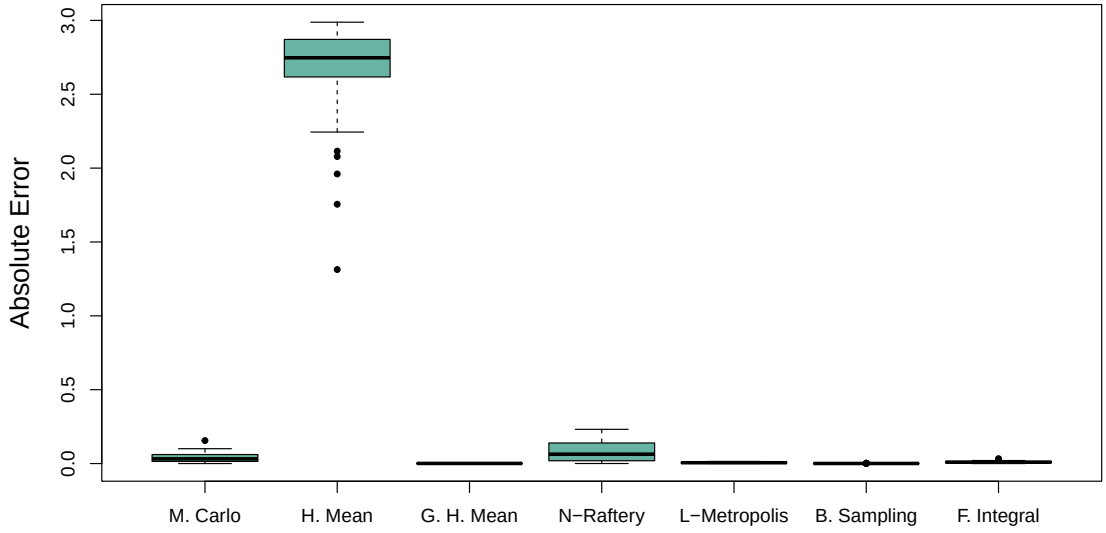

Figure 4.4: Box plots of the absolute errors of the 50 estimates

here, apart from the harmonic mean estimator. The generalised harmonic mean, Laplace-Metropolis, bridge sampling and Fourier integral estimators have performed particularly well, showing very thin box plots close to zero which represents close to zero errors across all 50 estimates. We can provide a numerical summary of the estimators by calculating the mean and standard deviation of the absolute errors on the logarithmic scale, which can be found in Table 4.1.

Methods	Mean	Std. Dev.
Monte Carlo	0.04057	0.03295
Harmonic Mean	2.66557	0.33388
Gen. Harmonic Mean	0.00087	0.00060
Newton-Raftery	0.08100	0.06686
Laplace-Metropolis	0.00549	0.00349
Bridge Sampling	0.00069	0.00045
Fourier Integral	0.01046	0.00700

Table 4.1: Means and standard deviations of the absolute errors for the 50 estimates of each method

We can see that the bridge sampling method has performed the best here with the lowest mean error and the lowest standard deviation, with the Generalised Harmonic Mean performing to similar ability. Although the Fourier Integral estimator hasn't performed as well, we note that its strength lies in its computational efficiency and we can perform the estimate with a much larger sample size with relative ease. To demonstrate this numerically, the total elapsed time to run the Bridge Sampling code in this setting is 86.532s whereas the elapsed time for the Fourier Integral estimate is 0.052s - a substantial difference. Note that as mentioned in the start of the chapter, there will be ways to optimise the bridge sampling code and this doesn't account for the time taken to obtain the posterior sample which is used in both methods, however it is clear that the Fourier Integral estimator is more computationally efficient. Increasing the posterior sample to $N = 10^7$ leads to the new Fourier integral estimator summaries that can be found in Table 4.2, where the results now align closely with that of the generalised harmonic mean and bridge sampling estimator.

Methods	Mean	Std. Dev.
Fourier Integral	0.001040	0.0008212

Table 4.2: Means and standard deviations of the absolute errors for the Fourier Integral estimates with $N = 10^7$

Although the Newton-Raftery estimator has improved the estimates in comparison to the harmonic mean estimator with the incorporation of the prior sample, it has not performed well in comparison to the rest of the estimators and noticeably has a higher mean and standard deviation score than the basic Monte Carlo estimator.

Chapter 5

Intractable Example: The Logistic Regression Model

5.1 The Logistic Regression Model

We will now introduce a set of logistic regression models $\boldsymbol{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_7\}$, where the model evidence is analytically intractable. We will introduce the data as independent variables consisting of 3 covariates, $\mathbf{x}_i = (x_{i,1}, x_{i,2}, x_{i,3})^{\mathrm{T}}$ $(i = 1, \ldots, n)$ with corresponding binomial response variables $y_1, \ldots, y_n$. The goal here is to perform model selection with the 7 estimation methods. We will select one of the models in the set $\boldsymbol{M}$ to be the true model and then simulate the binomial responses $y_1, \ldots, y_n$ using this model. We will then estimate the model evidence of each model and calculate the corresponding Bayes factors against the true model. With the aid of the interpretations in Table 2.1 we can judge whether the methods have successfully distinguished the true model from the other models in the set $\boldsymbol{M}$.

The logistic regression parameters are $\beta_0, \beta_1, \beta_2, \beta_3$ with β_0 being the intercept parameter and $\beta_1, \beta_2, \beta_3$ being the coefficient parameters. We can strategically omit combinations of the coefficient parameters in order to construct the models. The different combinations of omitting the coefficient parameters gives rise to a model space of size $2^3 = 8$. We will be working with all of the models in this space, apart from the trivial model which omits all 3 coefficient parameters. These seven models, which comprise $\boldsymbol{M}$, are listed below, where $\mathcal{M}_5$ is chosen as the true model for the data generation:

$$\begin{aligned}
&\mathcal{M}_1 : \boldsymbol{\beta} = (\beta_0, \beta_1, 0, 0) && \mathcal{M}_5 : \boldsymbol{\beta} = (\beta_0, \beta_1, 0, \beta_3) \quad \text{(True model)} \\
&\mathcal{M}_2 : \boldsymbol{\beta} = (\beta_0, 0, \beta_2, 0) && \mathcal{M}_6 : \boldsymbol{\beta} = (\beta_0, 0, \beta_2, \beta_3) \\
&\mathcal{M}_3 : \boldsymbol{\beta} = (\beta_0, 0, 0, \beta_3) && \mathcal{M}_7 : \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) \\
&\mathcal{M}_4 : \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, 0)
\end{aligned}$$

The binomial response $\mathbf{y} = (y_1, \ldots, y_n)^T$ will rely on $\mathbf{p} = (p_1, \ldots, p_n)^T$ and m where $\mathbf{p}$ is the vector containing the probability parameters for each of the individual responses

$y_i \in \mathbf{y}$ and m is the parameter denoting the number of trials of each binomial response. Note that it is commonplace to use n for the number of trials, however we will reserve this for the sample size of the data.
For a specific $p_i \in \mathbf{p}$, we define the odds ratio as the ratio of the probability of a successful trial to the probability of a trial not being successful:

$$\text{Odds ratio} = \frac{p_i}{1 - p_i}$$

In the logistic regression setting, the log of the odds ratio is a linear combination of the covariates. For example, with the model $\mathcal{M}_7 \in \boldsymbol{M}$, we have:

$$\text{logit}\left(p_i\right) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where the *logit* function maps the probability to a value on the entire space of real numbers $(-\infty, +\infty)$:

$$\text{logit}\left(x\right) = \log\left(\frac{x}{1 - x}\right) \quad (x \in [0, 1])$$

We will conveniently place the generated data into an $n \times 4$ matrix, with the first column consisting of a vector of 1s for the intercept parameter.

$$X_{n\text{x}4} = \begin{bmatrix} 1 & \mathbf{x}_1^{\mathrm{T}} \\ 1 & \mathbf{x}_2^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & x_{n,3} \end{bmatrix}$$

Then, we can efficiently compute the vector of probability parameters for the binomial response:

$$\begin{aligned} \log\left(\frac{\mathbf{p}}{1 - \mathbf{p}}\right) &= X\boldsymbol{\beta} \\ \frac{\mathbf{p}}{1 - \mathbf{p}} &= \exp(X\boldsymbol{\beta}) \\ \mathbf{p} &= \frac{\exp(X\boldsymbol{\beta})}{1 + \exp(X\boldsymbol{\beta})} \\ \mathbf{p} &= \frac{1}{1 + \exp(-X\boldsymbol{\beta})} \end{aligned}$$

Now, for each model $\mathcal{M}_i \in \boldsymbol{M}$, we can define the prior and likelihood in concise notation:

$$\begin{aligned} y_i &\sim \text{Bin}\left(m, p_i\right) \qquad (i = 1, \ldots, n) \\ \beta_k &\sim \mathcal{N}\left(\mu, \sigma^2\right) \qquad (k = 0, \ldots, 3) \end{aligned}$$

The model specifications we will use for the remainder of the chapter can be summarised in the following table:

size of data sets:	$n = 20$
posterior/ prior sample size:	$N = 10000$
response parameter:	$m = 20$
prior hyper-parameters:	$\mu = 0,\ \sigma^2 = 100$

As for the toy example in Section 4, we aim to introduce a vague prior through the σ^2 parameter. Setting this to 100 will ensure that the prior takes a flat shape with density distributed well through the tails.

5.2 Data Simulation

As in Section 4, we will simulate the data. We choose to simulate 20 sets of independent variables and corresponding response variables. The independent variables consisting of three covariates $\mathbf{x_i} = (x_{i,1}, x_{i,2}, x_{i,3})^{\mathrm{T}}$ $(i = 1, \ldots, 20)$ will be drawn independently from $x_{i,j} \sim \mathcal{N}(0, 5)$ $(i = 1, \ldots, 20$ and $j = 1, \ldots, 3)$. When placed into the $n \times 4$ matrix, this gives:

$$X_{n\mathrm{x}4} = \begin{bmatrix} 1 & \mathbf{x}_1^{\mathrm{T}} \\ 1 & \mathbf{x}_2^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & x_{n,3} \end{bmatrix}$$

We now use the true model $\mathcal{M}_5 : \boldsymbol{\beta} = (\beta_0, \beta_1, 0, \beta_3)$ to generate the 20 binomial responses. We set the logistic regression parameters $\boldsymbol{\beta} = (1, 2, 0, -1.5)$ for the purpose of data generation. Now, using these logistic regression parameters we must calculate the p values $\mathbf{p} = (p_1, \ldots, p_{20})$ using the derived formula:

$$\mathbf{p} = \frac{1}{1 + \exp(-X\boldsymbol{\beta})}$$

We can now directly simulate the response variables $y_1, \ldots, y_{20}$ from Binomial$(m, \mathbf{p})$ with m set to 20.

5.3 Implementing the Methods

In this section, we will now describe how we can implement the estimator methods within the framework of the logistic regression models, again working on the logarithmic scale. As before, note that there are ways in which the code used to implement these methods could be optimised for greater efficiency. The implementation code for the logistic regression models as well as the *Stan* file used to encode the models can be found through this link.

First, we must introduce the methods we use to assess the variability of the estimators in this intractable logistic regression example. Essentially, if we split the posterior,

prior and proposal samples into equally sized batches, we can calculate an estimate from each distinct batch. Following this, we will compute a mean estimate for each estimator as well as the standard deviation of the estimates, known as the Monte Carlo (MC) errors . From now on, we will refer to these batches as chains and for each model we will be generating 50 chains of samples to use for the estimators. For the chains, we will use the notation $\mathcal{B}_{prior(1)}, \dots \mathcal{B}_{prior(50)}$, $\mathcal{B}_{post(1)}, \dots, \mathcal{B}_{post(50)}$ *and* $\mathcal{B}_{prop(1)}, \dots, \mathcal{B}_{prop(50)}$ to denote the 50 chains of prior, posterior and proposal samples respectively, where $\boldsymbol{\beta}^{(1)}_{prior\,(1)}$ will denote the first sample of betas in the first prior chain. To avoid introducing excessive notation, we will not differentiate between the samples of each model. This will not be an issue as we will only be using $\mathcal{M}_4 : \beta = (\beta_0, \beta_1, \beta_2, 0)$ to demonstrate the implementation of the methods, where the same techniques can be applied to every model in the set $\boldsymbol{M}$. Note that the prior, posterior and proposal chains generated for each model will only include the coefficient parameters that have not been omitted.

5.3.1 Sampling from the Prior

For the prior sample, we will generate this by sampling directly from the prior distribution:

$$\beta_k \sim \mathcal{N}(0, 100) \qquad (k = 0, \dots, 3)$$

For example, in $\mathcal{M}_4$, for the ith prior chain $\mathcal{B}_{prior(i)}$ we have:

$$\mathcal{B}_{prior(i)} = \begin{bmatrix} \boldsymbol{\beta}^{(1)T}_{prior\,(i)} \\ \boldsymbol{\beta}^{(2)T}_{prior\,(i)} \\ \vdots \\ \boldsymbol{\beta}^{(N)T}_{prior\,(i)} \end{bmatrix} = \begin{bmatrix} \beta^{(1)}_{0\ prior(i)} & \beta^{(1)}_{1\ prior(i)} & \beta^{(1)}_{2\ prior(i)} \\ \beta^{(2)}_{0\ prior(i)} & \beta^{(2)}_{1\ prior(i)} & \beta^{(2)}_{2\ prior(i)} \\ \vdots & \vdots & \vdots \\ \beta^{(N)}_{0\ prior(i)} & \beta^{(N)}_{1\ prior(i)} & \beta^{(N)}_{2\ prior(i)} \end{bmatrix}$$

5.3.2 Sampling from the Posterior

With the logistic regression models, we no longer have an analytically tractable posterior distribution which we can sample from directly. Instead, we resort to Markov Chain Monte Carlo (MCMC) methods to obtain the posterior sample, needed for all of the methods apart from the basic Monte Carlo estimator. For this, we will encode the logistic regression models in the interface *Rstan*, which implements an MCMC method called Hamiltonian Monte Carlo to obtain a posterior sample, after we have inputted the data generated in Section 5.2 . Note that here we are simulating different sets of posterior chains for each model, and we must specify which of the coefficient parameters have been set to zero. For this, we will input the adjusted data matrices which remove the covariate data corresponding to zeroed coefficient parameters, and then set the number of coefficient parameters equal to the number of unspecified coefficient parameters. We will set the number of iterations to 11000 for each of the 50 posterior sample chains, with the first 1,000 iterations disregarded as warm up samples. For example, in $\mathcal{M}_4$, for

the ith posterior chain, $\mathcal{B}_{post(i)}$ we have:

$$\mathcal{B}_{post(i)} = \begin{bmatrix} \boldsymbol{\beta}^{(1)T}_{post\,(i)} \\ \boldsymbol{\beta}^{(2)T}_{post\,(i)} \\ \vdots \\ \boldsymbol{\beta}^{(N)T}_{post\,(i)} \end{bmatrix} = \begin{bmatrix} \beta^{(1)}_{0\ post(i)} & \beta^{(1)}_{1\ post(i)} & \beta^{(1)}_{2\ post(i)} \\ \beta^{(2)}_{0\ post(i)} & \beta^{(2)}_{1\ post(i)} & \beta^{(2)}_{2\ post(i)} \\ \vdots & \vdots & \vdots \\ \beta^{(N)}_{0\ post(i)} & \beta^{(N)}_{1\ post(i)} & \beta^{(N)}_{2\ post(i)} \end{bmatrix}$$

Some convergence diagnostics can be found in Figure 5.1 and Figure 5.2 for the posterior sample generated from model $\mathcal{M}_7$. Now we can proceed with the implementation of

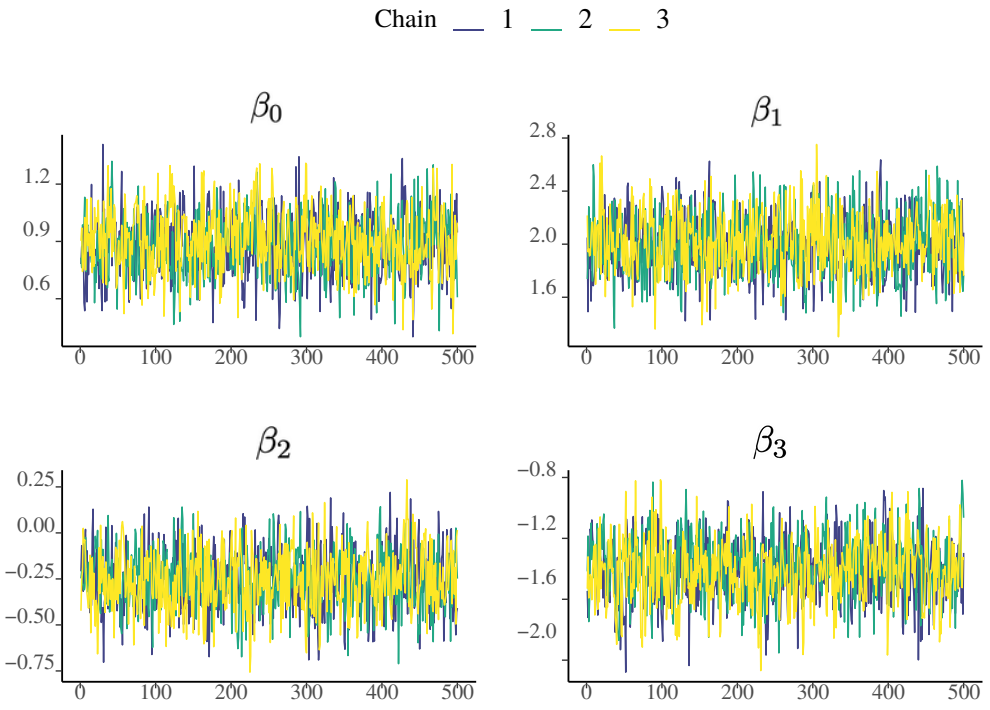

Figure 5.1: Trace plots of the first 500 post warm up iterations from chains 1,2 and 3

the model evidence estimators. We will be using $\mathcal{M}_4 : \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, 0)$ as the example model for implementing the techniques, where the same approach is used for every model in the set $\boldsymbol{M}$. The adjusted data matrix, denoted X' will be formed by removing the

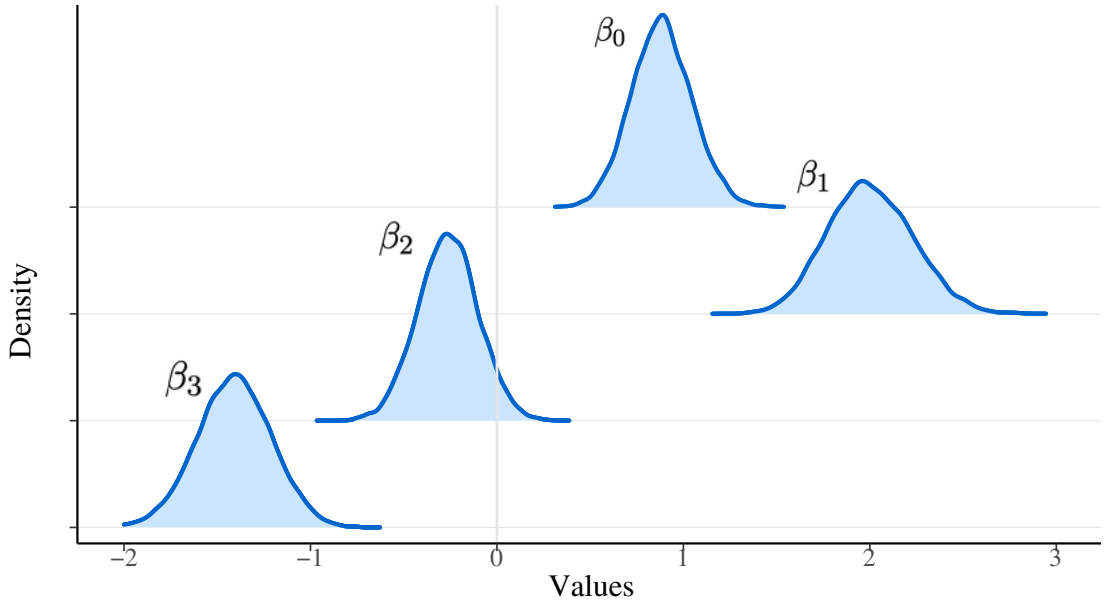

Figure 5.2: 3d density plots of the posterior beta samples using all post warm up iterations from the first posterior chain

third covariate from the original data:

$$X'_{nx3} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix}$$

5.3.3 Basic Monte Carlo Estimator

For the basic Monte Carlo estimator discussed in Section 3.1, we must calculate the likelihoods for each of the prior sample. In contrary to the Gamma-Poisson example, we must first calculate the vectors of p values corresponding to the prior samples, using the formula:

$$\mathbf{p} = \frac{1}{1 + \exp(-X\boldsymbol{\beta})}$$

For the ith prior chain, this gives the following p values:

$$\mathbf{p}^{(j)}_{prior(i)} = \frac{1}{1 + \exp(-X'\boldsymbol{\beta}^{(j)}_{prior(i)})}$$

After we have collected the vectors of p values for each prior sample, we can proceed with calculating the likelihoods. As laid out in Section 5.1, the likelihood takes the binomial

form. For the ith prior chain, the likelihoods can be derived as:

$$p(\mathbf{y}|\boldsymbol{\beta}^{(j)}_{prior(i)}) = \prod_{\ell=1}^{n} \text{Bin}\left(y_\ell; m, p^{(j)}_{\ell\ prior(i)}\right)$$

Now, the 50 Monte Carlo estimates are given by:

$$\log \hat{p}_{MC(1)} = \log\left(\frac{1}{N}\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\beta}^{(t)}_{prior(1)})\right)$$

$$\vdots$$

$$\log \hat{p}_{MC(50)} = \log\left(\frac{1}{N}\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\beta}^{(t)}_{prior(50)})\right)$$

5.3.4 Harmonic Mean Estimator

For the harmonic mean estimator we also must access the likelihoods, this time for the posterior samples in order to reach the derived formula in Section 3.2. Again, we begin by calculating the vectors of p values corresponding to the posterior samples before obtaining the likelihoods. The p values of the ith posterior chain are given by:

$$\mathbf{p}^{(j)}_{post(i)} = \frac{1}{1 + \exp(-X'\boldsymbol{\beta}^{(j)}_{post(i)})}$$

We can now calculate the likelihoods as we did for the prior samples. The ith posterior chain gives the likelihoods:

$$p(\mathbf{y}|\boldsymbol{\beta}^{(j)}_{post(i)}) = \prod_{\ell=1}^{n} \text{Bin}\left(y_\ell; m, p^{(j)}_{\ell\ post(i)}\right)$$

By taking the log of the final formula, we can obtain the final 50 estimates from the sets of likelihoods:

$$\log \hat{p}_{HM(1)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N} \frac{1}{p(\mathbf{y}|\boldsymbol{\beta}^{(t)}_{post(1)})}\right)$$

$$\vdots$$

$$\log \hat{p}_{HM(50)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N} \frac{1}{p(\mathbf{y}|\boldsymbol{\beta}^{(t)}_{post(50)})}\right)$$

5.3.5 Generalised Harmonic Mean Estimator

For the generalised harmonic mean estimator, we must introduce a proposal density. In the Toy Example, we achieved this by introducing simple Gaussian proposal densities

stemming from the sample means and sample variances of the posterior samples. Now that we are working in a multivariate parameter space, we will need to employ multivariate Gaussian distributions for the proposal densities. Specifically, we will be introducing 50 multivariate Gaussian densities of the form $g(\boldsymbol{\beta})$, each corresponding to a distinct posterior chain. We will set the mean of each Gaussian proposal to the sample mean of the posteriors samples from the corresponding chain and the variance equal to the sample variance of the posterior samples from the corresponding chain. For example, with the ith chain we have:

$$\overline{\boldsymbol{\beta}}_{post(i)} = \begin{bmatrix} \frac{1}{N}\sum_{t=1}^{N} \beta_{0\ post(i)}^{(t)} \\ \frac{1}{N}\sum_{t=1}^{N} \beta_{1\ post(i)}^{(t)} \\ \frac{1}{N}\sum_{t=1}^{N} \beta_{2\ post(i)}^{(t)} \end{bmatrix}$$

$$S_{post(i)} = \frac{1}{N}\left(\mathcal{B}_{post(i)} - \overline{\boldsymbol{\beta}}_{post(i)}\right)\left(\mathcal{B}_{post(i)} - \overline{\boldsymbol{\beta}}_{post(i)}\right)^{\mathrm{T}}$$

The 50 proposal densities then take the form:

$$g(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta};\, \overline{\boldsymbol{\beta}}_{post(i)}, S_{post(i)})$$

To apply the generalised harmonic mean formula from Section 3.3, we must obtain the likelihood, prior and proposal density values evaluated at each posterior sample. Again, we will recycle the likelihoods calculated for the harmonic mean estimator, meaning we only need the prior and proposal density values. The prior density values of the ith posterior chain are given as:

$$p(\boldsymbol{\beta}_{post(i)}^{(j)}) = \prod_{\ell=0}^{2} \mathcal{N}(\beta_{\ell\ post(i)}^{(j)}; 0, 100)$$

Similarly, we obtain the proposal densities using $g(\boldsymbol{\beta})$, with the results of the ith chain taking the form:

$$g(\boldsymbol{\beta}_{post(i)}^{(j)}) = \mathcal{N}(\boldsymbol{\beta}_{post(i)}^{(j)};\, \overline{\boldsymbol{\beta}}_{post(i)}, S_{post(i)})$$

Taking the log of the formula we get the 50 estimates as:

$$\log \hat{p}_{GHM(1)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N} \frac{g(\boldsymbol{\beta}_{post(1)}^{(t)})}{p(\mathbf{y}_1|\boldsymbol{\beta}_{post(1)}^{(t)})\, p(\boldsymbol{\beta}_{post(1)}^{(t)})}\right)$$

$$\vdots$$

$$\log \hat{p}_{GHM(50)} = -\log\left(\frac{1}{N}\sum_{t=1}^{N} \frac{g(\boldsymbol{\beta}_{post(50)}^{(t)})}{p(\mathbf{y}_{50}|\boldsymbol{\beta}_{post(50)}^{(t)})\, p(\boldsymbol{\beta}_{post(50)}^{(t)})}\right)$$

5.3.6 Newton-Raftery Estimator

We run the Newton-Raftery estimator using the same settings as with the Toy Example, setting the iterations to 1000 and δ to 0.2. We incorporate the posterior and prior samples together when calculating the estimates through the mixture density:

$$g(\lambda) = \delta p(\lambda) + (1-\delta)p(\lambda|\mathbf{y}) \qquad (0 < \delta < 1)$$

We will still be replacing the posterior samples with some of the prior sample to acheive a sample from this mixture density, however we must now approach this more carefully. To obtain the NR samples, we will randomly replace $50 \times N$ samples elements of each posterior chain with elements from the corresponding prior chain. Following this, we will have 50 NR chains with which we must calculate the likelihood values. Again, when we need likelihood values in are logistic regression setting, we must calculate the associate p values of the samples:

$$\mathbf{p}_{NR(i)}^{(j)} = \frac{1}{1+\exp(-X'\boldsymbol{\beta}_{NR(i)}^{(j)})}$$

We now summarise the likelihood values for the ith chain, which alongside the other sets of likelihood values will lead to the 50 estimates:

$$p(\mathbf{y}|\boldsymbol{\beta}_{NR(i)}^{(j)}) = \prod_{\ell=1}^{n} \text{Bin}\left(y_\ell; m, p_{\ell\ NR(i)}^{(j)}\right)$$

With these likelihoods, we can now apply the iterative equation on the newly formed NR chains:

$$\hat{p}(\mathbf{y}) = \frac{\sum_{t=1}^{N} p(\mathbf{y}|\boldsymbol{\beta}_{NR}^{(t)})\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\beta}_{NR}^{(t)})\}^{-1}}{\sum_{t=1}^{N}\{\delta\hat{p}(\mathbf{y}) + (1-\delta)p(\mathbf{y}|\boldsymbol{\theta}^{(t)})\}^{-1}}$$

After the findings in Section 4.2, we will initialise the iterative equation at $p_0 = 1$ for the 50 sets of likelihoods and then take logs to obtain the final 50 estimates of the form:

$$\log\hat{p}_{NR(1)}, \ldots, \log\hat{p}_{NR(50)}$$

5.3.7 Laplace-Metropolis Estimator

Applying the formula from Section 3.5 to the logistic regression setting, we have:

$$\hat{p}(\mathbf{y}) = (2\pi)^{d/2}|\mathbf{S}|^{1/2}p(\mathbf{y}|\boldsymbol{\beta}^{max})p(\boldsymbol{\beta}^{max})$$

$$\boldsymbol{\beta}^{max} = \max_{t=1,\cdots,N}\{p(y|\boldsymbol{\beta}_{post}^{(t)})p(\boldsymbol{\beta}_{post}^{(t)})\}$$

We first calculate the $\boldsymbol{\beta}^{max}$ corresponding to each of the 50 posterior chains. We can use the likelihood values calculated for the harmonic mean estimator and the prior values calculated for the generalised harmonic mean estimator to determine the product which gives the maximum value. We label the 50 values $\boldsymbol{\beta}^{max}_{(1)}, \ldots, \boldsymbol{\beta}^{max}_{(50)}$. We use the corresponding likelihood and prior product of these values for the formula. We must now calculate the $\mathbf{S}$ values which correspond to the sample covariance matrices of the posterior chains. These are the same sample covariance matrices that we calculated in Section 5.3.5 which we labelled $S_{post(1)}, \ldots, S_{post(50)}$. For $\mathcal{M}_4$, we have 1 intercept parameter and 2 coefficient parameters, so set d equal to 3. We then obtain the following 50 estimates after taking the log the estimating formula:

$$\begin{aligned}
\log \hat{p}_{LM(1)} &= \frac{3}{2}\log(2\pi) + \frac{1}{2}\log|S_{post(1)}| + \log\left(p(\mathbf{y}|\boldsymbol{\beta}^{max}_{(1)})p(\boldsymbol{\beta}^{max}_{(1)})\right) \\
&\vdots \\
\log \hat{p}_{LM(50)} &= \frac{3}{2}\log(2\pi) + \frac{1}{2}\log|S_{post(50)}| + \log\left(p(\mathbf{y}|\boldsymbol{\beta}^{max}_{(50)})p(\boldsymbol{\beta}^{max}_{(50)})\right)
\end{aligned}$$

5.3.8 Bridge Sampling Estimator

We will stay consistent and run the bridge sampling estimator for 1000 iterations with both N_1 and N_2 set to 1000, giving $s_1 = s_2 = \frac{1}{2}$

As before, we must acquire proposal samples alongside the posterior samples in order to reach the final estimates. For the proposal distribution, we again construct 50 multivariate normal distributions which take their means and variances from the posterior sample means and sample variances respectively. This provides the same proposal densities $g(\boldsymbol{\beta})$ as in Section 5.3.5. We sample directly form the corresponding proposal distributions $G(\boldsymbol{\beta})$ to obtain the proposal chains, stored in matrices $\mathcal{B}_{prop(1)}, \ldots, \mathcal{B}_{prop(50)}$ For the ith proposal chain, $\mathcal{B}_{prop(i)}$, we have:

$$\begin{bmatrix} \boldsymbol{\beta}^{(1)\mathrm{T}}_{prop\,(i)} \\ \boldsymbol{\beta}^{(2)\mathrm{T}}_{prop\,(i)} \\ \vdots \\ \boldsymbol{\beta}^{(N)\mathrm{T}}_{prop\,(i)} \end{bmatrix} = \begin{bmatrix} \beta^{(1)}_{0\ prop(i)} & \beta^{(1)}_{1\ prop(i)} & \beta^{(1)}_{2\ prop(i)} \\ \beta^{(2)}_{0\ prop(i)} & \beta^{(2)}_{1\ prop(i)} & \beta^{(2)}_{2\ prop(i)} \\ \vdots & \vdots & \vdots \\ \beta^{(N)}_{0\ prop(i)} & \beta^{(N)}_{1\ prop(i)} & \beta^{(N)}_{2\ prop(i)} \end{bmatrix}$$

After subbing in the values for N_1, N_2 as well as the corresponding s_1 and s_2 values, the iterative formula derived in Section 3.6 becomes:

$$\hat{p(\mathbf{y})} = \frac{\frac{2}{N}\sum_{t=1}^{N} \frac{p(\mathbf{y}|\lambda^{(t)}_{prop})p(\lambda^{(t)}_{prop})}{p(\mathbf{y}|\lambda^{(t)}_{prop})p(\lambda^{(t)}_{prop}) + \hat{p(\mathbf{y})}g(\lambda^{(t)}_{prop})}}{\frac{2}{N}\sum_{t=1}^{N} \frac{g(\lambda^{(t)}_{post})}{p(\mathbf{y}|\lambda^{(t)}_{post})p(\lambda^{(t)}_{post}) + \hat{p(\mathbf{y})}g(\lambda^{(t)}_{post})}} \tag{1}$$

The likelihood, prior and proposal density values needed for the posterior chains on the denominator of (1) have already been calculated in previous methods. Therefore, in order to implement the formula we simply need to derive the equivalent values for the proposal chains. We demonstrate the implementation for the ith chain, with the proposal density values taking the form:

$$g(\boldsymbol{\beta}^{(j)}_{prop(i)}) = \text{Normal}(\boldsymbol{\beta}^{(j)}_{prop(i)}; \overline{\boldsymbol{\beta}}_{post(i)}, s^2_{post(i)})$$

We now apply the prior density to obtain the prior density values: Finally, to generate the likelihood values of the proposal samples we must calculate the vectors of p values corresponding to the proposal samples, as we did for the posterior samples in Section 5.3.4 and prior samples in Section 5.3.3 For the ith proposal chain, we get the following p values:

$$\mathbf{p}^{(j)}_{prop(i)} = \frac{1}{1 + \exp(-X'\boldsymbol{\beta}^{(j)}_{prop(i)})}$$

the binomial likelihood values then take the form:

$$p(\mathbf{y}|\boldsymbol{\beta}^{(j)}_{prop(i)}) = \prod_{\ell=1}^{n} \text{Bin}\left(y_\ell; m, p^{(j)}_{\ell\ prop(i)}\right)$$

We can now push ahead with the iterative formula (1), setting the initialiser $\hat{p}_0 = 1$. After 1000 iterations, we take logs to get the final 50 estimates of the form:

$$\log \hat{p}_{BS(1)}, \ldots, \log \hat{p}_{BS(50)}$$

5.3.9 Fourier Integral Estimator

Now that we are working in multivariate settings with the logistic regression models, we use the corresponding formula derived in Section 3.7:

$$\hat{p}(\boldsymbol{\beta}^*|y) = \frac{1}{N\pi^d} \sum_{t=1}^{N} \prod_{j=1}^{d} \frac{\sin(R(\boldsymbol{\beta}^*_j - \boldsymbol{\beta}^{(t)}_j))}{\boldsymbol{\beta}^*_j - \boldsymbol{\beta}^{(t)}_j}$$

As we did for the toy Gamma-Poisson example, we will take the means of the posterior samples as the centres of the posterior density estimate, which take the form $\boldsymbol{\beta}^*_{(i)} = \overline{\boldsymbol{\beta}}_{post(i)}$. Now we must again decide on a value of R before proceeding with the estimating formula. We first produce a series plot of the first 5 density estimates under $\mathcal{M}_4$ which can be found in Figure 5.3. We can see that in the logistic regression setting the estimates do not converge as nicely as for Gamma-Poisson model. Here, we have a short window for R where the estimator behaves well, which is roughly from R = 15 to R = 25. If we make R too large, the estimates will be susceptible to fluctuations in the sine function. To proceed, we will set R to 20, as before. Finally, in order to obtain the model evidence estimates we must calculate the posterior and prior density values of the centres. These

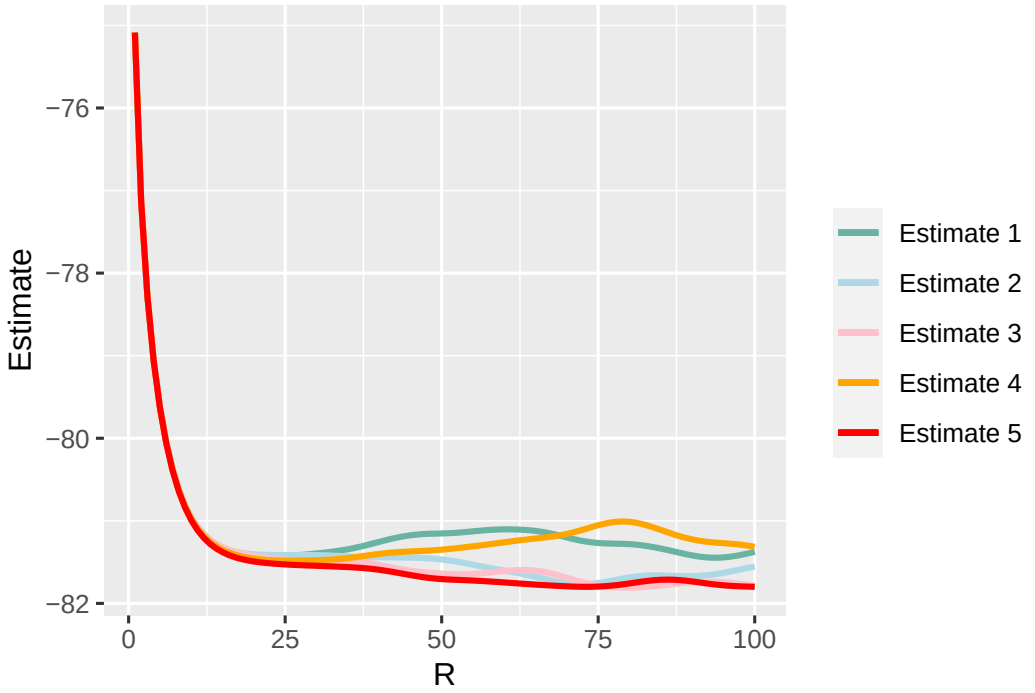

Figure 5.3: Series plot of the first 5 estimates when increasing R from 1 to 100

will take the form $p(\mathbf{y}|\boldsymbol{\beta}_i^*)$ and $p(\boldsymbol{\beta}_i^*)$ respectively. By rearranging Bayes' Theorem and taking logs we arrive at the final 50 model evidence estimates:

$$\begin{aligned} \log \hat{p}_{FI(1)} &= \log \hat{p}(\boldsymbol{\beta}_{(1)}^*|\mathbf{y}) + \log p(\boldsymbol{\beta}_1^*) - \log p(\mathbf{y}|\boldsymbol{\beta}_1^*) \\ &\vdots \\ \log \hat{p}_{FI(50)} &= \log \hat{p}(\boldsymbol{\beta}_{(50)}^*|\mathbf{y}) + \log p(\boldsymbol{\beta}_{50}^*) - \log p(\mathbf{y}|\boldsymbol{\beta}_{50}^*) \end{aligned}$$

5.4 Results (Logistic Regression Model)

Now that we have obtained the 50 batch estimates from each estimator, we can proceed with model selection. We first compute the means of the collections of batch estimates, which will be the final estimate for each of the methods. These can be found in Table 5.1. We immediately see that the basic Monte Carlo estimator has failed to produce numerical estimates for models $\mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_6, \mathcal{M}_7$. Upon further investigation, we find that a few of the batch estimates returned small values on the logarithmic scales whereas some produced zero estimates due the lack of prior simulations near the posterior mode. We also see a consistent grouping of the estimates for the general harmonic mean, Laplace Metropolis, bridge sampling and Fourier integral methods which suggests that they have located the model evidence values accurately. This is coupled with very low MC errors,

with the bridge sampling estimator again having the lowest values amongst the group. The Newton-Raftery estimator has performed outside of the top group, with larger MC errors and batch mean estimates which disagree with the previously mentioned group. However, it is clear that the Newton-Raftery estimator has marked an improvement on the harmonic mean estimator.

Using the batch mean estimates, we can now calculate the Bayes factors on the logarithmic scale for each model in comparison with the true model $\mathcal{M}_5$ using the formula:

$$\log \widehat{(BF_{5\,j})} = \log \hat{p}_{\mathcal{M}_5} - \log \hat{p}_{\mathcal{M}_j}$$

We can then compare the results against the interpretation table shown in Figure 2.1. We can see that for the General Harmonic Mean, Laplace Metropolis, bridge sampling and Fourier integral estimators, we have "decisive evidence" for the true model $\mathcal{M}_5$ over the other models in the set. Therefore, if we were performing model selection, these estimators would have successfully calculated the model evidence values to sufficient accuracy in order to confidently choose $\mathcal{M}_5$ over the other models. With the less accurate model evidence values calculated by the harmonic mean and Newton-Raftery estimators, we would still manage to confidently choose $\mathcal{M}_5$ over models $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_6$. However, the log Bayes factor of model $\mathcal{M}_5$ in comparison to $\mathcal{M}_7$ falls in the range $-1.10 \leq \log \widehat{(BF_{i\,j})} \leq 1.10$ which is considered "not worth more than a bare mention" in Jeffreys (1961). Therefore, despite having less accurate estimates of the model evidence, we are able to reduce the set of models to 2 possible models, with one of them being the true model that generated the data.

Model	M.Carlo $\log \hat{p}_{MC}$	H.Mean $\log \hat{p}_{HM}$	G.H.Mean $\log \hat{p}_{GHM}$	N.Raftery $\log \hat{p}_{NR}$	L.Metropolis $\log \hat{p}_{LM}$	B.Sampling $\log \hat{p}_{BS}$	F.Integral $\log \hat{p}_{FI}$
$\mathcal{M}_1$	**−103.900**	**−64.756**	**−75.093**	**−71.587**	**−75.080**	**−75.092**	**−75.073**
-	(29.358)	(0.479)	(0.001)	(0.175)	(0.015)	(0.001)	(0.020)
$\mathcal{M}_2$	**−196.826**	**−134.142**	**−145.058**	**−140.920**	**−145.049**	**−145.057**	**−144.997**
-	(47.472)	(0.644)	(0.001)	(0.269)	(0.013)	(0.0004)	(0.016)
$\mathcal{M}_3$	**−137.273**	**−93.904**	**−104.385**	**−100.705**	**−104.379**	**−104.384**	**−104.359**
-	(27.030)	(0.554)	(0.001)	(0.227)	(0.014)	(0.008)	(0.021)
$\mathcal{M}_4$	**Inf**	**−65.587**	**−81.451**	**−72.391**	**−81.437**	**−81.449**	**−81.431**
-	(NaN)	(0.431)	(0.001)	(0.247)	(0.015)	(0.001)	(0.032)
$\mathcal{M}_5$	**Inf**	**−37.674**	**−52.891**	**−44.438**	**−52.877**	**−52.890**	**−52.890**
-	(NaN)	(0.644)	(0.001)	(0.290)	(0.020)	(0.001)	(0.042)
$\mathcal{M}_6$	**Inf**	**−94.707**	**−110.657**	**−101.416**	**−110.645**	**−110.656**	**−110.621**
-	(NaN)	(0.831)	(0.001)	(0.229)	(0.014)	(0.001)	(0.025)
$\mathcal{M}_7$	**Inf**	**−37.394**	**−58.002**	**−44.078**	**−57.985**	**−58.000**	**−57.984**
-	(NaN)	(0.622)	(0.002)	(0.212)	(0.025)	(0.002)	(0.071)

Table 5.1: Final batch mean estimates of the marginal likelihoods for each model, with MC errors given in brackets

Model	M.Carlo $\widehat{BF_{5j_{MC}}}$	H.Mean $\widehat{BF_{5j_{HM}}}$	G.H.Mean $\widehat{BF_{5j_{GHM}}}$	N.Raftery $\widehat{BF_{5j_{NR}}}$	L.Metropolis $\widehat{BF_{5j_{LM}}}$	B.Sampling $\widehat{BF_{5j_{BS}}}$	F.Integral $\widehat{BF_{5j_{FI}}}$
$\mathcal{M}_1$	**-**	**27.082**	**22.202**	**27.149**	**22.203**	**22.203**	**22.183**
$\mathcal{M}_2$	**-**	**96.468**	**92.166**	**96.481**	**92.172**	**92.167**	**92.101**
$\mathcal{M}_3$	**-**	**56.230**	**51.494**	**56.267**	**51.502**	**51.495**	**51.469**
$\mathcal{M}_4$	**-**	**27.913**	**28.560**	**27.953**	**28.560**	**28.560**	**28.541**
$\mathcal{M}_5$							
$\mathcal{M}_6$	**-**	**57.033**	**57.766**	**56.977**	**57.768**	**57.766**	**57.731**
$\mathcal{M}_7$	**-**	**0.280**	**5.110**	**-0.360**	**5.108**	**5.110**	**5.094**

Table 5.2: Bayes factors of each model in comparison to the true model $\mathcal{M}_5$, given on the logarithmic scale

Chapter 6

Dealing with Real-World Data

The previous chapters have shown that model evidence estimation methods can be very effective at accurately determining the marginal likelihood, a useful tool in model selection. However, real-world datasets often involve a a high number of variables which make full models very complex and challenging to work with. In practice, it is common to implement dimensionality reduction techniques to retain a smaller set of crucial variables for regression analysis, before proceeding with the standard model selection procedure.

There are many approaches we can take to address this dimensionality reduction problem. For example, techniques such as lasso regression introduced by Tibshirani (1996) or elastic net regression proposed by Zou & Hastie (2005) impose a penalty on the absolute size of the regression parameters $\boldsymbol{\beta}$, effectively performing variable selection by shrinking the less important parameters to zero. Alternatively, principal component analysis (PCA) is a technique developed by Pearson (1901) and Hotelling (1933) which aims to transform the original variables to a new set of uncorrelated variables known as the principal components, and to keep the principal components which hold the most variation in the data.

In this chapter, we will demonstrate a technique recommended independently by Kendall et al. (1965) and Hotelling (1957) called principal component regression (PCR). This is a two-stage process where PCA is performed on the set of explanatory variables to obtain a smaller set of crucial principal components, before performing regression analysis, with the new explanatory variables being the reduced set of principal components.

To demonstrate PCR, we will be using the Communities and Crime data set from the UCI machine learning repository Redmond (2009) which combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCRtotal, leading to 1994 observations of 128 variables. The goal here is to perform regression analysis, with the response variable being the number of violent crimes per 100,000 of the population. The data set is unclean, con-

taining missing values, as well as 4 variables which are not predictive for the response variable. To amend this, we work with a cleaned version of the data from the *mogavs* package in R, which removes the non-predictive variables and imputes the missing values using the *mclust* package. This leaves a clean data set with now 123 variables, one of which being the aforementioned response variable. The code used for this chapter can be found through this link.

6.1 Part 1 - Principal Component Analysis

The first step of PCR is the PCA. As previously mentioned, the goal in this step is to perform PCA on the set of explanatory variables. Therefore, we first compile the data of the explanatory variables into a 1994×122 matrix X, where the rows of X correspond to the observations and the columns of X correspond to the various explanatory variables of the regression problem. Essentially, when performing PCA, we will find a d-dimensional projection of X which preserves as much variance of the data as possible. In doing this, we find an orthonormal basis $\mathbf{v}_1, ..., \mathbf{v}_d$ of a d-dimensional space, where $\mathbf{v}_1, ..., \mathbf{v}_d$ are the principal components of the analysis. To perform the analysis, we must first center and scale X so that all of its columns have mean $= 0$ and variance $= 1$. We then use the R package *prcomp* to perform PCA on X. A summary of the PCA can be found in Table 6.1.

	Standard deviation	**Proportion of variance**	**Cumulative proportion**
$\mathbf{v}_1$	**8.801**	**0.635**	**0.635**
$\mathbf{v}_2$	**2.964**	**0.072**	**0.707**
$\mathbf{v}_3$	**2.661**	**0.058**	**0.765**
$\mathbf{v}_4$	**2.230**	**0.041**	**0.806**
$\mathbf{v}_5$	**1.778**	**0.026**	**0.832**
$\mathbf{v}_6$	**1.372**	**0.015**	**0.847**
$\mathbf{v}_7$	**1.331**	**0.015**	**0.862**
$\mathbf{v}_8$	**1.213**	**0.012**	**0.874**
$\mathbf{v}_9$	**1.149**	**0.011**	**0.884**
$\mathbf{v}_{10}$	**1.026**	**0.009**	**0.893**
$\mathbf{v}_{11}$	**0.997**	**0.008**	**0.901**
$\mathbf{v}_{12}$	**0.964**	**0.008**	**0.909**

Table 6.1: Table depicting the breakdown of variance explained by the first 12 principal components

The results show that over 90% of the variance in the data is explained by the first 11 principal components alone, with the first principal component containing 63.5% of the variance. To determine the amount of principal components that we want to keep for the regression analysis, we may choose a cut off value i.e. 80% and retain the smallest number of principal components that result in a cumulative proportion of variance greater than this cut off value. Alternatively, we could produce a scree plot of the variance proportions for each principal component in descending order. We then choose the cutoff point as the point after which the values tend to level off. This is also known as the "elbow" of the scree plot. Note that both of these methods aim to keep the principal components that explain the highest proportion of variance, which is a technique suggested in Hadi & Ling (1998). For the latter alternative, the corresponding scree plot from the PCA can be found in Figure 6.1, where there is a subtle "elbow" point at the sixth principal component, indicating that we should proceed with 6 of the principal components. When deciding on how many principal components to retain,

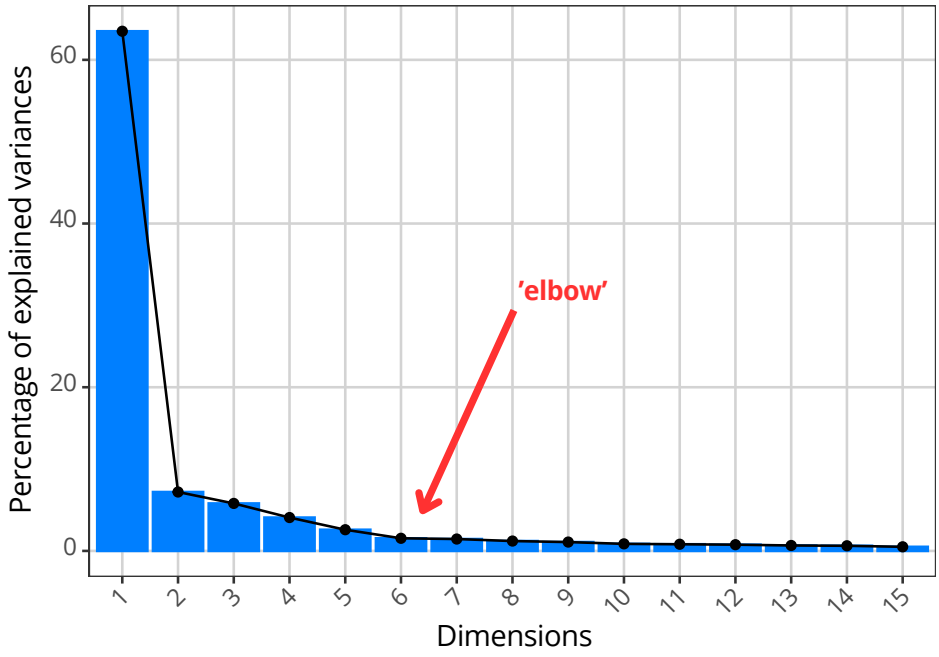

Figure 6.1: Scree plot depicting the percentage of explained variances in the first 15 principle components

there is a clear trade off between holding enough principal components to maintain the variability in the data, and reducing the dimensionality of the data so that we can minimise the complexity of further analysis. Choosing 6 principal components, as the scree plot suggests in Figure 6.1, will capture 84.7% of the data whilst effectively reducing the size of the explanatory data from 122 variables to just 6 components. This

demonstrates how we are able to strike a balance between retaining information and simplifying the data set, which is the main goal of PCA.

6.2 Part 2 - Performing Regression analysis

The second and final step of PCR is to perform regression analysis, using the 6 chosen principal components as the new explanatory variables in the regression problem. It is important to note that original explanatory variables are now incorporated into the 6 principal components, however they will not contribute equally. Essentially, the variables that explain the most variation in the data will be more heavily weighted in the principal components. We can visualise this by plotting the contributions of the original explanatory variables to the first two principal components v_1 and v_2. This plot can be

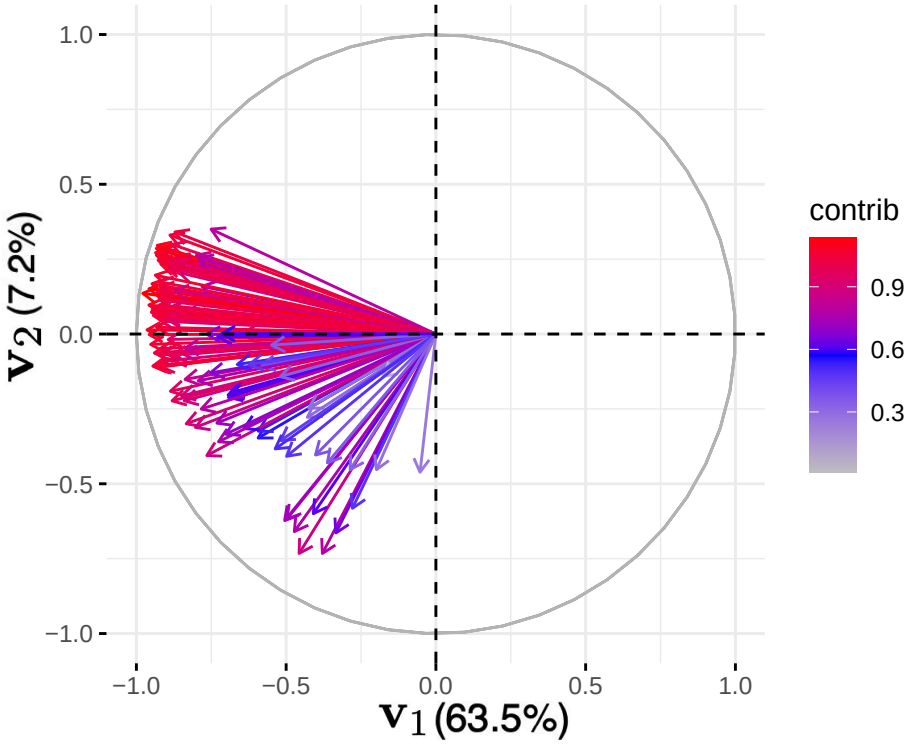

Figure 6.2: Visualisation of the data decomposed into the first 2 principal components

found in Figure 6.2 which shows that all of the original variables contribute negatively to v_1, whilst the loadings in v_2 split the variables into two distinct groups - those that contribute negatively to v_2 and those that contribute positively. A newer method by Zou et al. (2006) called Sparse PCA introduces a penalty, shrinking the loading of the less important variables in the principal components to zero, effectively reducing the amount

of variables that are incorporated at the end of the analysis.

Now that we are working with a more reasonable number of variables, it will be more computationally feasible to suggest regression models that fit the data and proceed with model selection methods that we have discussed in previous chapters.

We end the discussion with a final note on PCR. Throughout the chapter, we have assumed that the principal components that explain the most variance will be the most crucial in our regression problem. However, this may not always be the case. Moreover, since we only performed PCA on the explanatory variables in the first step, it is clear that there is no information from the response variable being used in this process. Kawano et al. (2018) addresses this issue through a one-stage procedure known as sparse principal component regression for generalized linear models (SPCR-glm) which introduces a basic loss function that combines the regression loss and the PCA loss.

Chapter 7

Conclusion

In this report, we explored the subject of Bayesian model selection and in particular the computation of the model evidence, which lies at the centre of this problem alongside the Bayes factor. We investigated a selection of methods aimed at estimating the model evidence. It is clear that some of the estimators such as the basic Monte Carlo estimator relied on simpler identities and formulas, which often resulted in less accurate outcomes when applied to the examples in Chapters 4 & 5. However, some of the methods discussed, such as the bridge sampling estimator, encompassed a more intensive set up but have led to estimates of a higher degree of accuracy. We also discussed some of the practical considerations that must be addressed when dealing with real-world data, highlighting techniques such as PCR and the newer sparse PCR, which can simplify complex data sets. The exploration of model evidence methods included a newer method in the field which utilised the Fourier transform in its set up. It is important to note that the continued emergence of newer methods highlights the dynamic and evolving nature of Bayesian model selection, where progress may lead to methods with increased accuracy and efficiency in the future.

Finally, for reference, the code used throughout this report to implement the methods in Chapters 4 & 5 as well as for PCR in chapter 6 the can be found through this link.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical transactions of the Royal Society of London*, ***53***, 370–418.

Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika*, ***76***(1), 183–183.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, ***90***(432), 1313–1321.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, ***96***(453), 270–281.

Dittrich, D., Leenders, R. T. A., & Mulder, J. (2019). Network autocorrelation modeling: A Bayes factor approach for testing (multiple) precise and interval hypotheses. *Sociological Methods & Research*, ***48***(3), 642–676.

Friel, N., & Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, ***70***(3), 589–607.

Friel, N., & Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, ***66***(3), 288–308.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal*, ***7***(1), 143–167.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, ***81***, 80–97.

Hadi, A. S., & Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, ***52***(1), 15–19.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo Methods*. Springer. (50–75)

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, ***24***(6), 417.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, ***10***(2), 69–79.

Jeffreys, H. (1961). *The Theory of Probability* (Third ed.). Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, ***90***(430), 773–795.

Kawano, S., Fujisawa, H., Takada, T., & Shiroishi, T. (2018). Sparse principal component regression for generalized linear models. *Computational Statistics & Data Analysis*, ***124***, 180–196.

Kendall, M. G., et al. (1965). *Course in Multivariate Analysis*.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of the American Statistical Association*, ***92***(438), 648–655.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, ***11***, 125–139.

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, ***56***(1), 3–26.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, ***2***(11), 559–572.

Redmond, M. (2009). Communities and Crime. *UCI Machine Learning Repository*.

Robert, C. P., & Wraith, D. (2009). Computational methods for Bayesian model choice. *American Insititue of Physics*, ***1193***, 251–262.

Rotiroti, F., & Walker, S. G. (2022). Computing marginal likelihoods via the Fourier integral theorem and pointwise estimation of posterior densities. *Statistics and Computing*, ***32***(5), 67.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Interntational Society for Bayesian Analysis*, ***1***(4), 833–859.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, ***58***(1), 267–288.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, ***81***(393), 82–86.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, ***67***(2), 301–320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, ***15***(2), 265–286.