ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Comparison of MCMC Methods for the Estimation of the Marginal Likelihood for Bayesian Model Evaluation

By

Konstantinos M. Perrakis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

January 2008

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Σύγκριση μεθόδων προσομοίωσης Monte Carlo με χρήση Μαρκοβιανών Αλυσίδων για την Εκτίμηση Περιθώριων Πιθανοφανειών για την Μπεϋζιανή Αξιολόγηση Μοντέλων

Κωνσταντίνος Μ. Περράκης

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιανουάριος 2008

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Ioannis Ntzoufras for his precious help and encouragement he generously offered me. With his guidance I explored a new field of statistics.

VITA

I was born in June 9th, 1978 in Athens where I got my High School Diploma. In 2002 I graduated from the Department of International and European Economic Studies of Athens University of Economics and Business. After a half a year of work in an enterprise as an account assistant and the fulfillment of my military service, I enrolled in the Master of Science in Statistics. I am currently working in a market research company as a statistical analyst.

ABSTRACT

Konstantinos Perrakis

Comparison of MCMC Methods for the Estimation of the Marginal Likelihood for Bayesian Model Evaluation

January 2008

Model selection is the final and perhaps the most significant stage of statistical inference. In this thesis we examine this aspect of statistical inference from a Bayesian perspective. Bayesian statistics are established on a theory whose origins date back to the 18th century, yet they were not broadly used for many years since in most of the cases researches had to confront intractable, high dimensional integrals. However, the evolution in computer technology and the advent of simulation methods made the implementation of Bayesian ideas practically feasible. In particular, the recent Markov Chain Monte Carlo methods proved to be powerful implementation tools for researchers.

In this thesis, we examine how Markov Chain Monte Carlo methods can advocate Bayesian analysis not only with respect to single model investigation but also with respect to model comparison. Formal Bayesian model comparison is based on the evaluation of the marginal likelihoods data of the models under comparison. Through these conditional probabilities, one can subsequently calculate quantities such as the Bayes Factors and the Posterior Odds of competing models. Our aim is to present and evaluate several simulation-based methods that intend to estimate the marginal likelihood.

ΠΕΡΙΛΗΨΗ

Περράκης Κωνσταντίνος

Σύγκριση μεθόδων προσομοίωσης Monte Carlo με χρήση Μαρκοβιανών Αλυσίδων για την Εκτίμηση Περιθώριων Πιθανοφανειών για την Μπεϋζιανή Αξιολόγηση Μοντέλων

Ιανουάριος 2008

Η επιλογή μοντέλου είναι το τελικό και ίσως το σημαντικότερο στάδιο της στατιστικής συμπερασματολογίας. Σε αυτήν την εργασία εξετάζουμε την επιλογή μοντέλου από την Μπεϋζιανή σκοπιά. Παρότι οι αρχές της Μπεϋζιανής θεωρίας ανάγονται στον 18^{ο} αιώνα, η στατιστική κατά Bayes δεν είχε ευρεία διάδοση για αρκετά χρόνια, καθώς στις περισσότερες περιπτώσεις οι ερευνητές ήταν αναγκασμένοι να αντιμετωπίσουν δυσεπίλυτα ολοκληρώματα υψηλής τάξης. Όμως η ανάπτυξη των υπολογιστών και η εμφάνιση μεθόδων προσομοίωσης έκανε την εφαρμογή της Μπεϋζιανής θεωρίας πρακτικά δυνατή. Ειδικότερα, οι πρόσφατες μέθοδοι προσομοιώσης Monte Carlo με την χρήση Μαρκοβιανών αλυσίδων (MCMC) συνέβαλλαν αποφασιστικά στην διεκπεραίωση της Μπεϋζιανής ανάλυσης.

Στην παρούσα διατριβή εξετάζουμε πως οι μέθοδοι MCMC μπορούν να υποβοηθήσουν την Μπεϋζιανη ανάλυση όχι μόνο όσον αφορά στην διερεύνηση μεμονωμένων μοντέλων, αλλά και στην σύγκριση μεταξύ μοντέλων. Η κατά Bayes σύγκριση μοντέλων βασίζεται στον υπολογισμό των περιθώριων πιθανοφανειών των παρατηρούμενων δεδομένων των υπο σύγκριση μοντέλων. Μέσω αυτών των δεσμευμένων πιθανοτήτων μπορούν να υπολογιστούν οι παράγοντες Bayes (Bayes Factors) και οι εκ των υστέρων λόγοι σχετικών πιθανοτήτων (Posterior Odds) μεταξύ των συγκρινόμενων μοντέλων. Στόχος της παρούσας διατριβής είναι η παρουσίαση και η αξιολόγηση συγκεκριμένων μεθόδων προσομοιώσης που αποσκοπούν στην εκτίμηση των περιθωρίων πιθανοφανειών.

CONTENTS

Chapter 1: Introduction 1
1.1 Purpose of the thesis 1
1.2 Structure of the thesis 1
Chapter 2: The Bayes Approach 3
2.1 Introduction 3
2.2 The Bayes' Theorem 4
2.3 Prior distribution 5
2.3.1 Elicited Priors 6
2.3.2 Conjugate Priors 6
2.3.3 Non-informative Priors 7
2.4 Summarizing posterior information 8
2.4.1 Location and dispersion measures 9
2.4.2 Credible sets 9
2.5 Predictive distribution 10
2.6 Bayesian p-values 11
2.7 Posterior Odds and Bayes Factors 13
2.8 Information Criteria 15
2.9 Conclusion 17
Chapter 3: Markov Chain Monte Carlo 19
3.1 Introduction 19
3.2 Asymptotic methods 20
3.3 Monte Carlo Integration 21
3.4 Markov Chain Monte Carlo Algorithms 23
3.4.1 Markov chains 23
3.4.2 The Metropolis and Metropolis-Hastings Algorithms 24
3.4.3 The Gibbs Sampler 29
3.4.4 The Metropolis within Gibbs Algorithm 31
3.4.5 Convergence Diagnostics 33
3.5 A linear regression example with normal data 35
3.5.1 M-H algorithm implementation 36
3.5.2 Gibbs Sampler implementation 44
3.5.3 Informal model checking through plots of predicted values 46
3.6 A logistic regression example with binomial data 47

3.6.1 Metropolis algorithm implementation 49
3.6.2 Metropolis within Gibbs implementation 52
3.6.3 Checking model discrepancies through test quantities 55
Chapter 4: Marginal Likelihood Estimators 59
4.1 Introduction 59
4.2 Harmonic mean estimator 60
4.3 Laplace-Metropolis estimator 61
4.4 Newton and Raftery's estimator 62
4.5 Bridge sampling estimator 63
4.6 Candidate's estimator 65
4.6.1 Marginal likelihood from the Gibbs output 66
4.6.2 Marginal likelihood from the Metropolis-Hastings output 68
4.7 Chen's estimator 69
Chapter 5: Illustration and Comparison of Methods in a Simple Regression Example 73
5.1 Models and prior selection 73
5.2 Simulating from the posterior 74
5.2.1 Details of the Gibbs sampler implementation 74
5.2.2 Details of the Metropolis-Hastings implementation 77
5.3 Implementation of the methods 81
5.3.1 Harmonic Mean Estimator 81
5.3.2 Laplace-Metropolis Estimator 82
5.3.3 Newton and Raftery Estimator 83
5.3.4 Bridge Sampling Estimators 84
5.3.5 Candidate's estimators 86
5.3.5.1 The Chib Estimator 87
5.3.5.2 The Chib and Jeliazkov Estimator 91
5.3.6 Chen estimator 93
5.4 Comparing the models 93
5.5 Comparing results 95
5.6 Summary 103
Chapter 6: Conclusions and Further Discussion 105
6.1 Conclusions concerning the marginal likelihood estimators 105
6.2 Further discussion 107
REFERENCES 109

LIST OF TABLES

2.1 Interpretations for Bayes Factors and for logBF 15
2.2 Interpretations for Bayes Factors and for 2lnBF 15

3.1 DC output and wind velocity observations 36
3.2 Initial M-H estimates of posterior means and standard deviations 41
3.3 M-H estimates of posterior means and standard deviations 43
3.4 M-H estimates of posterior quantiles and calculated R-roots 43
3.5 Gibbs estimates of posterior means and standard deviations 45
3.6 Gibbs estimates of posterior quantiles and calculated R-roots 45
3.7 Number of leuchaimia deaths and dose of radiation 48
3.8 Metropolis estimates of posterior means and standard deviations 51
3.9 Metropolis estimates of posterior quantiles and calculated R-roots 51
3.10 Metropolis within Gibbs estimates of posterior means and standard deviations 53
3.11 Metropolis within Gibbs estimates of posterior quantiles and calculated R-roots 53

5.1 Posterior summary for model 0 obtained from Gibbs sampling 76
5.2 Posterior summary for model 1 obtained from Gibbs sampling 76
5.3 Posterior summary for model 2 obtained from Gibbs sampling 77
5.4 Posterior summary for model 3 obtained from Gibbs sampling 77
5.5 Posterior summary for model 0 obtained from M-H simulation 79
5.6 Posterior summary for model 1 obtained from M-H simulation 80
5.7 Posterior summary for model 2 obtained from M-H simulation 80
5.8 Posterior summary for model 3 obtained from M-H simulation 80
5.9 True marginal and log-marginal likelihood values and posterior probabilities for 4 regression models 94
5.10 Pairwise comparisons of 4 regression models based on the true values of 2lnBF 94
5.11 Estimates of AIC, BIC and DIC for 4 regression models 95
5.12 Marginal likelihood estimates of 4 regression models 96
5.13 Posterior probability estimates of 4 regression models 97

5.14 Batched marginal likelihood estimates and MC errors for 4 regression models 98
5.15 Estimates of 2lnBF of model 2 versus model 3 101
5.16 Batched estimates of 2lnBF of model 2 vesus model 3, MC errors and 95% percentile condifence intervals 102

LIST OF FIGURES

3.1 Histogram of DC output and scatter plot of DC output and logarithm of wind velocity 36
3.2 Time series plots, autocorrelation plots and histograms resulting from a M-H simulation 41
3.3 Ergodic mean plots and histograms resulting from M-H simulation 43
3.4 Ergodic mean plots and histograms resulting from Gibbs sampling 45
3.5 Scatter plots of posterior draws from M-H and Gibbs simulations 46
3.6 Histograms of 20 replicated DC output datasets 47
3.7 Ergodic mean plots and histograms resulting from Metropolis simulation ... 51
3.8 Ergodic mean plots and histograms resulting from Metropolis within Gibbs simulation 54
3.9 Scatter plots of posterior draws from Metropolis and Metropolis within Gibbs simulations 55
3.10 Kernel smoothed densities of chi-square and deviance test quantities for observed and replicated data 57

5.1 Ergodic mean plots of 4 marginal likelihood estimates for model 0 99
5.2 Ergodic mean plots of 4 marginal likelihood estimates for model 1 99
5.3 Ergodic mean plots of 4 marginal likelihood estimates for model 2 100
5.4 Ergodic mean plots of 4 marginal likelihood estimates for model 3 100

Chapter 1: Introduction

1.1 Purpose of the thesis

Model selection is among the dominant issues in statistical analysis. This thesis is concerned with this issue from a Bayesian viewpoint. The formal Bayesian approach towards model selection is based on the calculation of marginal likelihoods, through which we can evaluate the Bayes Factors, the posterior probabilities and the posterior odds of competing models.

Unfortunately, direct calculation of marginal likelihoods is in most of the cases cumbersome or even impossible, since it requires analytic solutions of high dimensional integrals. As we will see Monte Carlo and subsequent Markov Chain Monte Carlo methods offer the most trustworthy estimates of high dimensional integrals. The main purpose of this thesis is to present and evaluate the available simulation-based estimators of the marginal likelihood.

1.2 Structure of the thesis

Chapter 2 focuses on the main aspects of Bayesian theory. In the first sections the reader is familiarized with the concepts of the prior and the posterior distribution which form the basis of Bayesian theory. In the following sections we present in brief the main inferential tools most commonly used in Bayesian analysis.

The subject of chapter 3 is Markov Chain Monte Carlo. The concepts of Monte Carlo integration and Markov chains are succinctly described; a thorough theoretical investigation of these fields is not provided since it would exceed the context of this thesis. Instead, attention is drawn to the basic Monte Carlo Markov Chain algorithms; the Metropolis-Hastings algorithm, the Gibbs sampler and the Metropolis within Gibbs algorithm. Theoretical aspects of the aforementioned algorithms are examined and their use is described in detail. We

then implement these algorithms in two regression problems, the first is a normal linear regression example and the second a logistic regression example.

In chapter 4 a review of the simulation-based marginal likelihood estimation methods is given. Each method is described in brief and practical implementation issues are discussed.

Implementation and evaluation of the methods is the main topic of chapter 5. The first data set presented in chapter 2 is re-examined by taking into account four competing regression models. Metropolis-Hastings and Gibbs sampling simulations are utilized in order to acquire posterior samples and posterior summaries for each model. The competing models are then compared based on the true marginal likelihood values. Finally, we present the results obtained from each estimation method and compare the estimates with the true values of the marginal likelihoods, the posterior probabilities and the Bayes Factors.

Conclusions are summarized in chapter 6. Each method is evaluated according to the corresponding results and from the overall implementation experience. Further discussion regarding alternative model selection approaches is also provided.

All computations and plots presented in this thesis were carried out in the R programming language, version 2.5.

Chapter 2: The Bayes Approach

2.1 Introduction

Bayesian theory is based on the original 1763 paper of Rev. Thomas Bayes, an English minister and mathematician. In this paper, inference for the parameters of a Binomial distribution is achieved by conditioning on the data which are also Binomial. The area generated some interest by Gauss, Laplace and other mathematicians of the time. In 1774 Laplace presented the general form of the Bayes theorem.

Unfortunately, most of the early 20th century statisticians ignored this field of work. In 1939, a physicist named Harold Jeffreys reintroduced Laplace's work. Jeffreys, along with the econometrician Arthur Bowley, argued on behalf of Bayesian ideas during this period. Bayesian methods achieved recognition only after 1950, when many statistical researchers began to advocate this methods as remedies for certain deficiencies of the classical or frequentistic approach, such as the interpretation of the classical confidence interval in any single data experiment or the violation of the Likelihood Principle (Carlin and Louis, 1996, p.2-5).

The main difference between the classical approach and the Bayesian approach is rather philosophical in nature. The classical approach assumes a probability distribution (likelihood) for the data and considers the unknown parameters as fixed. The frequentistic uncertainty originates from the repetition of samples, so the evaluation procedures are based on repeated sampling, imagining an infinite replication of the same inferential problem for fixed values of the unknown parameters. The Bayesian approach considers the unknown parameters as random variables, so it assumes a sampling distribution (likelihood) along with a *prior distribution* for the parameters. The Bayesian uncertainty comes from the parameters, so the evaluation procedure is based on an infinite sampling experiment of parameters drawn from the distribution which is conditional on the data that is, the parameters posterior distribution. In

general, a frequentist conditions on the parameters and then replicates over the data, while a Bayesian conditions on the data and then replicates over the parameters (Carlin and Louis, 1996, p.6).

2.2 The Bayes' Theorem

The Bayesian approach specifies for observed data $\mathbf{y} = (y_1, y_2, ..., y_n)$ a sampling distribution $p(\mathbf{y} \mid \boldsymbol{\theta})$, that is the likelihood of the data given the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$. The parameter vector $\boldsymbol{\theta}$ is also considered as a random quantity having a prior distribution $p(\boldsymbol{\theta})$. This is the distribution of $\boldsymbol{\theta}$ *before* the data are observed. The joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ can be expressed as a product of these two densities

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

From the basic properties of conditional probabilities we obtain the distribution of $\boldsymbol{\theta}$ given the data $\mathbf{y}$

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

where $p(\mathbf{y})$ is the *marginal likelihood* of the data given by $p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$ if $\boldsymbol{\theta}$ is discrete or by $p(\mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ if $\boldsymbol{\theta}$ is continuous. This formula is known as the *Bayes' Theorem* and $p(\boldsymbol{\theta} \mid \mathbf{y})$ denotes the posterior distribution of $\boldsymbol{\theta}$, that is the distribution of $\boldsymbol{\theta}$ *after* observing the data.

All Bayesian inference about $\boldsymbol{\theta}$ is based on the posterior distribution which contains information from both the experimental data and the prior beliefs about $\boldsymbol{\theta}$. It should be noted that the prior and posterior distributions are always relative to the observations considered at a given moment (Gamerman and Lopes, 2006, p.44); after observing $\mathbf{y}$ and obtaining the posterior, new observations

$\mathbf{y}^{new}$, related to $\boldsymbol{\theta}$ through an eventually different likelihood function, could become available. Then, the posterior of $\mathbf{y}$ can be considered as the prior for $\mathbf{y}^{new}$ and we can obtain the posterior of $\mathbf{y}^{new}$ by a new application of the Bayes' theorem.

The marginal likelihood provides the expected distribution of $\mathbf{y}$ as $p(\mathbf{y}) = E\left[p(\mathbf{y}\,|\,\boldsymbol{\theta})\right]$ and the expectation is taken with respect to the prior distribution of $\boldsymbol{\theta}$. It is also referred to as the *integrated likelihood* or as the *prior predictive* distribution. Marginal likelihood probabilities are of great importance, since they are required for the calculation of *Bayes Factors* (see section 2.7). According to Kass and Raftery (1995), $p(\mathbf{y})$ can be interpreted as the predictive probability of the data; that is, the probability of seeing the data that were actually observed, calculated *before* any data were available.

Since, $p(\mathbf{y})$ does not actually depend on $\boldsymbol{\theta}$, we can acquire the unnormalized posterior distribution from

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The above expression is often useful in Bayesian statistics, since it can significantly simplify the calculation of the posterior distribution.

2.3 Prior distribution

Determination of the prior distribution has vital importance in Bayesian statistics. Prior distributions are usually specified from information accumulated from past studies or from the opinions of subject area experts. When there is little or no available information about the parameters in question, then *vague* or *non-informative* distributions are adopted (Carlin and Louis, 1996, p.27-37).

2.3.1 Elicited Priors

A rational approach in specifying $p(\boldsymbol{\theta})$ is to initially distinguish the values of $\boldsymbol{\theta}$ which are deemed as 'possible to occur' and then to assign point masses which sum up to one in a way that reflects the prior beliefs. When $\boldsymbol{\theta}$ is continuous probability masses are assigned in intervals instead of points, resulting in a histogram prior for $\boldsymbol{\theta}$. Of course, this approach can be time consuming especially when $\boldsymbol{\theta}$ is multivariate. A simpler solution is to assume that the prior density belongs to a parametric distributional family $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, choosing $\boldsymbol{\eta}$ so that the resulting distribution expresses prior beliefs as nearly as possible.

2.3.2 Conjugate Priors

When the prior is a known distribution of the form $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, then some choices of $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ are more convenient for the calculation of the posterior distribution than others. More specifically, we may choose a member of the distributional family which is *conjugate* to the likelihood $p(\mathbf{y}|\boldsymbol{\theta})$. This results to a posterior distribution which belongs to the same distributional family with the prior. Obviously, the use of conjugate priors simplifies considerably the computation of the posterior.

An important result presented by Morris (1983) is that distributions belonging to the exponential family always have a conjugate prior. In most of the cases, the sampling distribution is drawn from the exponential family therefore conjugate priors are broadly used. Additional information on the conjugate property of distributions belonging to the exponential family can be found in Consonni and Veronese (1992) and Gutierrez-Pena and Smith (1995).

When the use of a single conjugate prior is not adequate in terms of expressing prior beliefs accurately enough, then a mixture of conjugate prior distributions may be used in order to improve the accuracy. Mixtures of conjugate priors are more flexible and still simplify calculations. As shown in

Dalal and Hall (1983) a mixture of conjugate priors leads to a mixture of posteriors.

2.3.3 Non-informative Priors

In many cases there is no reliable prior information concerning $\boldsymbol{\theta}$ or objective inference based solely on the data is desired. In these cases the prior density $p(\boldsymbol{\theta})$ should contain no information about $\boldsymbol{\theta}$ in the sense that no value of $\boldsymbol{\theta}$ should be favored over another. Such priors are called vague or non-informative.

When the parameter space is discrete and finite, that is $\theta \in \boldsymbol{\Theta} = \{\theta_1, \theta_2, ..., \theta_n\}$ then the distribution

$$p(\theta_i) = \frac{1}{n}, \quad i = 1, 2, ...n,$$

is obviously non-informative since all values of θ are equally probable.

When the parameter space is continuous and bounded, say $\theta \in \boldsymbol{\Theta} = [a, b]$, with $-\infty < a < b < \infty$ then a non-informative prior is given by the uniform distribution

$$p(\theta) = \frac{1}{b-a}, \quad a < \theta < b.$$

In the case of an unbounded parameter space like $\boldsymbol{\Theta} = (-\infty, \infty)$ the appropriate uniform prior has the form

$$p(\theta) = c, \quad c > 0.$$

This distribution is improper, since $\int p(\theta) d\theta = \infty$. Hence its use as a prior seems inappropriate. But if the integral of the likelihood with respect to θ is finite then the resulting posterior distribution is proper, so inference is still feasible.

One drawback of the uniform distribution is that it is not invariant to reparametrization. This means that $p(\theta)$ may be non-informative for θ, but $p(\gamma)$ may be informative to γ, where $\gamma = g(\theta)$. A solution to this is the use of the *Jeffreys'* non-informative prior which is invariant to transformations (Jeffreys, 1961). Jeffreys' prior has the form

$$p(\boldsymbol{\theta}) = \left|I(\boldsymbol{\theta})\right|^{1/2},$$

where $I(\boldsymbol{\theta})$ is the expected Fisher information matrix, having ij – element

$$I_{ij}(\boldsymbol{\theta}) = -E_{\mathbf{X}|\boldsymbol{\Theta}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\mathbf{x}\,|\,\boldsymbol{\theta})\right].$$

Calculating $I(\boldsymbol{\theta})$ can be cumbersome in high dimensional problems, so the common approach is to obtain a Jeffreys' prior for each parameter individually and then form the joint prior from the product of the individual priors. In addition, Jeffreys' prior seems to face difficulties in multi-parameter problems where only a subset or one or more parametric functions of the parameter vector $\boldsymbol{\theta}$ are of inferential interest and the remaining are nuisance parameters. Information over the use of Jeffreys' prior for Generalized Linear Models (GLM) can be found in Ibrahim and Laud (1991).

Bernando (1979) introduced the so-called *reference prior* approach for deriving non-informative priors in multiparameter cases by splitting the parameter vector into parameters of interest and nuisance parameters. Information over reference priors can be found in the paper of Berger and Bernando (1992) who extended further the idea of Bernando (1979). Alternative proposals in constructing reference priors were presented by Ghosh and Mukerjee (1992); comparison of different constructing approaches for reference priors can be found in the papers of Datta and Ghosh (1995, 1996). Additional information regarding the use and selection of non-informative priors is provided by Kass and Wasserman (1996).

2.4 Summarizing posterior information

After obtaining the posterior, it is meaningful to summarize the information provided by it. For this, we rely mainly in certain location and dispersion measures. In addition, *credible sets* or *credible intervals* of the parametric space are often presented.

2.4.1 Location and dispersion measures

Location and dispersion measures provide an image of the possible central values and of the variability of the posterior distribution, respectively.

The common choices for the location measures are the mean, the mode and the median of the posterior distribution. These measures correspond respectively to the expected value of $\boldsymbol{\theta}$, the most likely value of $\boldsymbol{\theta}$ and the value of $\boldsymbol{\theta}$ which divides the parametric space in two equal probability parts (Gamerman and Lopes, 2006, p.47). When the posterior is symmetric the mean and median will be identical; for symmetric and unimodal posterior distributions all three measures will coincide. In the case of an asymmetric posterior the median is often preferred since it is intermediate to the mode and the mean (Carlin and Louis, 1996, p.39). Most often, the mode will be numerically harder to find, especially when $\boldsymbol{\theta}$ is multivariate. As a result, the posterior mode is usually approximated through the use of maximization algorithms.

The main dispersion measures are the posterior variance, the standard deviation, the precision, the interquartile range and the curvature at the posterior mode. When $\boldsymbol{\theta}$ is multivariate the variance is given by the posterior covariance matrix; in this case the standard deviation is the vector of square roots of the diagonal elements of the covariance matrix. Posterior precision is given by the inverse of the covariance matrix, while the curvature at the mode is given by the matrix of the second derivatives of $-\log p(\boldsymbol{\theta} \mid \mathbf{y})$ evaluated at the posterior mode (Gamerman and Lopes, 2006, p.47).

2.4.2 Credible sets

The Bayesian analogue of a frequentist confidence interval is called a credible set. According to Carlin and Louis (1996, p.42) a $100\times(1-a)\%$ credible set for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a subset $C \subseteq \boldsymbol{\Theta}$ such that

$$1-a \leq \Pr(C \mid \mathbf{y}) = \int_C p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta},$$

where integration is replaced by summation over discrete components of $\boldsymbol{\theta}$. In contrast to the classical confidence interval interpretation this definition enables direct probability statements about the likelihood of $\boldsymbol{\theta}$ falling in C. The interpretation of this credible set is

"The probability that $\boldsymbol{\theta}$ lies in *C* given the observed data is at least $(1-a)$".

When the posterior is asymmetric or multimodal then it would be preferable to obtain the *Highest Posterior Density* (HPD) credible set, which groups together the "most likely" values of $\boldsymbol{\theta}$ and hence is narrower than the equal tail credible set. Yet, obtaining the HPD credible set is not straightforward, since it requires solving iteratively a non-linear equation. Wright (1986) presented an iterative method for univariate cases; Ghosh and Mukerjee (1995) and Hyndman (1996) introduced iterative solutions for multivariate cases.

2.5 Predictive distribution

An important issue in statistical analysis is the ability to make inference about future observations. In Bayesian statistics this comes naturally from the use of the *predictive distribution*.

Suppose that y_{n+1} is a future observation independent of $\mathbf{y} = (y_1, y_2, ..., y_n)$ conditional on $\boldsymbol{\theta}$. Then the distribution of $y_{n+1} \mid \mathbf{y}$ is given by

$$p(y_{n+1} \mid \mathbf{y}) = \int p(y_{n+1}, \boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \Leftrightarrow$$
$$p(y_{n+1} \mid \mathbf{y}) = \int p(y_{n+1} \mid \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \Leftrightarrow$$
$$p(y_{n+1} \mid \mathbf{y}) = \int p(y_{n+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$$

since y_{n+1} and $\mathbf{y}$ are conditionally independent.

The predictive distribution provides information for new observations given the likelihood, the prior and the data observed. It is also referred to as the *posterior predictive* distribution, in contrast to the marginal likelihood (the prior predictive distribution), since it is the expected distribution of a future

observation y_{n+1} as $p(y_{n+1}|\mathbf{y}) = E\left[p(y_{n+1}|\boldsymbol{\theta})\right]$ and the expectation is taken now with respect to the *posterior* distribution of $\boldsymbol{\theta}$.

The predictive distribution forms the basis of the *predictive inference* within the Bayesian paradigm. According to the predictive approach, inference about parameters is not possible since they are not observed. In contrast, the predictive distribution is defined in terms of *observable* values of the dependent variable and seems to be the natural instrument for decisions concerning model adequacy and model selection.

Several model selection methods and criteria have been generated by this approach; see Geisser and Eddy (1979), Laud and Ibrahim (1994), Greenberg and Parks (1997) and Gelfand and Ghosh (1998). These methods can be viewed as alternative options to *Bayes Factors* which are the formal Bayesian approach regarding the issue of model selection; see section 2.7.

2.6 Bayesian p-values

Bayesian p-values or *posterior predictive p-values* are based on posterior predictive checks. Posterior predictive checks are used to evaluate the fit of a model and are in fact generalizations of the classical tests in that they average over the posterior distribution rather than fixing the unknown parameter at some point $\hat{\boldsymbol{\theta}}$ (Gelman et al., 1993). To evaluate the fit of a model we compare the observed data $\mathbf{y}$ to the predictive or replicated data $\mathbf{y}^{rep}$ drawn from the predictive distribution. The discrepancy between observed and expected data is measured through *test quantities* $T(\mathbf{y},\boldsymbol{\theta})$ which can be functions of the unknown parameters as well as the data. When the distribution of the test quantity is free of $\boldsymbol{\theta}$, then $T(\mathbf{y},\boldsymbol{\theta}) \equiv T(\mathbf{y})$ is a pivotal quantity and the Bayesian p-value concurs with the frequentist p-value (Gelman et al., 1993).

The Bayesian p-value is defined by Rubin (1984) as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity

$$\text{Bayes p-value} = \Pr\left(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}\right),$$

where the probability is defined over the posterior distribution of $\boldsymbol{\theta}$ and the predictive distribution of $\mathbf{y}^{rep}$, thus

$$\text{Bayes p-value} = \int\int I_{\left(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})\right)} p(\mathbf{y}^{rep} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} d\mathbf{y}^{rep},$$

where $I_{\left(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})\right)}$ is the indicator function given by

$$I_{\left(T(\mathbf{y}^{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta})\right)} = \begin{cases} 1, & T(\mathbf{y}^{rep}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta}) \\ 0, & T(\mathbf{y}^{rep}, \boldsymbol{\theta}) < T(\mathbf{y}, \boldsymbol{\theta}). \end{cases}$$

It should be noted that Bayesian p-values serve *only* as measures of discrepancy between the assumed model and the observed data, providing information concerning model adequacy, and they should not be compared across models (Carlin and Louis, 1996, p.57).

Since Bayesian p-values are in fact measures of discrepancies, there are no general rules when choosing a test quantity. According to Gelman et al. (1993) the choice of test quantities should reflect our inferential interests; a test quantity can be any function of the data alone or of the data along with the unknown parameters which can possibly reveal discrepancies between observed and replicated data. Moreover, we can evaluate more than one discrepancy measures and judge the fit of a specific model from different perspectives. Nevertheless, general goodness-of-fit discrepancy measures are useful for routine checks of the overall fitness. Such a measure recommended by Gelman et al. (1993) is the chi-square discrepancy quantity given by

$$T(\mathbf{y}, \boldsymbol{\theta}) = \sum_i \frac{(y_i - E(y_i \mid \boldsymbol{\theta}))^2}{Var(y_i \mid \boldsymbol{\theta})},$$

where the summation is over the sample observations.

A model is considered suspect if the Bayesian p-value for a meaningful test quantity is near to 0 or 1. According to Gelman et al. (1995, p.173) major failures of a model, corresponding to tail area probabilities less than 0.01 or more than 0.99, can be addressed by expanding the model in an appropriate way. Lesser failures might also suggest model improvement or might be ignored if the main inferences are not affected.

For more information over Bayesian p-values see Gelman et al. (1993) with the associated comments and Meng (1994).

2.7 Posterior Odds and Bayes Factors

In the Bayesian framework hypothesis testing is strongly associated with model selection. There is no constrain to the number of hypotheses that may be simultaneously considered, so we usually switch notation from "hypotheses" H_i to "models" M_i, $i = 1,2,...,k$ (Carlin and Louis, 1996, p.47). Model selection and hypothesis testing are based on posterior probabilities, posterior odds and Bayes factors. Bayesian methods provide flexibility in hypothesis testing; according to Kass and Raftery (1995) we can evaluate evidence in favor of the null hypothesis, compare non-nested models, draw inferences by taking into account model uncertainty and determine which competing model provides better predictive results.

Consider two competing models M_0 and M_1, each with a corresponding parameter vector $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. These models specify the distribution of the data $p(\mathbf{y} \mid M_i)\left(\equiv p(\mathbf{y} \mid \boldsymbol{\theta}_i, M_i)\right)$, with $i = 0,1$. In addition, each model M_i has a prior probability $p(M_i)$, with $i = 0,1$ and $p(M_0) + p(M_1) = 1$. From Bayes theorem the posterior probability of a model is given by

$$p(M_i \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_i)p(M_i)}{p(\mathbf{y} \mid M_1)p(M_1) + p(\mathbf{y} \mid M_2)p(M_2)},$$

for $i = 0,1$.

The *posterior odds* PO_{01} of model M_0 versus model M_1 is given by

$$PO_{01} = \frac{p(M_0 \mid \mathbf{y})}{p(M_1 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid M_0)}{p(\mathbf{y} \mid M_1)} \times \frac{p(M_0)}{p(M_1)}.$$

The quantity $BF_{01} = \frac{p(\mathbf{y} \mid M_0)}{p(\mathbf{y} \mid M_1)}$ is called *Bayes factor* of model M_0 versus model M_1, while $\frac{p(M_0)}{p(M_1)}$ corresponds to the *prior odds* of model M_0 versus model M_1. Thus, we have that

$$\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds}.$$

The distribution $p(\mathbf{y} \mid M_i)$ is the marginal likelihood of the data, discussed in section 2.2, conditional on the model. By taking into account the dependence from the model, $p(\mathbf{y} \mid M_i)$ is given by

$$p(\mathbf{y} \mid M_i) = \int_{\boldsymbol{\theta}_i} p(\mathbf{y} \mid \boldsymbol{\theta}_i, M_i) p(\boldsymbol{\theta}_i \mid M_i) d\boldsymbol{\theta}_i.$$

The above model comparison can be extended to more than two competing models. Suppose we have $K+1$ competing models $M_0, M_1, M_2, ..., M_K$. Each model $M_1, M_2, ..., M_K$ is compared in turn with M_0, yielding Bayes factors $BF_{10}, BF_{20}, ..., BF_{K0}$. Then the posterior probability of the model M_i, for $i = 0, ..., K$, is given by

$$p(M_i \mid \mathbf{y}) = \frac{a_i BF_{i0}}{\sum_{r=1}^{K} a_r BF_{r0}}.$$

The term $a_r = p(M_r) / p(M_0)$, with $r = 0, ..., K$, is the prior odds of the corresponding model M_r against model M_0, with $BF_{00} = a_0 = 1$.

When trying to make inference about a quantity of interest which is well defined for every model we can deal with model uncertainty by using the posterior model probabilities as weights (Kass and Raftery, 1995). According to the authors this technique, known as *model averaging*, yields consistently and substantially better predictions than the methods based on individual models; for more information on model averaging see Hoeting et al. (1999) and Raftery et al. (1997).

Interpretations of Bayes factors provided by Kass and Raftery (1995) are given in Tables 2.1 and 2.2. According to the authors the categories presented in Table 2.2 seem to furnish appropriate guidelines for most of the cases.

$\log_{10} BF_{10}$	BF_{10}	Evidence Against M_0
0 to 0.5	1 to 3.2	Not worth than a bare mention
0.5 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
Greater than 2	Greater than 100	Decisive

Table 2.1 *Interpretations for Bayes factors and for the common logarithm of Bayes factors.*

$2 \ln BF_{10}$	BF_{10}	Evidence Against M_0
0 to 2	1 to 3	Not worth than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
Greater than 10	Greater than 150	Very Strong

Table 2.2 *Interpretations for Bayes factors and for twice the natural logarithm of Bayes factors.*

2.8 Information Criteria

An alternative and often easier solution when comparing different models is through the use of information criteria. The most popular criteria are the

Bayesian version of *Akaike's Information Criterion* (AIC) (Akaike, 1974), the *Bayes Information Criterion* (BIC) (Schwarz, 1978) also known as the *Schwarz Criterion* and the most recent *Deviance Information Criterion* (DIC) (Spiegelhalter et al., 2002). All of these information criteria are based on the evaluation of the *deviance*. The deviance of model m is defined as $D(\boldsymbol{\theta}_m) = -2\log p(\mathbf{y} \mid \boldsymbol{\theta}_m)$.

Akaike's information criterion (AIC, Akaike, 1974) is defined as

$$AIC(m) = D(\hat{\boldsymbol{\theta}}_m) + 2d_m,$$

where $D(\hat{\boldsymbol{\theta}}_m)$ is the minimum value of the deviance of model m and d_m is the number of estimated parameters.

The Bayesian information criterion (BIC, Schwarz, 1978) is estimated by

$$BIC_m = D(\hat{\boldsymbol{\theta}}_m) + d_m \log(n),$$

where n is the number of observations. The BIC is also used as a rough approximation to the logarithm of the Bayes factor; see Kass and Raftery (1995). Both AIC and BIC penalize for the number of parameters and in general tend to choose the less complex models. According to Brooks (2002) Bayesian variations of AIC and BIC based on posterior summaries of the deviance are given by

$$AIC(m)_{\bar{D}} = \overline{D(\boldsymbol{\theta}_m)} + 2d_m,\ BIC(m)_{\bar{D}} = \overline{D(\boldsymbol{\theta}_m)} + d_m \log(n)$$

and

$$AIC(m)_{D(\bar{\theta})} = D(\bar{\boldsymbol{\theta}}_m) + 2d_m,\ BIC(m)_{D(\bar{\theta})} = D(\bar{\boldsymbol{\theta}}_m) + d_m \log(n).$$

The term $\overline{D(\boldsymbol{\theta}_m)}$ is the posterior expectation of the deviance which actually summarizes the fit of model m, while the term $D(\bar{\boldsymbol{\theta}}_m)$ is the deviance of model m evaluated at the posterior mean.

Recently, Spiegelhalter et al. (2002) introduced the Deviance information criterion (DIC) which is calculated as

$$DIC(m) = \overline{D(\boldsymbol{\theta}_m)} - p_m,$$

where p_m represents the "effective" number of parameters. The above expression of DIC is equivalent to

$$DIC(m) = 2\overline{D(\boldsymbol{\theta}_m)} - D(\bar{\boldsymbol{\theta}}_m).$$

This criterion is useful in locating the best model within a group of models; it does not indicate whether a model is correct or not (Lopes, 2002).

2.9 Conclusion

In this chapter we tried to summarize the main aspects of Bayesian theory. As we have seen, Bayesian theory is rather autonomous in nature since it is strictly based on basic properties of conditional probabilities. As such it is relatively easy to understand. In fact, it requires only comprehension of three basic distributions; the prior distribution, the likelihood and the posterior distribution. All subsequent aspects of Bayesian theory are products of the relationships between these distributions.

We also presented brief descriptions of common inferential tools used in Bayesian data analysis. In particular, we discussed the use of location and dispersion measures along with the use of credible sets in order to summarize posterior information. The purpose of Bayesian p-values based on posterior predictive checks was examined and finally we showed how Bayes factors, posterior odds and information criteria can be utilized in terms of model selection.

In the following chapter we focus on the basic *Markov Chain Monte Carlo* algorithms. We then implement the algorithms in a linear regression example and a logistic regression example.

Chapter 3: Markov Chain Monte Carlo

3.1 Introduction

Bayesian inference is strictly based on the posterior distribution. Having at hand the posterior distribution, we can easily calculate any summary of interest or even graphically present the posterior for inferential purposes. As we have seen, the computation of the posterior distribution comes down to the evaluation of complex, often high dimensional, integrals. In many circumstances this integration cannot be derived analytically. In addition, posterior summarization often involves computing moments or percentiles, which leads to the evaluation of more integrals. Due to the above problems, Bayesian statistics were not broadly used for many years.

First attempts to use the Bayesian approach relied mainly on conjugate prior distributions. The use of a conjugate to the likelihood prior distribution is an easy and acceptable solution for some cases. In particular, priors from the beta or the gamma distributions have proven to be quite flexible in expressing prior beliefs, since they can produce various forms of densities. The use of Generalized Linear Models (GLM) was also available due to the fact that distributions which belong to the exponential family always have a conjugate prior (Morris, 1983). So, many common problems could be solved with the use of appropriate conjugate distributions.

As large sample theory became more popular, it was also implemented in Bayesian statistics. Asymptotic methods were used in order to obtain analytic approximations of the posterior distribution. The simplest method is to use a *normal approximation* to the posterior. This approximation is essentially a Bayesian version of the central limit theorem. A more complicated asymptotic method is the *Laplace approximation* which provides more accurate posterior approximations.

The recent development of simulation methods and the evolution of computer technology provided statisticians with new computational orientated

methods. Concerning Bayesians, it was no longer necessary to use a conjugate prior in order to calculate the posterior distribution. Moreover, asymptotic approximations were not the only alternative now; the ability to generate random draws from the posterior gave the researches the option to calculate directly any summary of interest, even get an estimate of the full joint posterior density. This approach is generally known as *Monte Carlo integration* or simply *Monte Carlo*.

3.2 Asymptotic methods

Large sample or asymptotic theory indicates that if the posterior distribution is unimodal and roughly symmetric, then the posterior distribution can be approximated by a normal distribution around its mode as sample size increases. That is, for $n \to \infty$

$$p(\boldsymbol{\theta} \mid y) \approx N\left(\tilde{\boldsymbol{\theta}}, \left[J\left(\tilde{\boldsymbol{\theta}}\right)\right]^{-1}\right),$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode of $\boldsymbol{\theta}$ and $J\left(\boldsymbol{\theta}\right)$ is the observed information matrix given by

$$J\left(\boldsymbol{\theta}\right) = -\frac{d^2}{d\boldsymbol{\theta}^2} \log p(\boldsymbol{\theta} \mid y).$$

This approximation results from a Taylor expansion of $\log p(\boldsymbol{\theta} \mid y)$ at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ (Gamerman and Lopes, 2006, p.83). Usually a maximization algorithm like the Newton-Raphson or the Expectation-Maximization (EM) algorithm is being used in order to locate the posterior mode. Then we can adopt the normal approximation or even a Student's t approximation, if sample size is not adequately large (Gelman et al., 1995, p. 275). Of course, the question how large should a sample be in order to use the normal approximation is not straightforward to answer. Nevertheless, the approximation is quite accurate if $\boldsymbol{\theta}$ is a low dimensional vector. This means that the method works better for conditional and marginal distributions rather than for full joint distributions. If the dimension of $\boldsymbol{\theta}$ is high then $\boldsymbol{\theta}$ can be partitioned into $\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_k\right)$

subvectors and we can approximate all or some of the lower dimension conditional densities $p(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)}, \mathbf{y})$, where $\boldsymbol{\theta}_{(-i)} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, ..., \boldsymbol{\theta}_k)$ for $i = 1, 2, ...k$. On the other hand, a marginal distribution of one component of $\boldsymbol{\theta}$, is actually an average over all other components stretching this distribution closer to normality (Gelman et al., 1995, p.97).

Another, more complex, approach is the Laplace approximation. This method produces in general better point estimates if the posterior density is significantly far from the normal one. The Laplace approximation is subjected to the same limitations with the normal approximation, regarding the issues of sample size and dimensionality; for details on the Laplace approximation see Tierney and Kadane (1986), Gamerman and Lopes (2006, p. 88-92) and Carlin and Louis (1996, p.146).

3.3 Monte Carlo Integration

The basic idea of Monte Carlo (MC) integration is attractively simple; suppose y is a random variable, from which we can generate random draws, that is $y \sim p(y)$. Then for every quantity of interest $\gamma \equiv E[g(y)] = \int g(y)p(y)dy$ which cannot be calculated analytically, we can draw $y_1, y_2, ..., y_N \overset{iid}{\sim} p(y)$ (N is now a *simulated* sample) and calculate

$$\hat{\gamma} \equiv \frac{1}{N}\sum_{i=1}^{N} g(y_i).$$

The estimate $\hat{\gamma}$ is a strongly consistent estimate of γ in that

$$\hat{\gamma} \to \gamma \text{ as } N \to \infty,$$

this means that $\hat{\gamma}$ converges to γ with probability 1 as $N \to \infty$. In addition from the Central Limit Theorem we have that

$$\sqrt{N}\frac{\hat{\gamma} - \gamma}{\sigma} \to N(0,1) \text{ as } N \to \infty.$$

Strong consistency follows directly from the Strong Law of Large Numbers. One can immediately see that MC estimates improve their precision at rate $O\left(N^{-1/2}\right)$ as the simulated sample size is increased. Unlike asymptotic results, the value of N is under the control of the researcher and can be increased by drawing more values from $p(\cdot)$. For more information see Gamerman and Lopes (2006, p.96).

Another contrast with asymptotic methods is that we can evaluate the accuracy of $\hat{\gamma}$ for any fixed value of *N*. Since $\hat{\gamma}$ is itself a sample mean of independent observations, we have that $Var(\hat{\gamma}) = \frac{\sigma^2}{N}$, where $\sigma^2 = Var\left[g(y)\right]$. But $Var\left[g(y)\right]$ can be estimated by the sample variance of the $g(y_i)$ values, where $i = 1, 2, ..., N$. Thus, a standard error estimate for $\hat{\gamma}$ is given by

$$\widehat{s.e}(\hat{\gamma}) \equiv \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}\left[g(y_i) - \hat{\gamma}\right]^2},$$

(Carlin and Louis, 1996, p.150). This quantity is called *Monte Carlo error*. By this term we refer to the standard error of the estimated due to the fact that we use a simulated sample. Monitoring the Monte Carlo error is essential; this quantity should be low in order to evaluate the parameter of interest with increased precision.

There are various Monte Carlo methods, such as the *Inverse CDF method*, *Importance Sampling*, *Rejection Sampling* and *Weighted Bootstrap*. The first method requires knowledge of the density or cumulative density functions, the latter three are based on the existence of an approximating density. An overview of these methods will not be presented here, since the main purpose of this chapter is to present the Markov Chain Monte Carlo (MCMC) methods. More information about these methods can be found in Carlin and Louis (1996, p.153-158) and Gamerman and Lopes (2006, p.25-34).

The main characteristic of these simulation methods is that all of them are non-iterative. Therefore, we generate a sample of size *N* one time and then stop. Another attribute is that they do not perform well for high dimensional problems. In such cases it is difficult to identify an approximating density. Even if an

approximating density is found it rarely leads to satisfactory results. Sometimes, these methods are used in combination with some asymptotic approximation in order to acquire initial values for the MCMC simulation.

3.4 Markov Chain Monte Carlo Algorithms

MCMC methods have recently become very popular in Bayesian statistics. They owe their popularity to their ability to accurately approximate high dimensional integrals through simulation. The basic idea of MCMC simulation is to formulate a Markov chain from a specific starting point. This chain converges to a stationary distribution due to the properties of Markov chains.

3.4.1 Markov chains

A Markov chain is a stochastic process $\left\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ..., \boldsymbol{\theta}^t\right\}$ with two important properties:

1. The distribution of $\boldsymbol{\theta}$ in period $t+1$, given the $\boldsymbol{\theta}$ for all preceding periods depends only on the $\boldsymbol{\theta}$ in the latest time period t. That is,

$$f\left(\boldsymbol{\theta}^{t+1} \mid \boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}, ..., \boldsymbol{\theta}^1\right) = f\left(\boldsymbol{\theta}^{t+1} \mid \boldsymbol{\theta}^t\right).$$

Alternatively one could say that given the present state of a Markov chain, past and future states are independent.

2. If a Markov chain is irreducible, aperiodic and positive recurrent, then as t tends to infinity ($t \to \infty$) the distribution of $\boldsymbol{\theta}^{(t)}$ tends to a stationary distribution.

The stationary distribution is often called the equilibrium distribution of the Markov chain. Irreducible and aperiodic, mean that there is positive probability moving from any state to any other state and that there are no absorbing states from which the chain cannot escape (Carlin and Louis, 1996, p.75). Positive

recurrent, means that the probability of returning to the state from which we started equals to 1 and that the expected time of return is finite (Gamerman and Lopes, 2006, p.118). Further details for the properties of Markov chains and examples can be found in Gamerman and Lopes (2006, p,113-136) and Gilks et al. (1996, p.59-71).

Within the Bayesian framework, we wish to generate observations $\left\{\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,...,\boldsymbol{\theta}^t\right\}$ of the Markov chain, as a dependent sample from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. In order to achieve that, the equilibrium distribution of the simulated Markov chain must actually concur to the posterior distribution. Once this is accomplished, we must discard an initial part of the simulated observations from $\boldsymbol{\theta}^1$ to $\boldsymbol{\theta}^{t_0}$ and keep $\left\{\boldsymbol{\theta}^{t_0+1},\boldsymbol{\theta}^{t_0+2},...,\boldsymbol{\theta}^t\right\}$. The discarded part is the so called "burn-in" period, that is, the period the Markov chain has not yet converged to its equilibrium distribution. Finally, it is essential to use some kind of a diagnostic tool, in order to check whether or not our Markov chain has reached its equilibrium distribution.

The two basic MCMC methods are the *Metropolis-Hastings* algorithm and the *Gibbs Sampler*. There is also the *Metropolis within Gibbs* algorithm which is in fact a combination of the two previous algorithms mentioned.

3.4.2 The Metropolis and Metropolis-Hastings Algorithms

The basic Metropolis algorithm was introduced by Metropolis *et al*. (1953), later Hastings (1970) presented a generalized version of that algorithm, the Metropolis-Hastings algorithm.

The Metropolis-Hastings (M-H) algorithm is based on the existence of a proposal distribution (also called 'candidate' or 'jumping' distribution) $q_t(\boldsymbol{\theta}^{t-1},\boldsymbol{\theta}^t)$ which is actually part of a certain transition kernel (the Metropolis algorithm is restricted in symmetric proposal distributions). According to Markov chain theory, iterations from the transition kernel converge to the equilibrium distribution when the number of iterations is large; for information

on the transition kernel see Chib and Greenberg (1995), Brooks (1998) and Gilks et al. (1996, p.7).

Our aim is to sample from $p(\boldsymbol{\theta}|\mathbf{y})$, the target distribution which we may know only up to a constant multiple. Given any states $\boldsymbol{\theta}^a, \boldsymbol{\theta}^b$, $q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a)$ is actually the probability of transition from state $\boldsymbol{\theta}^b$ to state $\boldsymbol{\theta}^a$. If the proposal distribution satisfies the relation $p(\boldsymbol{\theta}^b|\mathbf{y})q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a) = p(\boldsymbol{\theta}^a|\mathbf{y})q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)$ for all states $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$, then iterations from q_t for $t \to \infty$ converge to the target distribution $p(\boldsymbol{\theta}|\mathbf{y})$. This sufficient condition is called the reversibility condition, and can be intuitively explained as 'the unconditional probability of moving from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$ when $\boldsymbol{\theta}^b$ is generated by $p(\cdot|\mathbf{y})$ equals the unconditional probability of moving from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$ when $\boldsymbol{\theta}^a$ is generated by $p(\cdot|\mathbf{y})$' (Chib and Greenberg, 1995).

This condition is not always satisfied; therefore the Metropolis-Hastings algorithm introduces a probability of transition or move a_{MH} as part of the transition kernel. For example, if the unconditional probability of moving from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$ is greater than the unconditional probability of moving from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$, that is $p(\boldsymbol{\theta}^b|\mathbf{y})q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a) > p(\boldsymbol{\theta}^a|\mathbf{y})q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)$, then this means that transitions from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$ are made too often while transitions from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$ are made rarely (Chib and Greenberg, 1995). Therefore, the transition probability $a_{MH}(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)$ from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$ is set equal to 1 in order to have more transitions from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$, and the transition probability $a_{MH}(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a)$ from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$ must then balance the two sides of the reversibility condition. That is,

$$
\begin{aligned}
&p(\boldsymbol{\theta}^b|\mathbf{y})q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a)a_{MH}(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a) = p(\boldsymbol{\theta}^a|\mathbf{y})q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)a_{MH}(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) \Leftrightarrow \\
&p(\boldsymbol{\theta}^b|\mathbf{y})q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a)a_{MH}(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a) = p(\boldsymbol{\theta}^a|\mathbf{y})q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) \Leftrightarrow \\
&a_{MH}(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a) = \frac{p(\boldsymbol{\theta}^a|\mathbf{y})q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b)}{p(\boldsymbol{\theta}^b|\mathbf{y})q_t(\boldsymbol{\theta}^b, \boldsymbol{\theta}^a)}.
\end{aligned}
$$

Thus, balance is obtained between the two sides of the reversibility condition. For more details see Chib and Greenberg (1995), Brooks (1998), Besag (2001) and Gilks et al. (1996).

In a more general notation, for any states x, y the transition probability from state *x* to state *y* is given by

$$a_{MH}(x,y) = \min\left[\frac{p(y)q(y,x)}{p(x)q(x,y)}, 1\right].$$

This implies that knowledge of the normalizing constant of the target distribution is not needed, since it cancels out in the ratio

$$\frac{p(y)q(y,x)}{p(x)q(x,y)}.$$

In the case of the Metropolis algorithm we have that $q(x,y) = q(y,x)$, since the proposal distribution is symmetric, therefore the transition probability reduces to

$$a_{M}(x,y) = \min\left[\frac{p(y)}{p(x)}, 1\right].$$

A relatively simple proof that the Metropolis algorithm sequence $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ..., \boldsymbol{\theta}^t\}$ converges to the target distribution is provided by Gelman et al. (1995, p.325). First, it is shown that the simulated sequence is a Markov chain that converges to a unique stationary distribution. The proof of this is trivial, since the selection of the proposal $q_t(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}^{t-1})$ ensures the properties of irreducibility, aperiodicity and positive recurrency, properties that hold for most random walks. Then, it is shown that the stationary distribution equals the target distribution.

Consider two states $\boldsymbol{\theta}^a, \boldsymbol{\theta}^b$ generated from $p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y})$ with $p(\boldsymbol{\theta}^b \mid \mathbf{y}) \geq p(\boldsymbol{\theta}^a \mid \mathbf{y})$, then the unconditional probability of a transition from $\boldsymbol{\theta}^a$ to $\boldsymbol{\theta}^b$ is

$$\begin{aligned} p(\boldsymbol{\theta}^{t-1} = \boldsymbol{\theta}^a, \boldsymbol{\theta}^t = \boldsymbol{\theta}^b \mid \mathbf{y}) &= p(\boldsymbol{\theta}^a \mid \mathbf{y}) q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) a_M(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) \\ &= p(\boldsymbol{\theta}^a \mid \mathbf{y}) q_t(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b), \end{aligned}$$

because the probability of move $a_M(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) = 1$, since

$$p(\boldsymbol{\theta}^b \mid \mathbf{y}) \geq p(\boldsymbol{\theta}^a \mid \mathbf{y}) \Rightarrow \frac{p(\boldsymbol{\theta}^b \mid \mathbf{y})}{p(\boldsymbol{\theta}^a \mid \mathbf{y})} \geq 1.$$

The unconditional probability of a transition from $\boldsymbol{\theta}^b$ to $\boldsymbol{\theta}^a$ is

$$
\begin{aligned}
p(\boldsymbol{\theta}^{t-1}=\boldsymbol{\theta}^b,\boldsymbol{\theta}^t=\boldsymbol{\theta}^a\,|\,\mathbf{y}) &= p(\boldsymbol{\theta}^b\,|\,\mathbf{y})q_t(\boldsymbol{\theta}^b,\boldsymbol{\theta}^a)a_M(\boldsymbol{\theta}^b,\boldsymbol{\theta}^a) \\
&= p(\boldsymbol{\theta}^b\,|\,\mathbf{y})q_t(\boldsymbol{\theta}^b,\boldsymbol{\theta}^a)\frac{p(\boldsymbol{\theta}^a\,|\,\mathbf{y})}{p(\boldsymbol{\theta}^b\,|\,\mathbf{y})} \\
&= p(\boldsymbol{\theta}^a\,|\,\mathbf{y})q_t(\boldsymbol{\theta}^b,\boldsymbol{\theta}^a).
\end{aligned}
$$

This two probabilities are equal since q_t is symmetric ($q_t(\boldsymbol{\theta}^b,\boldsymbol{\theta}^a)=q_t(\boldsymbol{\theta}^a,\boldsymbol{\theta}^b)$). This means that the joint distribution of $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t-1}$ is symmetric and therefore $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t-1}$ have the same marginal distributions. So, $\int_{\boldsymbol{\theta}^{t-1}} p(\boldsymbol{\theta}^{t-1},\boldsymbol{\theta}^t)d\boldsymbol{\theta}^{t-1}=p(\boldsymbol{\theta}^t\,|\,\mathbf{y})$ is *also* the stationary distribution of the Markov chain. The same proof can be also utilized for the Metropolis-Hastings algorithm by simply replacing the probability of transition.

To simulate a Metropolis sample of size N we use the following steps:

1. Set initial values $\boldsymbol{\theta}^0$.
2. For $t=1,2,...,N$:
   a. Generate $\boldsymbol{\theta}^*$ from the proposal density $q_t(\boldsymbol{\theta}^{t-1},\boldsymbol{\theta})$.
   b. Calculate $a_M=\min\left(\frac{p(\boldsymbol{\theta}^*\,|\,\mathbf{y})}{p(\boldsymbol{\theta}^{t-1}\,|\,\mathbf{y})},1\right)$.
   c. Set $\boldsymbol{\theta}^t=\begin{cases}\boldsymbol{\theta}^* & \text{, with probability } a_M \\ \boldsymbol{\theta}^{t-1} & \text{, with probability } 1-a_M.\end{cases}$

Likewise, to simulate a Metropolis-Hastings sample of size N we use the following steps:

1. Set initial values $\boldsymbol{\theta}^0$.
2. For $t=1,2,...,N$:
   a. Generate $\boldsymbol{\theta}^*$ from the proposal density $q_t(\boldsymbol{\theta}^{t-1},\boldsymbol{\theta})$.
   b. Calculate $a_{MH}=\min\left(\frac{p(\boldsymbol{\theta}^*\,|\,\mathbf{y})q_t(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{t-1})}{p(\boldsymbol{\theta}^{t-1}\,|\,\mathbf{y})q_t(\boldsymbol{\theta}^{t-1},\boldsymbol{\theta}^*)},1\right)$.
   c. Set $\boldsymbol{\theta}^t=\begin{cases}\boldsymbol{\theta}^* & \text{, with probability } a_{MH} \\ \boldsymbol{\theta}^{t-1} & \text{, with probability } 1-a_{MH}.\end{cases}$

Step (c) of each algorithm requires the generation of a uniform random number u from $U(0,1)$. If $u \le a(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^{*})$ we set $\boldsymbol{\theta}^{t} = \boldsymbol{\theta}^{*}$, else we set $\boldsymbol{\theta}^{t} = \boldsymbol{\theta}^{t-1}$.

Regarding the issue of proposal distribution selection, according to Gelman et al. (1995, p.326) a good jumping density possesses the following properties:

- It is easy to sample from $q_t(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$.
- The probability of transition $a\left(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^{*}\right)$, can be easily calculated.
- Each move or jump goes a reasonable distance in the parameter space.
- The jumps are not rejected too frequently.

The most commonly used choice is the *random walk chain*, in which the candidate $\boldsymbol{\theta}^{*}$ is drawn according to the process $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t} + \mathbf{z}_{t+1}$. In this case the candidate value equals the present value plus noise. A usual choice for a random walk chain is a multivariate normal distribution, $\boldsymbol{\theta}^{*} \sim N(\boldsymbol{\theta}^{t}, \boldsymbol{\Sigma})$ or a multivariate t distribution, $\boldsymbol{\theta}^{*} \sim t_{\nu}(\boldsymbol{\theta}^{t}, \boldsymbol{\Sigma})$. These densities have the advantage of being symmetric, so utilizing them simplifies calculations. The use of a random walk chain requires only the determination of the covariance matrix $\boldsymbol{\Sigma}$.

Another candidate-generating family of distributions, arises from a simpler process, that is $\boldsymbol{\theta}_{t+1} = \mathbf{z}_{t+1}$. In this case the candidate value is independent of the current value and $\mathbf{z}^{*}\left(\equiv \boldsymbol{\theta}^{*}\right)$ has a multivariate density, which can again, be a multivariate normal or t distribution, $\boldsymbol{\theta}^{*} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\boldsymbol{\theta}^{*} \sim t_{\nu}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ respectively. This is called an *independence chain* and requires both location and scale parameters to be determined. For a more detailed description and other approaches on proposal densities see Chib and Greenberg (1995) and Gamerman and Lopes (2006, p.198-205).

The issue of determining the scale of the proposal distribution is one that has not yet been fully explored. This issue is of vital importance, regarding the speed of convergence. If the scale is set very large then a lot of candidate values

will be far away from the high density region of the posterior parameter space, and thus the acceptance ratio will be very low. On the other hand, given that the starting values are not extreme in regard to posterior parameter space, a very small scale will result in a higher acceptance ratio but still we will have to face the problem of undersampled low density regions, since the chain will need more time to reach those regions (Chib and Greenberg, 1995; Carlin and Louis, 1996).

One approach suggested, is to run an initial chain, obtain a crude estimate $\tilde{\boldsymbol{\Sigma}}$ and use this estimate as the scale of the proposal distribution (Carlin and Louis, 1996). Gelman *et al*. (1995, p.334) proposed for problems with normal target and normal random walk proposal densities, $\left(2.4^2/d\right)\tilde{\boldsymbol{\Sigma}}$ to be the most efficient scale, with d being the number of dimensions and $\tilde{\boldsymbol{\Sigma}}$ some estimate of $\boldsymbol{\Sigma}$. This rule results to acceptance ratios around 0.45 and 0.25, respectively for one-dimensional and multi-dimensional problems which are up to six dimensions. In absence of general rules, the selection of scale is in most of the cases a calibrating process; we increase or decrease the scale in order to achieve an acceptance ratio usually within the rate of 0.3-0.5 for univariate distributions (Gilks et al., 1996).

3.4.3 The Gibbs Sampler

The Gibbs sampler algorithm was introduced by Geman and Geman (1984). This algorithm is actually a special case of the Metropolis-Hastings algorithm but it is often presented separately due to its popularity and easy-to-use nature.

In Gibbs sampling each component of the vector $\boldsymbol{\theta}$ is drawn separately. Therefore, when $\boldsymbol{\theta}$ is of dimensionality d, we have d steps in every iteration. The components of $\boldsymbol{\theta}$ are actually generated from their full conditional distribution, this means that component θ_j is sampled from $p\left(\theta_j^t \mid \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)$, with $\boldsymbol{\theta}_{(-j)}^{t-1} = \left(\theta_1^t, \theta_2^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_{d-1}^{t-1}, \theta_d^{t-1}\right)$. This cycling process ends when all d components have been drawn. One can see that Gibbs sampler is a special case of a Metropolis-Hastings algorithm for a single component with proposal

distribution $q_{j,t}(\boldsymbol{\theta}^{t-1}, \theta_j^t)$, the full conditional distribution $p\left(\theta_j^t \mid \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)$. According to Gelman *et al*. (1995, p.328) we have

$$q_{j,t}(\boldsymbol{\theta}^{t-1}, \theta_j^*) = \begin{cases} p\left(\theta_j^* \mid \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right), & \text{if } \boldsymbol{\theta}_{(-j)}^* = \boldsymbol{\theta}_{(-j)}^{t-1} \\ 0, & \text{otherwise.} \end{cases}$$

The above relation means that the j component of $\boldsymbol{\theta}$ is updated when all components of vectors $\boldsymbol{\theta}^*, \boldsymbol{\theta}^{t-1}$, except j, match. In this case the probability of transition becomes

$$a_{Gibbs} = \frac{p(\boldsymbol{\theta}^* \mid \mathbf{y}) p\left(\theta_j^{t-1} \mid \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)}{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y}) p\left(\theta_j^* \mid \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)} = \frac{\dfrac{p(\boldsymbol{\theta}^* \mid \mathbf{y}) p\left(\theta_j^{t-1}, \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)}{p\left(\boldsymbol{\theta}_{(-j)}^{t-1} \mid \mathbf{y}\right)}}{\dfrac{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y}) p\left(\theta_j^*, \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)}{p\left(\boldsymbol{\theta}_{(-j)}^{t-1} \mid \mathbf{y}\right)}} = \frac{p(\boldsymbol{\theta}^* \mid \mathbf{y}) p\left(\theta_j^{t-1}, \boldsymbol{\theta}_{(-j)}^{t-1}, \mathbf{y}\right)}{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y}) p\left(\theta_j^*, \boldsymbol{\theta}_{(-j)}^*, \mathbf{y}\right)}$$

$$= \frac{p(\boldsymbol{\theta}^* \mid \mathbf{y}) p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y})}{p(\boldsymbol{\theta}^{t-1} \mid \mathbf{y}) p(\boldsymbol{\theta}^* \mid \mathbf{y})} = 1.$$

Thus, in Gibbs sampling the acceptance ratio equals 1 which means that the proposed move is always accepted.

To simulate a Gibbs sample of size *N*, for $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$, we use the following steps:

1. Determine $d-1$ initial values $\boldsymbol{\theta}^0 = \left(\theta_2^0, ..., \theta_d^0\right)$.

2. For $t = 1, 2, ..., N$:

   Generate θ_1^t from $p\left(\theta_1 \mid \theta_2^{t-1}, ..., \theta_d^{t-1}, \mathbf{y}\right)$,

   Generate θ_2^t from $p\left(\theta_2 \mid \theta_1^t, \theta_3^{t-1}, ..., \theta_d^{t-1}, \mathbf{y}\right)$,

   Generate θ_3^t from $p\left(\theta_3 \mid \theta_1^t, \theta_2^t, \theta_4^{t-1}, ..., \theta_d^{t-1}, \mathbf{y}\right)$,

   $\vdots$ $\qquad\qquad$ $\vdots$

   Generate θ_d^t from $p\left(\theta_d \mid \theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t, ..., \theta_{d-1}^t, \mathbf{y}\right)$.

The ordering of the generating components has no affect on the convergence of the algorithm. For additional information concerning the Gibbs

sampler see Casella and George (1992), Brooks (1998), Gelfand (2000), Besag (2001), and Gilks et al. (1996).

3.4.4 The Metropolis within Gibbs Algorithm

As we have seen, the Gibbs sampler algorithm requires that all full conditionals are of known form and easy to generate from, in order to be applicable. In many cases though, we do not know the exact form of all or some full conditionals.

Suppose we know all full conditionals of $p(\boldsymbol{\theta}|\mathbf{y})$ except the one for component θ_i. The idea of the Metropolis within Gibbs algorithm, (some authors refer to it as *Univariate Metropolis* or *Metropolis Steps*), is to run a Gibbs sampler for the known full conditionals and use a Metropolis step, within the former, in order to update component θ_i. The target density of θ_i can be easily approximated, since it is analogous to the posterior, $p\left(\theta_i|\boldsymbol{\theta}_{(-i)},\mathbf{y}\right)\propto p\left(\boldsymbol{\theta}|\mathbf{y}\right)$, when all components except θ_i are held constant to their given value. This means that in the *i*-th step of each Gibbs iteration, we run a Metropolis chain of size *T* from which we keep the last value θ_i^T and then proceed with the outer Gibbs loop for component θ_{i+1}.

Thus, to simulate a Metropolis within Gibbs sample of size *N*, for $\boldsymbol{\theta}=\left(\theta_1,\theta_2,...,\theta_d\right)$, we use the following steps:

1. Determine $d-1$ initial values $\boldsymbol{\theta}^0=\left(\theta_2^0,...,\theta_d^0\right)$.

2. For $t=1,2,...,N$:

Generate θ_1^t from $p\left(\theta_1|\theta_2^{t-1},...,\theta_d^{t-1},\mathbf{y}\right)$,

Generate θ_2^t from $p\left(\theta_2|\theta_1^t,\theta_3^{t-1},...,\theta_d^{t-1},\mathbf{y}\right)$,

$\vdots$ $\qquad\qquad$ $\vdots$

Generate θ_{i-1}^t from $p\left(\theta_{i-1}|\theta_1^t,\theta_2^t,...,\theta_{i-2}^t,\theta_i^{t-1},...,\theta_d^{t-1},\mathbf{y}\right)$,

Use an inner Metropolis-Hastings loop in order to acquire θ_i^t.

For $l = 1, 2, ..., T$:

a. Generate θ_i^* from the proposal density $q_t(\theta_i^{l-1}, \theta_i)$.

b. Calculate $a_{MH} = \min\left(\frac{p(\theta_i^* \mid \boldsymbol{\theta}_{(-i)}, \mathbf{y}) q_t(\theta_i^*, \theta_i^{l-1})}{p(\theta_i^{l-1} \mid \boldsymbol{\theta}_{(-i)}, \mathbf{y}) q_t(\theta_i^{l-1}, \theta_i^*)}, 1 \right)$.

c. Set $\theta_i^l = \begin{cases} \theta_i^* & , \text{ with probability } a_{MH} \\ \theta_i^{l-1} & , \text{ with probability } 1 - a_{MH}. \end{cases}$

When $l = T$ set $\theta_i^t = \theta_i^T$ and continue with the outer Gibbs loop as follows.

Generate θ_{i+1}^t from $p\left(\theta_{i-1} \mid \theta_1^t, \theta_2^t, ..., \theta_i^t, \theta_{i+2}^{t-1}, ..., \theta_d^{t-1}, \mathbf{y}\right)$,

$\vdots \qquad\qquad \vdots$

Generate θ_d^t from $p\left(\theta_d \mid \theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t, ..., \theta_{d-1}^t, \mathbf{y}\right)$.

The convergence of this algorithm is not perfectly clear, since it is not a mixture or cycle of two separate algorithms. The Metropolis within Gibbs is rather, a deterministic combination of algorithms, which by themselves alone would not converge. Still, if the proposals used for each component are irreducible and aperiodic then each component will tend to its equilibrium distribution and convergence will occur (Carlin and Louis, 1996, p.182).

Another issue of concern is the selection of *T* for the inner Metropolis loop. A very large *T* would of course lead to a confident selection of θ_i^T, but it would be useless, in terms of overall convergence, especially in the early stages of the outer Gibbs loop. A very small *T*, like 1, would delay overall convergence, since it will be unlikely that θ_i^T will originate from the correct full conditional distribution (Carlin and Louis, 1996, p.182). In practice though, the selection $T = 1$ is often adopted.

3.4.5 Convergence Diagnostics

There is a variety of MCMC convergence diagnostics, which can be used in order to determine whether or not the equilibrium distribution of the Markov chain has been reached. These diagnostic tools have different characteristics. They can be quantitative and produce a single numeric summary, or qualitative like graphs and time series plots. Some require a single chain to be produced, while others a small number of parallel chains. Most of these methods base their approach on bias considerations, but there are also methods which check the variance or precision of the estimates. Extensive information on diagnostic tools can be found in Brooks and Roberts (1998), Mengersen et al. (1998), Cowles and Carlin (1996) and Brooks et al. (1997).

In this thesis we rely mainly on the diagnostic tool introduced by Gelman and Rubin (1992); see also Brooks and Gelman (1998). This approach addresses the issue of variance and requires a small number of parallel chains who must be initially overdispersed with respect to the target density. It can be used in either Metropolis-Hastings or Gibbs sampler simulations.

For each estimand θ of interest, the draws from the J parallel chains are labeled θ_{ij}, with $i = 1, 2, ..., N$ and $j = 1, 2, ..., J$. Then, we calculate the between and within chain variation, B and W respectively. They are,

$$B = \frac{N}{J-1}\sum_{j=1}^{J}\left(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot}\right)^2, \text{ where } \bar{\theta}_{\cdot j} = \frac{1}{N}\sum_{i=1}^{N}\theta_{ij}, \ \bar{\theta}_{\cdot\cdot} = \frac{1}{J}\sum_{j=1}^{J}\bar{\theta}_{\cdot j},$$

$$W = \frac{1}{J}\sum_{j=1}^{J}s_j^2, \text{ where } s_j^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\theta_{ij} - \bar{\theta}_{\cdot j}\right)^2.$$

An estimate of the marginal posterior variance $Var(\theta \mid \mathbf{y})$ can be given by a weighted average of B and W, which is $\widehat{Var}(\theta \mid \mathbf{y}) = \frac{N-1}{N}W + \frac{1}{N}B$. This estimate is unbiased under stationarity but overestimates the variance under the assumption that the starting distribution is overdispersed. Furthermore, the within chain variance is an underestimate of the marginal posterior variance, because each chain alone will not have time to visit the entire posterior space

and thus will have lower variability. Yet, for $N \to \infty$ the expectation of W approaches $Var(\theta | \mathbf{y})$ and therefore, convergence can be monitored by the scale reduction measure

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{Var}(\theta | \mathbf{y})}{W}},$$

which declines to 1 for $N \to \infty$ (Gelman et al., 1995, p.332).

Convergence can be also be checked by monitoring the Monte Carlo error of the estimates, since small values of it indicate that we have calculated the quantity of interest with precision. Calculating the MC error as described in section 3.3 - from the MCMC sample variance - would probably be anticonservative; the sampling chain will likely feature positive autocorrelation leading to an underestimate of the simulated sample's standard deviation (Carlin and Louis, 1996, p.194).

The computationally easiest way of estimating MC error is through the *batch mean* method. For any quantity $h(\boldsymbol{\theta})$ of interest we simply partition the MCMC sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_N$ into K batches $B_1, B_2, ..., B_K$ of size N_B. Thus, we have that $N = KN_B$. For each batch we estimate the corresponding sample mean from

$$\overline{h(\boldsymbol{\theta})}_{B_k} = \frac{1}{T_B} \sum_{t=(k-1)T_B+1}^{kT_B} h(\boldsymbol{\theta}^{(t)}),$$

where $k = 1, 2, ..., K$. Then, an estimate for the MC error of $\widehat{h(\boldsymbol{\theta})}$ is given by

$$\widehat{se}\left(\widehat{h(\boldsymbol{\theta})}\right) = \sqrt{\frac{1}{K(K-1)} \sum_{k=1}^{K} \left[\overline{h(\boldsymbol{\theta})}_{B_k} - \overline{h(\boldsymbol{\theta})}\right]^2},$$

where $\overline{h(\boldsymbol{\theta})}$ is given by $\overline{h(\boldsymbol{\theta})} = \frac{1}{K} \sum_{k=1}^{K} \overline{h(\boldsymbol{\theta})}_{B_k}$. Note that K must be large enough to ensure proper estimation of the variance (the usual choice is $30 \le K \le 50$) and N_B must also be large enough in order to ensure that the batch means are roughly independent. According to Carlin and Louis (1996, p.195) the latter can be determined by checking whether the lag 1 autocorrelation of the B_k is less than 0.05. If this is not the case, N_B must be increased.

In the remaining of this chapter we will make use of the aforementioned algorithms with two examples. In the first example we implement the Metropolis-Hastings and Gibbs Sampler algorithms on a normal linear regression problem and in the second one we use the Metropolis and Metropolis within Gibbs algorithms on a logistic regression example.

3.5 A linear regression example with normal data

The data presented here are wind velocity observations measured in miles per hour and electricity observations measured in volts (Montgomery et al., 2001, p.182). Our concern is the effect of wind velocity on the production of electricity from a water mill. Although wind velocity is by itself positively correlated with electricity production capacity, we will use the logarithm of wind velocity which has an even higher coefficient of correlation equal to 0.978.

The model used is the simple linear regression model $y = a + bx + \varepsilon$, with y the dependent variable, that is electricity production capacity (DC output), and x the explanatory variable, the logarithm of wind velocity. The Maximum Likelihood (ML) estimates of a, b and the standard deviation estimates are

$$\hat{a} = -0.83 \quad (0.111)$$
$$\hat{b} = 1.417 \quad (0.06)$$
$$\hat{\sigma} = 0.137.$$

The values in brackets correspond to the coefficients standard deviations. The 25 observations of wind velocity and DC output are given in Table 3.1. From the histogram of DC output observations presented in Figure 3.1 we notice that the distribution of the dependent variable seems to be skewed to the right, yet we cannot be absolutely sure due to the relatively small sample size. The plot right to the histogram indicates the strong positive correlation between the logarithm of wind velocity and DC output.

Obs.	*DC output* (volts)	*Wind Velocity (mph)*	*Logarithm of Wind Velocity*	Obs.	*DC output* (volts)	*Wind Velocity (mph)*	*Logarithm of Wind Velocity*
1.	*1.582*	*5.00*	*1.609*	**13.**	*1.562*	*4.60*	*1.526*
2.	*1.822*	*6.00*	*1.792*	**14.**	*1.737*	*5.80*	*1.758*
3.	*1.057*	*3.40*	*1.224*	**15.**	*2.088*	*7.40*	*2.001*
4.	*0.500*	*2.70*	*0.993*	**16.**	*1.137*	*3.60*	*1.281*
5.	*2.236*	*10.00*	*2.303*	**17.**	*2.179*	*7.85*	*2.061*
6.	*2.386*	*9.70*	*2.272*	**18.**	*2.112*	*8.80*	*2.175*
7.	*2.294*	*9.55*	*2.257*	**19.**	*1.800*	*7.00*	*1.946*
8.	*0.558*	*3.05*	*1.115*	**20.**	*1.501*	*5.45*	*1.696*
9.	*2.166*	*8.15*	*2.098*	**21.**	*2.303*	*9.10*	*2.208*
10.	*1.866*	*6.20*	*1.825*	**22.**	*2.310*	*10.20*	*2.322*
11.	*0.653*	*2.90*	*1.065*	**23.**	*1.194*	*4.10*	*1.411*
12.	*1.930*	*6.35*	*1.848*	**24.**	*1.144*	*3.95*	*1.374*
				25.	*0.123*	*2.45*	*0.896*

Table 3.1 *DC Output observations measured in volts and wind velocity observations in physical scale measured in miles per hour and in logarithmic scale.*

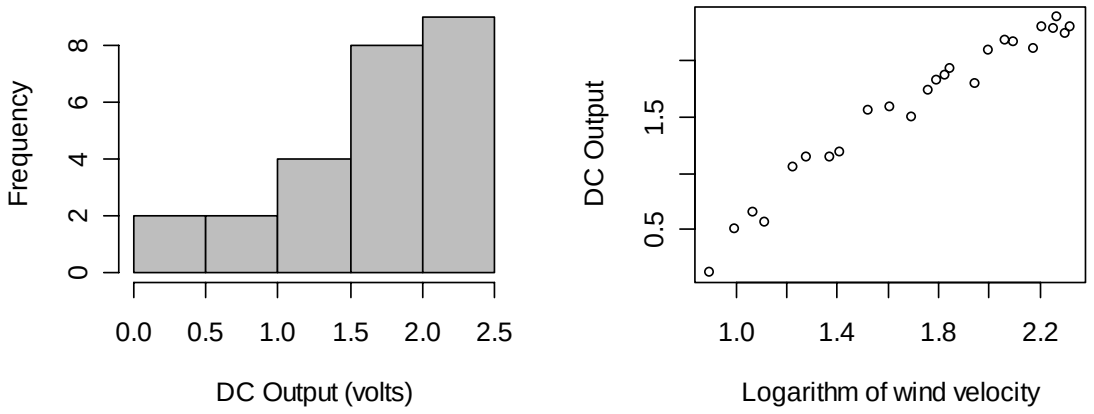

Figure 3.1 *Histogram of the 25 DC output observations (left) and scatter plot of the DC output and the wind velocity observations transformed into logarithmic scale (right).*

3.5.1 M-H algorithm implementation

As described before, our model assumption is $y = a + bx + \varepsilon$, with $\varepsilon \sim N(0, \sigma^2)$. The selected priors for parameters a and b are

$$a \mid \sigma^2 \sim N(0, k\sigma^2), \quad b \mid \sigma^2 \sim N(0, k\sigma^2),$$

with k considered as a multiplying constant.

We select a gamma prior for the inverse of σ^2 that is,

$$\sigma^{-2} \sim G(a_0, b_0),$$

where a_0 is the *shape* parameter and b_0 is the *rate* parameter. The distribution of σ^{-2} is given by

$$p(\sigma^{-2}) = \frac{b_0^{a_0}}{\Gamma(a_0)}\left(\sigma^{-2}\right)^{a_0-1}\exp\left(-b_0\sigma^{-2}\right) \text{ for } \sigma^{-2} > 0.$$

This means that the prior for σ^2 is an inverse gamma distribution with *shape* parameter a_0 and *scale* parameter b_0 that is, $\sigma^2 \sim IG(a_0, b_0)$. So, the prior distribution of σ^2 is given by

$$p(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}\left(\sigma^{2}\right)^{-a_0-1}\exp\left(\frac{-b_0}{\sigma^2}\right) \text{ for } \sigma^{2} > 0.$$

We further assume that parameters *a* and *b* are conditionally independent given σ^2. The prior distribution $p(a, b, \sigma^2)$ is actually conjugate to the likelihood. Therefore, this a case for which the posterior distribution can be calculated analytically. Our aim though, is to demonstrate the use of the Metropolis-Hastings algorithm regardless of prior selection.

The use of a hyper prior for parameters a and b means that we presume dependence between these parameters and the variance parameter. The prior design aims to reflect our ignorance concerning the three parameters. The selection of k is set large equal to 1000 and we use a gamma distribution with both location and rate parameters equal to 10^{-3}. This means that σ^{-2} is a random variable with mean equal to 1 and variance equal to 1000 and it also results to the fact that $\sigma^2 \sim IG(10^{-3}, 10^{-3})$, which is an inverse gamma random variable with infinite mean.

The joint posterior distribution of a, b, σ^2, under the assumption of conditional independence for parameters *a* and *b*, is

$$\begin{aligned} p(a, b, \sigma^2 \mid \mathbf{y}) &\propto p(a, b, \sigma^2)\, p(\mathbf{y} \mid a, b, \sigma^2) \\ &\propto p(a, b \mid \sigma^2)\, p(\sigma^2)\, p(\mathbf{y} \mid a, b, \sigma^2) \\ &\propto p(a \mid \sigma^2)\, p(b \mid \sigma^2)\, p(\sigma^2)\, p(\mathbf{y} \mid a, b, \sigma^2) \end{aligned}$$

$$p(a,b,\sigma^2 \mid \mathbf{y}) \propto \frac{1}{\sigma}\exp\left\{-\frac{a^2}{2k\sigma^2}\right\}\frac{1}{\sigma}\exp\left\{-\frac{b^2}{2k\sigma^2}\right\}\left(\sigma^2\right)^{-(a_0+1)}\exp\left\{-\frac{b_0}{\sigma^2}\right\}\times$$
$$\times\prod_{i=1}^{n}\frac{1}{\sigma}\exp\left\{\frac{\left(y_i-a-bx_i\right)^2}{2\sigma^2}\right\}$$
$$\propto \frac{\left(\sigma^2\right)^{-(a_0+1)}}{\sigma^{n+2}}\exp\left\{-\frac{b_0}{\sigma^2}\right\}\exp\left\{-\frac{a^2+b^2}{2k\sigma^2}-\frac{\sum_{i=1}^{n}\left(y_i-a-bx_i\right)}{2\sigma^2}\right\}.$$

After some algebra on the logarithm of this density we conclude to

$$\log p(a,b,\sigma^2 \mid \mathbf{y}) = -(2a_0+n+4)\log\sigma-\frac{b_0}{\sigma^2}-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}y_i^2+a^2(n+k^{-1})-\right.$$
$$\left.-2a\left(\sum_{i=1}^{n}y_i+b\sum_{i=1}^{n}x_i\right)-2b\sum_{i=1}^{n}y_ix_i+b^2\left(\sum_{i=1}^{n}x_i^2+k^{-1}\right)\right)+\text{Constant.}$$

The next step is to select the proposal distributions for the three parameters in question. For parameters a and b we choose a bivariate normal random walk proposal, that is $\begin{pmatrix}a^t\\b^t\end{pmatrix}\sim\mathbf{N}_2\left(\begin{pmatrix}a^{t-1}\\b^{t-1}\end{pmatrix},\boldsymbol{\Sigma}\right)$, with $\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_a^2 & \rho\sigma_a\sigma_b\\ \rho\sigma_a\sigma_b & \sigma_b^2\end{pmatrix}$. Sampling from this distribution is simple, since $q(a^t,b^t)=q(b^t \mid a^t)q(a^t)$. Therefore, we initially generate

$$a^t \sim N(a^{t-1},\sigma_a^2)$$

and then we generate b^t from the distribution which is conditional on a^t, that is

$$b^t \mid a^t \sim N\left(b^{t-1}+\rho\frac{\sigma_b}{\sigma_a}(a^t-a^{t-1}),\sigma_b^2(1-\rho^2)\right).$$

The bivariate normal distribution is symmetric and therefore, not needed for the calculation of the transition probability. The selection of the proposal distribution for σ is more complicated, since standard deviation is strictly positive. Therefore, we cannot use a normal proposal distribution on σ as above, but we can use such a proposal on $\log\sigma$, which takes values on the real line

unrestricted. So, we generate $\log\sigma^t$ from a normal random walk proposal $\log\sigma^t \sim N\left(\log\sigma^{t-1}, \sigma_\sigma^2\right)$ and then sample σ^t through the exponential transformation on $\log\sigma^t$. An important remark one should have in mind, is that although the proposal for $\log\sigma$ is again symmetric and therefore not needed in the calculation of the transition probability, the latter does not stand for the Jacobian of the exponential transformation which is $\frac{1}{\sigma}$. Thus, the probability of move is given by

$$a_{MH} = \frac{p\left(a^*, b^*, \sigma^{2*} \mid \mathbf{y}\right)\left(\frac{1}{\sigma^{t-1}}\right)}{p\left(a^{t-1}, b^{t-1}, \sigma^{2^{t-1}} \mid \mathbf{y}\right)\left(\frac{1}{\sigma^*}\right)}$$

and on logarithmic scale

$$\log a_{MH} = \log p\left(a^*, b^*, \sigma^{2*} \mid \mathbf{y}\right) + \log\sigma^* - \log p\left(a^{t-1}, b^{t-1}, \sigma^{2^{t-1}} \mid \mathbf{y}\right) - \log\sigma^{t-1}.$$

An equivalent sampling technique would be to sample σ directly from a log-normal random walk proposal that is $\sigma^t \sim Log-N\left(\sigma^{t-1}, \sigma_\sigma^2\right)$ with proposal density given by

$$q_{Log-N}(\sigma^t) = \frac{1}{\sigma^t\sqrt{2\pi}\sigma_\sigma}\exp\left(-\frac{(\log\sigma^t - \sigma^{t-1})^2}{2\sigma_\sigma^2}\right) \text{ for } \sigma^t > 0.$$

Yet, then we would be forced to compute probability densities from the log-normal distribution since in this case the probability of move is given by

$$a_{MH} = \frac{p\left(a^*, b^*, \sigma^{2*} \mid \mathbf{y}\right) q_{Log-N}(\sigma^{t-1})}{p\left(a^{t-1}, b^{t-1}, \sigma^{2^{t-1}} \mid \mathbf{y}\right) q_{Log-N}(\sigma^*)}.$$

An alternative approach would be a gamma random walk proposal distribution $G(a_{prop}, b_{prop})$ with shape parameter a_{prop} and rate parameter b_{prop}. Knowing that this distribution has mean equal to a_{prop}/b_{prop}, we set the shape parameter as $a_{prop} = b_{prop}\sigma^{t-1}$. Thus, we sample σ from a gamma random walk proposal $\sigma^t \sim G(\sigma^{t-1}b_{prop}, b_{prop})$ which has mean σ^{t-1} and density

$$q_{Gamma}(\sigma^t) = \frac{b_{prop}^{\,b_{prop}\sigma^{t-1}}}{\Gamma\left(b_{prop}\sigma^{t-1}\right)}\left(\sigma^t\right)^{b_{prop}\sigma^{t-1}-1}\exp\left(-b_{prop}\sigma^t\right).$$

The probability of move is now given by

$$a_{MH} = \frac{p\left(a^*, b^*, \sigma^{2^*} \mid \mathbf{y}\right) q_{Gamma}(\sigma^{t-1})}{p\left(a^{t-1}, b^{t-1}, \sigma^{2^{t-1}} \mid \mathbf{y}\right) q_{Gamma}(\sigma^*)}$$

and calculated on logarithmic scale

$$\log a_{MH} = \log p\left(a^*, b^*, \sigma^{2^*} \mid \mathbf{y}\right) + \log q_{Gamma}(\sigma^{t-1}) - \log p\left(a^{t-1}, b^{t-1}, \sigma^{2^{t-1}} \mid \mathbf{y}\right) - \\ - \log q_{Gamma}(\sigma^*).$$

The role of the rate parameter b_{prop} is as significant as that of the parameter σ_σ^2 of the normal proposal for $\log\sigma$. The bigger the selection of b_{prop}, the more symmetric and less variable will the proposal get, tending to a normal distribution in $\mathbb{R}^+$. A small selection of b_{prop} on the other hand, will result to a proposal density which will be skewed to the left with bigger spread.

Our first approach will be to run an initial small chain of size 1000 in order to acquire a first impression of the posterior parameter space. We choose starting values which are not far distant from the MLE estimates $a^{(0)} = -0.5,\ b^{(0)} = 1$ and $\sigma^{(0)} = 0.1$ and keep the second half of the M-H chain for inferential purposes. The proposal scale for parameters a and b, which at this stage are kept uncorrelated, is set equal to 1 ($\widehat{\sigma}_a = \widehat{\sigma}_b = 1$) and the scale parameter of the gamma proposal distribution of

parameter σ is set equal to 25 ($b_{prop} = 25$). If the acceptance ratio is lower than 20% by the time the chain reaches the half of its course, then we decrease the spread of the proposals in order to increase acceptance ratio. In this case, scale parameters are divided or multiplied with 10 giving $\hat{\sigma}_a = \hat{\sigma}_b = 0.1$ and $b_{prop} = 250$, respectively.

The resulting acceptance ratio is low equal to 9.6 %. Mean and standard deviation estimates from this first simulation are shown in Table 3.1. This initial run also reveals a strong negative correlation, near -0.95, for parameters a and b.

Parameter	***a***	***b***	***σ***
Mean	-0.776	1.385	0.146
St. Deviation	0.109	0.060	0.021

Table 3.2 *Mean and standard deviation estimates for parameters a, b and σ acquired from a Metropolis-Hastings sample of size 500.*

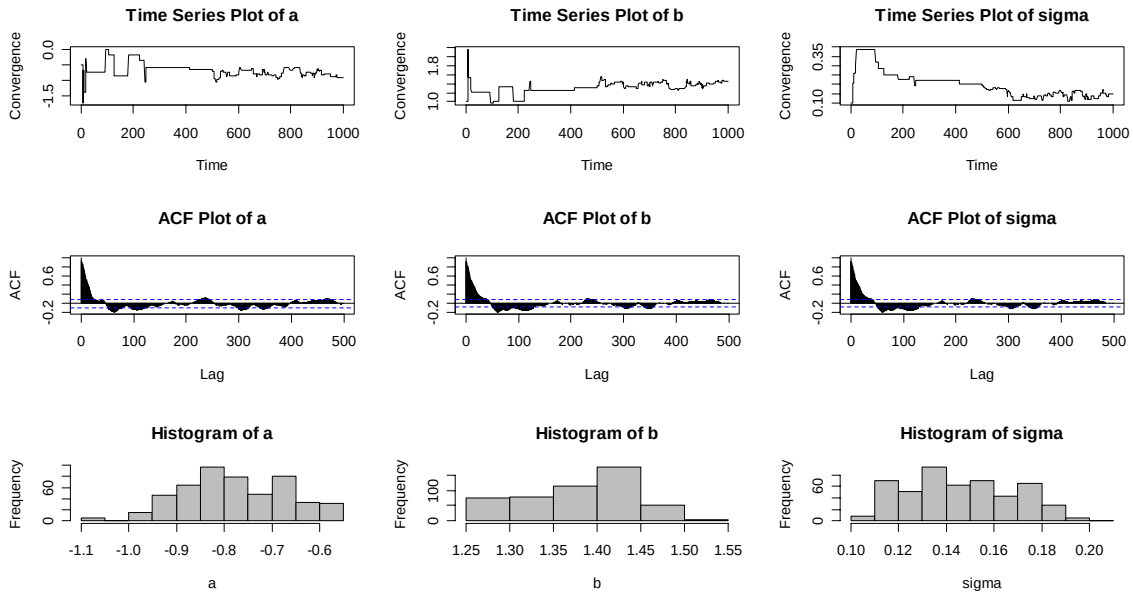

Figure 3.2 *Time series plots (up) of the 1000 iterations for parameters a, b and σ (sigma), autocorrelation plots (middle) and histograms (down) resulting from the last 500 iterations.*

Time series plots of the three parameters are shown in Figure 3.2. We can notice that the acceptance ratio is affected positively from the decrease in the proposals spread, since about half the distance, at iteration 500, the chain fluctuates more

than before. The rest of the plots in Figure 3.2, unlike time series plots, refer to the second half of the Metropolis-Hastings sample; that is the part we keep for inferential purposes. The ACF plots are a graphical representation of the parameters autocorrelation function. The histograms provide us a rough image of the posterior space.

Based on this information we can initialize a multi chain Metropolis-Hastings simulation from starting points that are over-dispersed in regard to posterior space. We now use five parallel chains of 9000 iterations each. The first 1000 iterations of each chain are discarded for the "burn-in" period. The initial values used are the following:

$$\text{Chain 1}: \left(a_1^0, b_1^0, \sigma_1^0\right) = (-1.2, 1.2, 0.5)$$
$$\text{Chain 2}: \left(a_2^0, b_2^0, \sigma_2^0\right) = (-1, 1.4, 0.12)$$
$$\text{Chain 3}: \left(a_3^0, b_3^0, \sigma_3^0\right) = (-0.9, 1.6, 0.14)$$
$$\text{Chain 4}: \left(a_4^0, b_4^0, \sigma_4^0\right) = (-0.7, 1.8, 0.16)$$
$$\text{Chain 5}: \left(a_5^0, b_5^0, \sigma_5^0\right) = (-0.5, 2, 0.18)$$

We use a multivariate normal random walk proposal for parameters a, b and $\log\sigma$, presuming zero correlations between the latter and the former two parameters ($\rho_{a,\sigma} = \rho_{b,\sigma} = 0$). The scale parameters of the proposal distribution are set equal to the standard deviation estimates acquired from the initial M-H run, $\widehat{\sigma}_a = 0.109$, $\widehat{\sigma}_b = 0.06$, $\widehat{\sigma}_\sigma = 0.021$ and the correlation coefficient $\rho_{a,b}$ is set equal to -0.99, thus we presume a very strong negative correlation between a and b.

Convergence of the five parallel chains can be monitored through the first series of plots presented in Figure 3.3. We notice that the ergodic mean of each chain converges to a value which is common for all chains, as the number of iterations becomes large. Also, the histograms of the 40000 draws from the posterior distribution, in Figure 3.3, provide us a clear image of the parameters posterior space.

Point estimates, posterior quantiles and the $\hat{R}$ root reduction measure for each parameter are summarized in Tables 3.3 and 3.4, presented below. We notice that the calculation of the $\hat{R}$ root measure is close to unity for all parameters, which implies successful convergence of the algorithm.

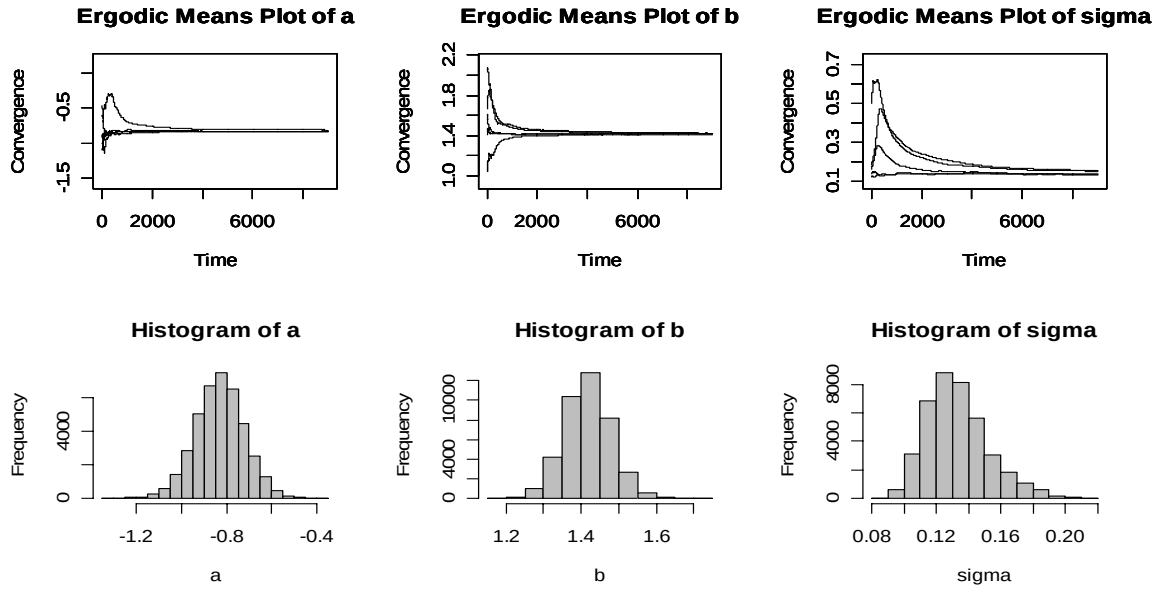

Figure 3.3 *Ergodic mean plots (up) for five parallel M-H chains each of size 9000 and histograms (down) of parameters a, b, and σ (sigma) acquired by an M-H sample of 40000 draws.*

Parameter	***a***	***b***	***σ***
Mean	-0.831	1.416	0.133
St. Deviation	0.109	0.061	0.019

Table 3.3 *Mean and standard deviation estimates for parameters a, b and σ.*

Parameter	**Posterior Quantiles**					
	0%	**25%**	**Median**	**75%**	**100%**	**R root**
a	-1.311	-0.902	-0.831	-0.759	-0.393	1.001
b	1.153	1.376	1.416	1.457	1.701	1.000
σ	0.086	0.119	0.131	0.144	0.217	1.002

Table 3.4 *Posterior quantiles and the calculated R root reduction measure for each parameter.*

The five parallel chains achieve acceptance ratios of 62.4%, 62.3%, 61.5%, 60.9% and 62.5% respectively.

3.5.2 Gibbs Sampler implementation

The normal setting used previously is one of the special cases for which the Gibbs sampler algorithm can be utilized, since all of the full conditionals are of known form. The prior design is the same as before with $a \sim N(0, k\sigma^2)$, $b \sim N(0, k\sigma^2)$, $\sigma^2 \sim IG(a_0, b_0)$ and prior parameters k, a_0, b_0 set equal to 1000, 10^{-3} and 10^{-3} respectively. The full conditional distributions for parameters a, b and σ are

$$p(a \mid b, \sigma^2, \mathbf{y}) \sim N\left(w_a(\bar{y} - b\bar{x}), w_a \frac{\sigma^2}{n} \right), \text{ with } w_a = \frac{k}{k + n^{-1}}$$

$$p(b \mid a, \sigma^2, \mathbf{y}) \sim N\left(w_b \frac{\sum_{i=1}^{n} x_i y_i - an\bar{x}}{\sum_{i=1}^{n} x_i^2}, w_b \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \right), \text{ with } w_b = \frac{k}{k + \left(\sum_{i=1}^{n} x_i^2 \right)^{-1}}$$

$$p(\sigma^2 \mid a, b, \mathbf{y}) \sim IG\left(a_0 + \frac{n}{2}, b_0 + \frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{2} \right).$$

We will use 5 parallel chains of 5000 iterations and discard the first 1000 values. To demonstrate the efficiency of Gibbs sampling, we use some initial values which are located in extreme regions of posterior space.

$$\text{Chain 1} : \left(a_1^0, b_1^0, \sigma_1^0\right) = (-4, -5, 0)$$
$$\text{Chain 2} : \left(a_2^0, b_2^0, \sigma_2^0\right) = (-2, -3, 0.5)$$
$$\text{Chain 3} : \left(a_3^0, b_3^0, \sigma_3^0\right) = (2, 0, 1)$$
$$\text{Chain 4} : \left(a_4^0, b_4^0, \sigma_4^0\right) = (3, 1, 1.5)$$
$$\text{Chain 5} : \left(a_5^0, b_5^0, \sigma_5^0\right) = (4, 3, 2)$$

Descriptive statistics, posterior quantiles and the $\hat{R}$ root reduction measure are shown in Tables 3.5 and 3.6. As we can see the resulting estimates for parameters a and b are quite similar to those acquired by the M-H algorithm. This is not the case for

estimates of σ. As we can see the estimates of the mean and of posterior quantiles are larger than the M-H estimates.

Parameter	*a*	*b*	*σ*
Mean	-0.833	1.418	0.142
St. Deviation	0.114	0.064	0.022

Table 3.5 *Mean and standard deviation estimates acquired by a Gibbs sample of size 20000.*

Parameter	Posterior Quantiles					
	0%	25%	Median	75%	100%	R root
a	-1.316	-0.906	-0.833	-0.759	-0.366	1.001
b	1.141	1.377	1.418	1.459	1.691	1.001
σ	0.087	0.127	0.139	0.155	0.274	0.999

Table 3.6 *Posterior quantiles and the calculated R root reduction measure for parameters a, b and σ.*

Ergodic mean plots and histograms of the 20000 draws from the Gibbs sample are presented below, in Figure 3.4.

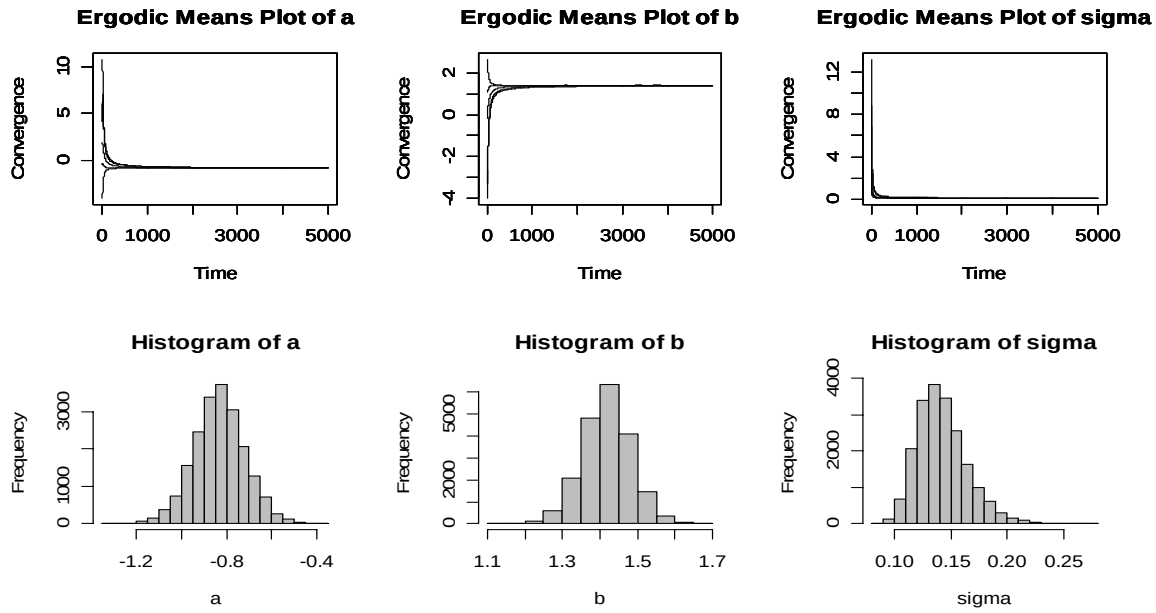

Figure 3.4 *Ergodic mean plots (up) for five parallel Gibbs chains each of size 5000 and histograms (down) for parameters a, b and σ (sigma) resulting from a Gibbs sample of size 20000.*

Scatter plots of parameters a and b, resulting from both MCMC methods, are presented in Figure 3.5.

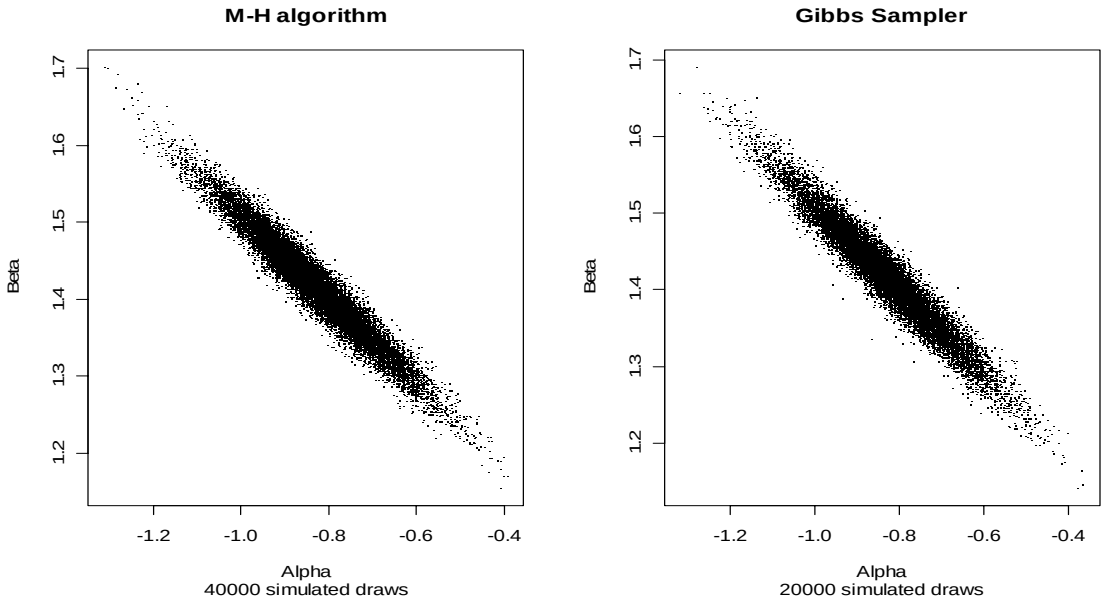

Figure 3.5 *Scatter plots of parameters a and b resulting from 40000 M-H draws (left) and from 20000 Gibbs draws (right).*

3.5.3 Informal model checking through plots of predicted values

One first approach on checking discrepancy between the fitted model and the observed data according to Gelman *et al*. (1995, p.167), is to simulate draws from the predictive distribution and then compare them with the observed data. Systematic differences between observed and replicated data imply potential failures of the model.

In our example, simulating from the predictive distribution is easy; for every set of $\left(a^l, b^l, \sigma^l\right)$, with $l = 1,...,20000$, acquired from the Gibbs sampler we simulate $\mathbf{y}^{rep(l)} = \left(y_1^{rep(l)}, y_2^{rep(l)}, ..., y_{25}^{rep(l)}\right)$ from $N\left(a^{(l)} + b^{(l)}\mathbf{x}, \sigma^{2(l)}\right)$. We then select in random manner 20 replications ($\mathbf{y}^{rep}$). The histograms of these replicated datasets are presented in Figure 3.6. We notice that most of them are skewed to the right.

In addition, six out of twenty histograms seem quite similar to that of the observed data. Thus, draws from the predictive distribution indicate no evidence of systematic differences between what is observed and what is predicted through our model.

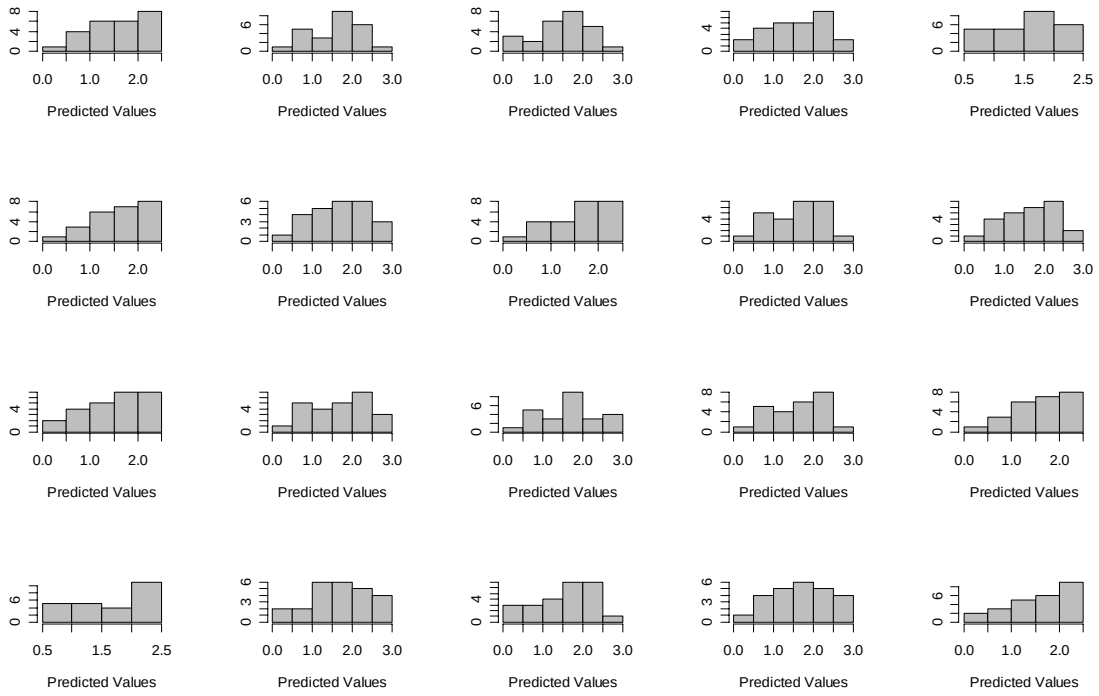

Figure 3.6 *Histograms of 20 replicated datasets. Most of the histograms are skewed to the right similar to the histogram of observed data.*

A more formal way of examining model discrepancy and lack of fit is through the use of test quantities; see section 2.6. We will see how test quantities can be utilized in the subsequent logistic regression example.

3.6 A logistic regression example with binomial data

The data in this second example are presented by Cox and Snell (1989). They concern number of deaths, during the period 1950-1959 caused from leuchaimia and other types of cancer, for the survivors of Hiroshima who where between 25 and 64 years old at year 1950. The death occurrences are presented according to

the dose of radiation which is originally given in unequal intervals. The data are shown below in Table 3.7.

Dose of Radiation (rads)	**Intevals**	*0*	*1-9*	*10-49*	*50-99*	*100-200*	*200+*
	Points	*0*	*4.5*	*29.5*	*74.5*	*149.5*	*249.5*
Number of Deaths	**Leuchamia**	*13*	*5*	*5*	*3*	*4*	*18*
	Other Cancer	*378*	*200*	*151*	*47*	*31*	*33*
	Total	*391*	*205*	*156*	*50*	*35*	*51*

Table 3.7 *Numbers of deaths caused from leuchaimia and other types of cancer according to the dose of radiation. The dose of radiation is measured in rads and given in unequal intervals in the original dataset (1st row). For the logistic regression we actually use points that lie near the center of the intervals (2nd row).*

A common model used for such types of problems is the logit model. For y being the dependent variable, that is number of deaths caused from leuchaimia, and x the explanatory variable, dose of radiation, then

$$\text{logit}(p_i) = a + bx_i \text{, with } y_i \mid p_i \sim Bin(n_i, p_i) \text{ for } i = 1,...,6 .$$

The probability of dying from leuchamia is p_i, while n_i is the total number of deaths caused by cancer, the subscript $i = 1,...,6$ represents the corresponding dose of radiation. The term $\text{logit}(p_i)$ is the logarithm of the *odds* of the unknown binomial probabilities p_i that is,

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) \text{ for } i = 1,...,6 .$$

In this dataset the explanatory variable is given in unequal intervals, therefore we will actually use the points that lie near the center of each interval shown in Table 3.7. The ML estimates of a and b are

$$\hat{a} = -3.564 \ (0.212) \text{ and } \hat{b} = 0.011 \ (0.001),$$

with the values in brackets corresponding to the coefficients standard deviations.

3.6.1 Metropolis algorithm implementation

For the model $\log it(p_i) = a + bx_i$, $i = 1,...,6$ we use the independent priors $a \sim N(0,k)$ and $b \sim N(0,k)$. The multiplying constant k is set again equal to 1000 in order to reflect prior ignorance. The joint posterior distribution is,

$$\begin{aligned}
p(a,b \mid \mathbf{y}) &\propto p(a,b)p(\mathbf{y} \mid a,b) = p(a)p(b)p(\mathbf{y} \mid a,b) \\
&\propto \exp\left\{-\frac{a^2+b^2}{2k}\right\}\prod_{i=1}^{6} p_i^{y_i}(1-p_i)^{n_i-y_i} \\
&\propto \exp\left\{-\frac{a^2+b^2}{2k}\right\}\prod_{i=1}^{6} p_i^{y_i}(1-p_i)^{n_i-y_i} \\
&\propto \exp\left\{-\frac{a^2+b^2}{2k}\right\}\prod_{i=1}^{6}\left(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{y_i}\left(1-\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{n_i-y_i} \\
&\propto \exp\left\{-\frac{a^2+b^2}{2k}\right\}\prod_{i=1}^{6}\left(\frac{e^{a+bx_i}}{1+e^{a+bx_i}}\right)^{y_i}\left(\frac{1}{1+e^{a+bx_i}}\right)^{n_i-y_i} \\
&\propto \exp\left\{-\frac{a^2+b^2}{2k}\right\}\prod_{i=1}^{6}\left(e^{a+bx_i}\right)^{y_i}\left(\frac{1}{1+e^{a+bx_i}}\right)^{n_i}.
\end{aligned}$$

On logarithmic scale the expression simplifies to

$$\log p(a,b \mid \mathbf{y}) = -\frac{a^2+b^2}{2k} + a\sum_{i=1}^{6} y_i + b\sum_{i=1}^{6} y_i x_i - \sum_{i=1}^{6} n_i \log\left(1+\exp\{a+bx_i\}\right) + \text{Constant.}$$

Presuming that we have no prior knowledge of the parameters posterior space, we will generate three parallel chains of size 20000, from the following starting points:

$$\begin{aligned}
&\text{Chain 1}: \left(a_1^0, b_1^0\right) = (-3,-1) \\
&\text{Chain 2}: \left(a_2^0, b_2^0\right) = (0,0) \\
&\text{Chain 3}: \left(a_3^0, b_3^0\right) = (3,1).
\end{aligned}$$

We use a bivariate normal random walk proposal, that is

$$\begin{pmatrix} a^t \\ b^t \end{pmatrix} \sim \mathbf{N}_2\left(\begin{pmatrix} a^{t-1} \\ b^{t-1} \end{pmatrix}, \boldsymbol{\Sigma}\right), \text{ with } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix}.$$

Determining the scale of the proposal density will be more difficult, in absence of an initial run. In our previous example scale parameters σ_a^2 and σ_b^2 were set equal to 1 and later on they were decreased in order to increase acceptance ratio. In addition, the initial values set for parameters a and b, where not extreme. As we have seen, this resulted to a relatively low acceptance ratio equal to 9.6%, therefore such a strategy will be probably inefficient if our main purpose is inferential.

We initially set $\sigma_a^2 = \sigma_b^2 = 25$ and $\rho = 0$; the scale selection seems large enough in order to explore the posterior space with adequacy, while correlation is set equal to zero in the absence of any prior knowledge. When each chain reaches the first fourth of its course, that is at iteration 5000, we use the past 5000 draws to obtain estimates of σ_a, σ_b and ρ. Then, the proposal distribution for each separate chain becomes

$$\mathbf{N}_2\left(\begin{pmatrix} a^{t-1}{}_{(i)} \\ b^{t-1}{}_{(i)} \end{pmatrix}, \boldsymbol{\Sigma}_{(i)}\right), \text{ with } \boldsymbol{\Sigma}_{(i)} = \begin{pmatrix} \widehat{\sigma^2_{a\,(i)}} & \hat{\rho}_{(i)}\hat{\sigma}_{a(i)}\hat{\sigma}_{b(i)} \\ \hat{\rho}_{(i)}\hat{\sigma}_{a(i)}\hat{\sigma}_{b(i)} & \widehat{\sigma^2_{a\,(i)}} \end{pmatrix} \text{and } i = 1,2,3.$$

Estimates $\widehat{\sigma_a^2}$, $\widehat{\sigma_b^2}$ and $\hat{\rho}$ are updated twice again, at iterations 10000 and 15000, from the corresponding 5000 previous draws.

The probability of transition is affected only from the unnormalized posterior distribution since a normal random walk proposal is used. Thus, it is given by

$$a_M = \frac{p\left(a^*, b^* \mid \mathbf{y}\right)}{p\left(a^{t-1}, b^{t-1} \mid \mathbf{y}\right)}.$$

Calculated on logarithmic scale the probability of move is given by

$$\log a_M = \log p\left(a^*, b^* \mid \mathbf{y}\right) - \log p\left(a^{t-1}, b^{t-1} \mid \mathbf{y}\right).$$

We keep the second half of each chain for inferential purposes. The resulting acceptance ratios are 23.5%, 28.6% and 25.6% respectively. Descriptive statistics, posterior quantiles and the $\hat{R}$ root measure for parameters α and b are presented in Tables 3.8 and 3.9, ergodic means plots and histograms are presented in Figure 3.7.

Parameter	***a***	***b***
Mean	-3.586	0.012
St. Deviation	0.215	0.001

Table 3.8 *Mean and standard deviation estimates for parameters a and b acquired by a Metropolis sample of size 30000.*

Parameter	**Posterior Quantiles**					
	0%	**25%**	**Median**	**75%**	**100%**	**R root**
a	-4.491	-3.731	-3.581	-3.437	-2.846	1.0002
b	0.006	0.011	0.012	0.013	0.017	1.0000

Table 3.9 *Posterior quantiles and the calculated R root reduction measure for each parameter of interest.*

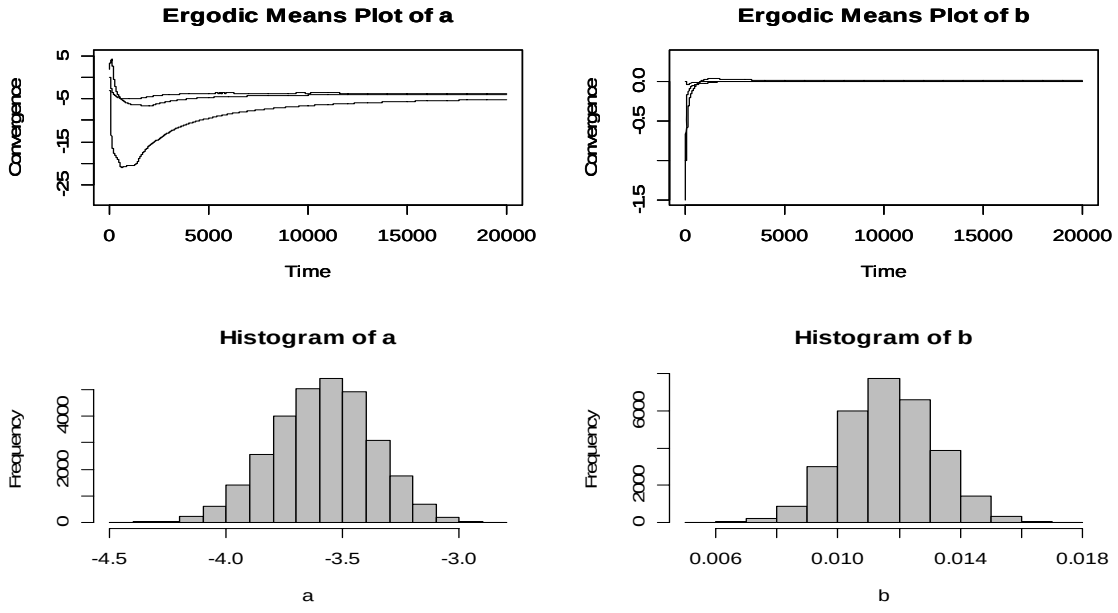

Figure 3.7 *Ergodic mean plots (up) for parameters a and b from three parallel Metropolis chains of 20000 iterations each. Histograms (down) for parameters a, b resulting from a Metropolis sample of 30000 draws.*

Point estimates of the means and standard deviations are almost identical to the ML estimates while the correlation estimate equals -0.668. Overall convergence, according to the calculation of the $\hat{R}$ roots and the examination of the ergodic mean plots presented in Figure 3.7, seems satisfactory.

An acceptance ratio of approximately 80% can be achieved by setting $\sigma_a = 0.1,\ \sigma_b = 0.0005$ and $\rho = -0.66$ along with initial values located within the posterior space.

3.6.2 Metropolis within Gibbs implementation

Now we will demonstrate the Metropolis within Gibbs algorithm for the same logistic regression example. The term Univariate Metropolis is in this case a more proper description of the algorithm in use, since none of the two full conditionals are of known form. So, we will actually use two separate Metropolis algorithms for parameters *a* and *b*. As we have seen, the joint posterior distribution is

$$p(a,b \mid \mathbf{y}) \propto \exp\left\{-\frac{a^2+b^2}{2k}\right\} \prod_{i=1}^{6} \left(e^{a+bx_i}\right)^{y_i} \left(\frac{1}{1+e^{a+bx_i}}\right)^{n_i}.$$

Thus, the full conditionals of parameters *a* and *b* are

$$p(a \mid b,\mathbf{y}) \propto \exp\left\{-\frac{a^2}{2k}\right\} \prod_{i=1}^{6} \exp\{ay_i\}\left(1+\exp\{a+bx_i\}\right)^{-n_i},$$

$$p(b \mid a,\mathbf{y}) \propto \exp\left\{-\frac{b^2}{2k}\right\} \prod_{i=1}^{6} \exp\{by_i x_i\}\left(1+\exp\{a+bx_i\}\right)^{-n_i}$$

and on logarithmic scale

$$\log p(a \mid b, \mathbf{y}) = -\frac{a^2}{2k} + a\sum_{i=1}^{6} y_i - \sum_{i=1}^{6} n_i \log\left(1+\exp\{a+bx_i\}\right) + \text{Constant},$$

$$\log p(b \mid a, \mathbf{y}) = -\frac{b^2}{2k} + b\sum_{i=1}^{6} y_i x_i - \sum_{i=1}^{6} n_i \log\left(1+\exp\{a+bx_i\}\right) + \text{Constant}.$$

We now use univariate proposal densities $a^t \sim N(a^{t-1}, \sigma_a^2)$ and $b^t \sim N(b^{t-1}, \sigma_b^2)$, with $\sigma_a^2 = \sigma_b^2 = 25$. The updating scheme presented above is now utilized for each parameter separately. The probabilities of transition on logarithmic scale are

$$\log a_M = \log p\left(a^* \mid b^{t-1}, \mathbf{y}\right) - \log p\left(a^{t-1} \mid b^{t-1}, \mathbf{y}\right)$$
$$\log a_M = \log p\left(b^* \mid a^*, \mathbf{y}\right) - \log p\left(b^{t-1} \mid a^*, \mathbf{y}\right).$$

Three parallel chains of size 20000 with the same initial values produce acceptance ratios of 41.3%, 42%, 42.4% for parameter *a* and 32.1%, 38.7%, 33.6% for parameter *b*. Results are summarized in Tables 3.10, 3.11 and Figure 3.8.

Parameter	***a***	***b***
Mean	-3.581	0.012
St. Deviation	0.210	0.001

Table 3.10 *Mean and standard deviation estimates resulting from a Metropolis within Gibbs simulation of size 30000.*

Parameter	**Posterior Quantiles**					
	0%	**25%**	**Median**	**75%**	**100%**	**R root**
a	-4.550	-3.722	-3.575	-3.434	-2.924	0.9999
b	0.005	0.011	0.012	0.013	0.018	1.0000

Table 3.11 *Posterior quantiles and the calculated R root reduction measure for parameters a and b.*

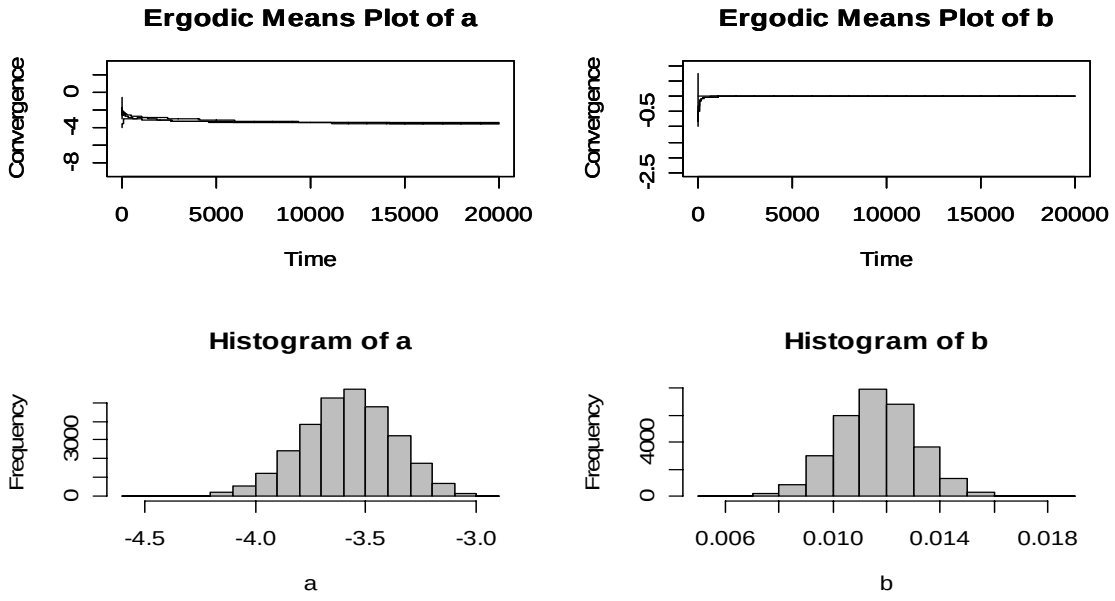

Figure 3.8 *Ergodic mean plots (up) for parameters a, b from three parallel Metropolis within Gibbs chains of 20000 iterations each. Histograms (down) for parameters a, b resulting from a simulated sample of size 30000.*

Acceptance ratios of approximately 80% for parameter a and 85% for parameter b can be achieved, by setting initial values located within the posterior space and initial scale parameters $\sigma_a = 0.1,\ \sigma_b = 0.0005$.

Estimates of means, standard deviations and quantiles obtained from the two algorithms are similar. This does not hold for the correlation estimates; from the Metropolis algorithm we obtain $\widehat{\rho}_M = -0.668$, while from Metropolis within Gibbs we obtain $\widehat{\rho}_{MG} = -0.606$. In order to check the efficiency of these estimates we run a single chain of 50000 iterations using both methods and conclude to $\widehat{\rho}_M \simeq \widehat{\rho}_{MG} \simeq -0.67$, a value not far from the initial estimate of the Metropolis algorithm. This non negligible difference between the two estimates can be noticed in the scatter plots of Figure 3.9.

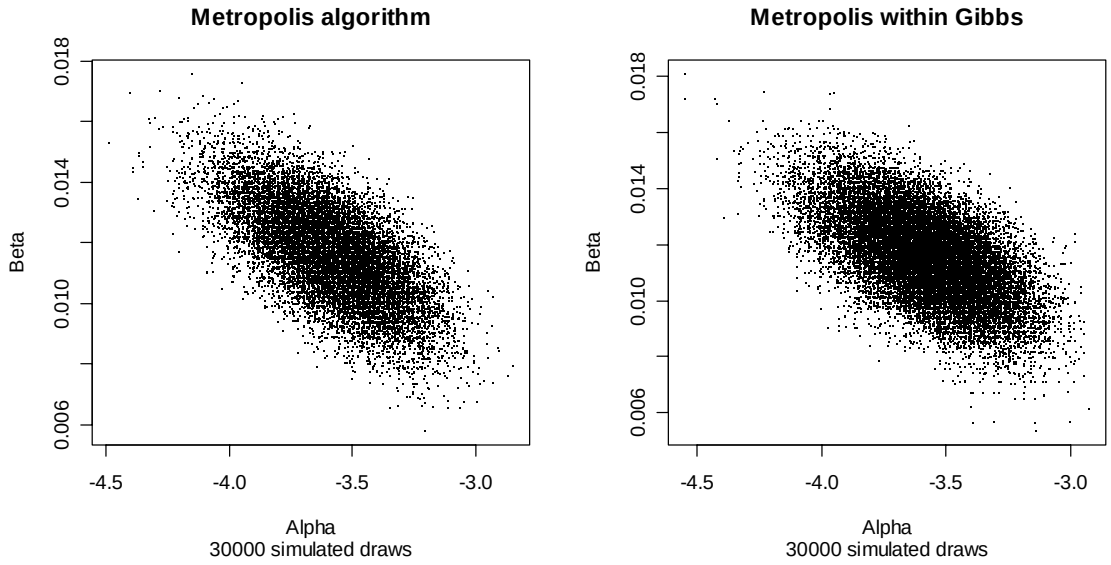

Figure 3.9 *Scatter plots for parameters a, b resulting from 30000 Metropolis draws (left) and 30000 Metropolis within Gibbs draws (right).*

3.6.3 Checking model discrepancies through test quantities

We will now use the Metropolis sample in order to evaluate the *chi square* and the *deviance* test quantities.

As discussed in section 2.6 the chi-square discrepancy measure suggested by Gelman et al. (1993) is given by

$$T(\mathbf{y},\boldsymbol{\theta}) = \sum_i \frac{(y_i - E(y_i \mid \boldsymbol{\theta}))^2}{Var(y_i \mid \boldsymbol{\theta})}.$$

Under the model hypothesis the parameter vector is $\boldsymbol{\theta} = \mathbf{p}$, with

$$\mathbf{p} = \frac{\exp\{a + b\mathbf{x}\}}{1 + \exp\{a + b\mathbf{x}\}}.$$

So, the chi-square test quantity has the form

$$T(\mathbf{y},\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(\frac{y_i - n_i p_i}{n_i p_i (1 - p_i)} \right)^2,$$

since $E(y_i \mid \boldsymbol{\theta}) = n_i p_i$ and $Var(y_i \mid \boldsymbol{\theta}) = n_i p_i (1 - p_i)$.

Another test quantity which is frequently used for overall goodness-of-fit checks is the deviance quantity. As already mentioned in section 2.8 the deviance of any model is generally obtained from $D(\boldsymbol{\theta}) = -2\log p(\mathbf{y}\,|\,\boldsymbol{\theta})$. Thus, in general the deviance quantity is given by

$$T(\mathbf{y},\boldsymbol{\theta}) = -2\log p(\mathbf{y}\,|\,\boldsymbol{\theta}).$$

In binomial logistic regression the deviance from the *maximal* model can be calculated analytically (Dobson, 1990, p.112), so we actually evaluate the deviance quantity from

$$T(\mathbf{y},\boldsymbol{\theta}) = \sum_i \left[y_i \log\left(\frac{y_i}{n_i p_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i p_i}\right)\right].$$

We evaluate the Bayesian p-value, defined as $\Pr\left(T(\mathbf{y}^{rep},\boldsymbol{\theta}) \geq T(\mathbf{y},\boldsymbol{\theta})\,|\,\mathbf{y}\right)$ (see section 2.6), through simulation. For this, we initially generate draws of replicated data; for each set of parameters $\left(a^l, b^l\right)$, $l = 1,...,30000$, we first calculate $\mathbf{p}^l = \dfrac{\exp\left\{a^l + b^l\mathbf{x}\right\}}{1+\exp\left\{a^l + b^l\mathbf{x}\right\}}$ and then we generate $\mathbf{y}^{rep(l)} = \left(y_1^{rep(l)}, y_2^{rep(l)},..., y_6^{rep(l)}\right)$ from $Bin(\mathbf{n},\mathbf{p}^l)$. Finally, we evaluate the test quantities $T(\mathbf{y},\boldsymbol{\theta}^l)$ and $T(\mathbf{y}^{rep(l)},\boldsymbol{\theta}^l)$ for $l = 1,...,30000$ and then obtain the corresponding Bayesian p-values by simply counting the number of cases for which, inequality $T(\mathbf{y}^{rep(l)},\boldsymbol{\theta}^l) \geq T(\mathbf{y},\boldsymbol{\theta}^l)$ holds.

The resulting p-values are 0.589 for the chi-square test quantity and 0.802 for the deviance test quantity. The chi square p-value implies that observed and replicated data seem to be in agreement. The p-value obtained from the deviance test is larger, yet it cannot be considered as an extreme tail area probability. Therefore, we could say that the model is not suspicious for major discrepancies. Kernel smoothed densities of each test quantity are presented in Figure 3.10 below.

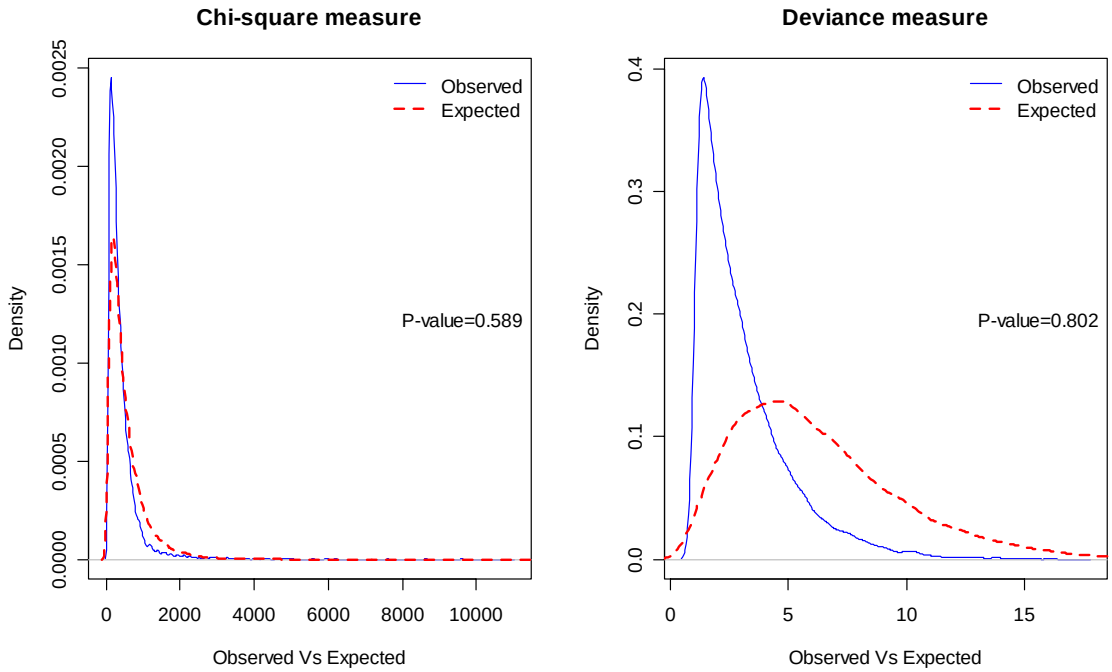

Figure 3.10 *Kernel smoothed densities of the chi-square (left) and deviance (right) test quantities derived from observed data (blue line) and replicated data (red dashed line).*

Chapter 4: Marginal Likelihood Estimators

4.1 Introduction

As discussed in section 2.7, hypothesis testing and model comparison are based on the calculation of posterior odds ratios and consequently on Bayes factors. In order to calculate Bayes factors we must obtain the marginal likelihood of the data under each competing model. Without the notational dependence from the model, the marginal density of the data $\mathbf{y}$ with parameter vector $\boldsymbol{\theta}$ is given by

$$p(\mathbf{y}) = \int_{\Theta} p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} .$$

Most often this integration cannot be derived analytically, so we usually have to obtain an estimate of the marginal likelihood using alternative approaches such as asymptotic or simulation based methods.

One can also notice that the integral is with respect to the prior distribution of $\boldsymbol{\theta}$ and not with respect to the posterior distribution of $\boldsymbol{\theta}$. Thus, the simplest Monte Carlo integration estimate is given by

$$\hat{p}(\mathbf{y}) = M^{-1} \sum_{m=1}^{M} p(\mathbf{y} \mid \boldsymbol{\theta}_m) ,$$

where the values $\{\boldsymbol{\theta}_m\}$, for $m = 1,...,M$, are now a sample from the prior distribution. Unfortunately, this estimate is very unstable when the posterior distribution is concentrated relative to the prior which is most often the case. Thus, this estimate is dominated by a few values of $\boldsymbol{\theta}$ which have a large likelihood value (Kass and Raftery, 1995).

In this chapter we will focus on "direct" MCMC methods. These methods utilize MCMC outputs from separate models in order to acquire the estimates of marginal likelihoods and consequently estimates of Bayes factors. Alternative options are asymptotic approximations to the marginal likelihood such as the *Laplace method* or the *Schwarz Criterion* which can be used as an approximation for Bayes factors; for details see Kass and Raftery (1995). There also exist MCMC model selection methods which simulate over both parameter and model

space such as the *Reversible Jump MCMC* (RJMCMC) algorithm (Green, 1995), the *Carlin and Chib* algorithm (Carlin and Chib, 1995) and the *Metropolised Carlin and Chib* algorithm (Dellaportas et al., 2002). These methods bypass marginal likelihoods and deliver directly the posterior probabilities of each model. Recently, Congdon (2004) also presented a method of estimating posterior model probabilities.

4.2 Harmonic mean estimator

The harmonic mean estimator is based on the importance sampling method and uses the posterior as the importance sampling function (Newton and Raftery 1994). The marginal density can be expressed as

$$p(\mathbf{y}) = \frac{1}{p(\mathbf{y})^{-1}} = \frac{1}{\int p(\mathbf{y})^{-1} p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{1}{\int \frac{p(\mathbf{y})^{-1}}{p(\boldsymbol{\theta} \mid \mathbf{y})} p(\boldsymbol{\theta} \mid \mathbf{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{1}{\int p(\mathbf{y} \mid \boldsymbol{\theta})^{-1} p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}}.$$

Thus, we have the identity $p(\mathbf{y}) = \left\{ \int p(\mathbf{y} \mid \boldsymbol{\theta})^{-1} p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta} \right\}^{-1}$ which results to the harmonic mean estimator

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

where $\boldsymbol{\theta}^{(t)}$, for $t = 1, 2, ..., N$, are posterior draws from the MCMC output. According to Newton and Raftery (1994) this estimator converges surely to $p(\mathbf{y})$, as $t \to \infty$. However $\hat{p}_H$ does not, in general, satisfy a Gaussian central limit theorem, hence values $\boldsymbol{\theta}^{(t)}$ with small likelihood have a large effect on the final result.

Although the harmonic mean estimator is quite unstable, it is easy to calculate and according to Kass and Raftery (1995) produces results that are accurate enough for interpretation on logarithmic scale.

4.3 Laplace-Metropolis estimator

The Laplace-Metropolis estimator combines the asymptotic result of the Laplace method with the MCMC output. Under the assumptions of the Laplace method the marginal density can be approximated by

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

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode, $\tilde{\boldsymbol{\Sigma}}$ is minus the inverse Hessian matrix evaluated at $\tilde{\boldsymbol{\theta}}$ and d is the dimension of $\boldsymbol{\theta}$ (Tierney and Kadane, 1986; Kass and Raftery, 1995).

The Laplace-Metropolis estimator $\hat{p}_{LM}$ discussed in Lewis and Raftery (1997) is given by the above equation with $\tilde{\boldsymbol{\theta}}$ estimated by $\boldsymbol{\theta}^{\max}$ the point that maximizes $p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)})$ among the N posterior draws and $\hat{\boldsymbol{\Sigma}}$ estimated by $\mathbf{S}$ the sample covariance matrix of the output. Hence, the Laplace-Metropolis estimator is given by

$$\hat{p}_{LM} = (2\pi)^{d/2} \left|\mathbf{S}\right|^{1/2} p(\mathbf{y} \mid \boldsymbol{\theta}^{\max}) p(\boldsymbol{\theta}^{\max}),$$

where $\boldsymbol{\theta}^{\max} = \left\{ \boldsymbol{\theta} : p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \max_{t=1,\ldots,N} \left\{ p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)}) \right\} \right\}$ and

$$\mathbf{S} = \frac{1}{N-1} \sum_{t=1}^{N} \left(\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}} \right) \left(\boldsymbol{\theta}^{(t)} - \bar{\boldsymbol{\theta}} \right)^{T}.$$

Alternative choices for $\tilde{\boldsymbol{\theta}}$ can be obtained by the multivariate posterior median or by a nonparametric density estimate of the posterior mode or even by the posterior mean when the posterior is a symmetric distribution. According to Lewis and Raftery (1997) the resulting estimator performed well in numerical experiments.

4.4 Newton and Raftery's estimator

Newton and Raftery (1994) suggested an estimator, based again on the importance sampling method, which would be less unstable than $\hat{p}_H$. The importance sampling function is now a mixture of the prior and posterior densities

$$g(\boldsymbol{\theta}) = \delta p(\boldsymbol{\theta}) + (1-\delta) p(\boldsymbol{\theta} \mid \mathbf{y}),$$

where $0 < \delta < 1$. One can notice that the choice $\delta = 0$ results to harmonic mean estimator. Sampling from g is achieved by randomly replacing $\delta \times N$ values of the posterior sample with independent draws from the prior (Lopes, 2002).

Using $g(\boldsymbol{\theta})$ as the importance sampling function results to the estimate $\hat{p}_{NR(1)}$. In order to obtain $\hat{p}_{NR(1)}$ we must first specify an initial value and then iterate the equation

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

According to Newton and Raftery (1994) this estimator retains the efficiency of $\hat{p}_H$ but it also satisfies a Gaussian central limit theorem, unlike the latter.

In order to avoid sampling from both the prior and the posterior, Newton and Raftery (1994) suggest to use all N values of the posterior sample and imaging that further $\delta N/(1-\delta)$ values are drawn from the prior, all with likelihoods $p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)})$ equal to their expected value $p(\mathbf{y})$. This yields the estimator $\hat{p}_{NR(2)}$ which is obtained by iterating the equation

$$\hat{p}_{NR(2)} = \frac{\delta N/(1-\delta) + \sum_{t=1}^{N} p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) \left\{ \delta \hat{p}_{NR(2)} + (1-\delta) p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) \right\}^{-1}}{\delta N/(1-\delta)\, \hat{p}_{NR(2)} + \sum_{t=1}^{N} \left\{ \delta \hat{p}_{NR(2)} + (1-\delta) p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) \right\}^{-1}}.$$

As with the first estimator, an initial value must be set in order to start the iterative process.

According to the authors both estimators performed well for values of δ as small as 0.01 without displaying the instability of $\hat{p}_H$.

4.5 Bridge sampling estimator

Innovative methods for computing the ratio of normalizing constants based on bridge sampling were studied by Meng and Wong (1996).

Suppose that we have two densities $p_i(\boldsymbol{\theta})$, $i = 1,2$ which are known up to a normalizing constant so that $p_i(\boldsymbol{\theta}) = \frac{q_i(\boldsymbol{\theta})}{c_i}$, $\boldsymbol{\theta} \in \Omega_i \subset R^d$, where Ω_i is the support of $p_i(\boldsymbol{\theta})$ and that our interest is in calculating the ratio c_1/c_2.

We additionally presume that $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$ have a common support that is,

$$\int_{\Omega_1 \cap \Omega_2} p_1(\boldsymbol{\theta}) p_2(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$$

and that samples from these densities are available. Then for any arbitrary function $a(\boldsymbol{\theta})$ defined on $\Omega_1 \cap \Omega_2$ which satisfies

$$0 < \left| \int_{\Omega_1 \cap \Omega_2} a(\boldsymbol{\theta}) p_1(\boldsymbol{\theta}) p_2(\boldsymbol{\theta}) d\boldsymbol{\theta} \right| < \infty,$$

we have

$$\frac{\int_{\Omega_2} a(\boldsymbol{\theta}) q_1(\boldsymbol{\theta}) p_2(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Omega_1} a(\boldsymbol{\theta}) q_2(\boldsymbol{\theta}) p_1(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{c_1}{c_2} \times \frac{\int_{\Omega_1 \cap \Omega_2} a(\boldsymbol{\theta}) p_1(\boldsymbol{\theta}) p_2(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Omega_1 \cap \Omega_2} a(\boldsymbol{\theta}) p_2(\boldsymbol{\theta}) p_1(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

which yields the key identity

$$\frac{c_1}{c_2} = \frac{E_2\left[a(\boldsymbol{\theta}) q_1(\boldsymbol{\theta})\right]}{E_1\left[a(\boldsymbol{\theta}) q_2(\boldsymbol{\theta})\right]},$$

where the expectation in the numerator is with respect to $p_2(\boldsymbol{\theta})$ and the expectation in the denominator is with respect to $p_1(\boldsymbol{\theta})$.

In our context we choose a density $g(\boldsymbol{\theta})$ which has the same support as the posterior. So, the corresponding densities are

$$p_1(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}),\ q_1(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),\ \text{with}\ c_1 = p(\mathbf{y})$$

and

$$p_2(\boldsymbol{\theta}) = q_2(\boldsymbol{\theta}) = g(\boldsymbol{\theta}),\ \text{with}\ c_2 = 1.$$

Thus, we have

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

from which we obtain the estimate

$$\hat{p}(\mathbf{y}) = \frac{L^{-1} \sum_{l=1}^{L} a(\boldsymbol{\theta}^{*(l)}) p(\mathbf{y} \mid \boldsymbol{\theta}^{*(l)}) p(\boldsymbol{\theta}^{*(l)})}{N^{-1} \sum_{t=1}^{N} a(\boldsymbol{\theta}^{(t)}) g(\boldsymbol{\theta}^{(t)})}.$$

The values $\boldsymbol{\theta}^{(t)}$ are draws from the available MCMC sample, while the values $\boldsymbol{\theta}^{*(l)}$ are a sample of size L drawn from $g(\boldsymbol{\theta})$. In general, we want the density $g(\boldsymbol{\theta})$ to be an accurate approximation of the posterior and also easy to sample from (Lopes, 2002).

According to Meng and Wong (1996) different selections of the arbitrary function $a(\boldsymbol{\theta})$ produce different bridge sampling estimators. Some of these estimators, also reviewed by Lopes (2002), are the following:

- For $a(\boldsymbol{\theta}) = \{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) g(\boldsymbol{\theta})\}^{-1}$ we obtain an estimate which resembles the harmonic mean, that is

$$\hat{p}_{HB} = \frac{L^{-1} \sum_{l=1}^{L} g(\boldsymbol{\theta}^{*(l)})^{-1}}{N^{-1} \sum_{t=1}^{N} \{p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)})\}^{-1}}.$$

- For $a(\boldsymbol{\theta}) = \{p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})g(\boldsymbol{\theta})\}^{-1/2}$ we obtain the geometric estimator

$$\widehat{p}_{GB} = \frac{L^{-1}\sum_{l=1}^{L}\left\{p(\mathbf{y}\,|\,\boldsymbol{\theta}^{*(l)})p(\boldsymbol{\theta}^{*(l)})/g(\boldsymbol{\theta}^{*(l)})\right\}^{1/2}}{N^{-1}\sum_{t=1}^{N}\left\{g(\boldsymbol{\theta}^{(t)})/p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})\right\}^{1/2}}.$$

- The optimal estimator $\widehat{p}_{OptB}$ of Meng and Wong (1996) is obtained by an iterative procedure. We specify an initial value, usually $\widehat{p}_{OptB} = \widehat{p}_{GB}$, and iterate the equation

$$\widehat{p}_{OptB} = \frac{\sum_{l=1}^{L} W_2^{(l)} \Big/ \left(s_1 W_2^{(l)} + s_2 \widehat{p}_{OptB}\right)}{\sum_{t=1}^{N} 1 \Big/ \left(s_1 W_1^{(t)} + s_2 \widehat{p}_{OptB}\right)},$$

where $s_1 = N/(N+L),\ s_2 = L/(N+L),\ W_2^{(l)} = p(\mathbf{y}\,|\,\boldsymbol{\theta}^{*(l)})p(\boldsymbol{\theta}^{*(l)})/g(\boldsymbol{\theta}^{*(l)})$ and $W_1^{(t)} = p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})/g(\boldsymbol{\theta}^{(t)})$.

Additional information on other bridge sampling estimates and discussion over efficiency issues can be found in Meng and Wong (1996).

4.6 Candidate's estimator

The candidate's approach is based on a simple identity following from Bayes theorem that is

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

This alternative formula for calculating the marginal density of $\mathbf{y}$ was first mentioned in Besag (1989). The advantage of this formula is that it holds for any value $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$, thus an estimate of the marginal density on logarithmic scale is

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

We can notice that this expression requires only the evaluation of the log-likelihood function and the prior along with an estimate of the posterior density at point $\boldsymbol{\theta}^*$. According to Chib (1995) this estimate does not suffer from any instability problem since it is actually a density value that is averaged. In addition, the entire estimation error arises from the estimation of the posterior ordinate $\hat{p}(\boldsymbol{\theta}^* \mid \mathbf{y})$. Estimation of the posterior ordinate from a Gibbs output was fully analyzed by Chib (1995). Later, Chib and Jeliazkov (2001) presented a method for estimating the posterior ordinate from a Metropolis-Hastings output. These two methods are being reviewed next.

Regarding the selection of the point $\boldsymbol{\theta}^*$, Chib (1995) recommends choosing a high density value such as the posterior mode or the maximum likelihood estimate or even the posterior mean provided that it is not located in a low density region.

4.6.1 Marginal likelihood from the Gibbs output

The presentation of this method in Chib (1995) includes parameter blocks along with a latent data block. We will present the three vector blocks example of Chib (1995) by replacing the latent data block with a parameter block.

Suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, then the Gibbs sampler is defined through the full conditional densities

$$p(\boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3), \;\; p(\boldsymbol{\theta}_2 \mid \mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3), \;\; p(\boldsymbol{\theta}_3 \mid \mathbf{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

The goal is to estimate $p(\boldsymbol{\theta}^* \mid \mathbf{y})$ which can be expressed as

$$p(\boldsymbol{\theta}^* \mid \mathbf{y}) = p(\boldsymbol{\theta}_3^* \mid \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_1^*, \mathbf{y}) p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y}) p(\boldsymbol{\theta}_1^* \mid \mathbf{y}).$$

The first ordinate $p(\boldsymbol{\theta}_3^* \mid \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_1^*, \mathbf{y})$ can be calculated directly, since the full conditional density of $\boldsymbol{\theta}_3$ is known when using a Gibbs sampler. The other two ordinates can be expressed as

$$p(\boldsymbol{\theta}_1^* \mid \mathbf{y}) = \int p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \mathbf{y}) p(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3 \mid \mathbf{y}) d\boldsymbol{\theta}_2 d\boldsymbol{\theta}_3,$$

$$p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y}) = \int p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3, \mathbf{y}) p(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^*, \mathbf{y}) d\boldsymbol{\theta}_3.$$

The marginal ordinate $p(\boldsymbol{\theta}_1^* \mid \mathbf{y})$ can be estimated by averaging the full conditional density of $\boldsymbol{\theta}_1$ with the posterior draws of $(\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, this yields the estimate

$$\hat{p}(\boldsymbol{\theta}_1^* \mid \mathbf{y}) = N^{-1} \sum_{t=1}^{N} p(\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \mathbf{y}).$$

The reduced ordinate $p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y})$ can be estimated with a similar technique which requires draws from the distribution of $\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^*, \mathbf{y}$. Therefore, we continue sampling for additional *L* iterations from the full conditional densities

$$p(\boldsymbol{\theta}_2 \mid \mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3) \text{ and } p(\boldsymbol{\theta}_3 \mid \mathbf{y}, \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2),$$

where $\boldsymbol{\theta}_1$ is now constant, equal to $\boldsymbol{\theta}_1^*$. According to Chib (1995) the draws $\boldsymbol{\theta}_3^{(l)}$ from this reduced Gibbs run follow the density $p(\boldsymbol{\theta}_3 \mid \boldsymbol{\theta}_1^*, \mathbf{y})$. Thus, we can now estimate the reduced ordinate $p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y})$ with

$$\hat{p}(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^{(l)}, \mathbf{y}).$$

So, the marginal density estimate, denoted as $\hat{p}_{Chib}$, is

$$\hat{p}_{Chib} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*) p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}_3^* \mid \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_1^*, \mathbf{y}) \hat{p}(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y}) \hat{p}(\boldsymbol{\theta}_1^* \mid \mathbf{y})}$$

and on logarithmic scale

$$\log \hat{p}_{Gibbs} = \log p(\mathbf{y} \mid \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \log p(\boldsymbol{\theta}_3^* \mid \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_1^*, \mathbf{y}) - \log \hat{p}(\boldsymbol{\theta}_2^* \mid \boldsymbol{\theta}_1^*, \mathbf{y}) - \log \hat{p}(\boldsymbol{\theta}_1^* \mid \mathbf{y}).$$

Although this procedure leads to an increase in the number of iterations, it is rather straightforward since it does not require additional programming (Chib, 1995). Further note that the reduced Gibbs run is not necessary for two blocks

parametric vectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The general case for an arbitrary number of blocks $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_B)$ is treated accordingly; see Chib (1995).

4.6.2 Marginal likelihood from the Metropolis-Hastings output

Chib and Jeliazkov (2001) extended the previous method in order to be implemented in M-H output. They have illustrated the method for one parameter block, two parameter blocks along with multiple latent variable blocks and multiple parameter blocks. We will demonstrate the use of this method for the simple case of one block sampling.

Let $q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \mathbf{y})$ denote the proposal density for a transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$, where q is allowed to depend on the data $\mathbf{y}$, then the M-H transition probability is

$$a_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left[1, \frac{p(\boldsymbol{\theta}' \mid \mathbf{y}) q(\boldsymbol{\theta}', \boldsymbol{\theta} \mid \mathbf{y})}{p(\boldsymbol{\theta} \mid \mathbf{y}) q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \mathbf{y})}\right].$$

In addition, let $p(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \mathbf{y})$ denote the sub-kernel of the M-H algorithm that is

$$p(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \mathbf{y}) = a_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}') q(\boldsymbol{\theta}, \boldsymbol{\theta}' \mid \mathbf{y}),$$

then from the reversibility of this sub-kernel the equation

$$p(\boldsymbol{\theta} \mid \mathbf{y}) p(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \mathbf{y}) = p(\boldsymbol{\theta}^* \mid \mathbf{y}) p(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y}),$$

holds for any $\boldsymbol{\theta}^*$. By integrating both sides of this expression with respect to $\boldsymbol{\theta}$ we obtain that the posterior ordinate equals

$$p(\boldsymbol{\theta}^* \mid \mathbf{y}) = \frac{\int p(\boldsymbol{\theta} \mid \mathbf{y}) p(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \mathbf{y}) d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}} \Leftrightarrow$$

$$p(\boldsymbol{\theta}^* \mid \mathbf{y}) = \frac{\int p(\boldsymbol{\theta} \mid \mathbf{y}) a_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) q(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \mathbf{y}) d\boldsymbol{\theta}}{\int a_{MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}}.$$

This yields the estimator

$$\widehat{p}(\boldsymbol{\theta}^* \mid \mathbf{y}) = \frac{N^{-1}\sum_{t=1}^{N} a_{MH}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^* \mid \mathbf{y})}{J^{-1}\sum_{j=1}^{J} a_{MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)})},$$

where $\boldsymbol{\theta}^{(t)}$ are draws from the M-H output and $\boldsymbol{\theta}^{(j)}$ are draws from the distribution $q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y})$, given the fixed value $\boldsymbol{\theta}^*$.

According to Chib and Jeliazkov (2001) values $\boldsymbol{\theta}^{(j)}$ that do not lie in the support of $p(\boldsymbol{\theta} \mid \mathbf{y})$ are included in the average of the denominator with the value $a_{MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(j)}) = 0$. The authors also comment that although the sample sizes of the numerator and denominator are let to be different, in practice they are set to be equal and that sampling from $q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y})$ usually consumes a small amount of time.

Thus, we acquire a marginal likelihood estimate, denoted as $\widehat{p}_{C-J}$, which is given on logarithmic scale by

$$\log \widehat{p}_{C-J} = \log p(\mathbf{y} \mid \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \log \widehat{p}(\boldsymbol{\theta}^* \mid \mathbf{y}).$$

4.7 Chen's estimator

The method presented by Chen (2005) is in fact a generalization of the Candidate's estimator method. Unlike the two previous methods, Chen's method does not require the specific form of the MCMC sampling process to be known. Although the presentation of the method by Chen (2005) includes treatment of a latent data block, we restrict ourselves to the simple case (without latent data).

Let $g(\boldsymbol{\theta})$ be a proper density function, then for any point $\boldsymbol{\theta}^*$ the likelihood function $p(\mathbf{y} \mid \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^*$ can be expressed as

$$p(\mathbf{y} \mid \boldsymbol{\theta}^*) = \int p(\mathbf{y} \mid \boldsymbol{\theta}^*) g(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

since $\int g(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. This equality can be re-expressed as

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

Then, from the identity $p(\mathbf{y}) = p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}) / p(\boldsymbol{\theta}\,|\,\mathbf{y})$ we have that

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

Solving this equation with respect to $p(\mathbf{y})$ yields on logarithmic scale

$$\log p(\mathbf{y}) = \log p(\mathbf{y}\,|\,\boldsymbol{\theta}^*) - \log E\left[\frac{g(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \frac{p(\mathbf{y}\,|\,\boldsymbol{\theta}^*)}{p(\mathbf{y}\,|\,\boldsymbol{\theta})} \right],$$

where the expectation is with respect to the posterior distribution $p(\boldsymbol{\theta}\,|\,\mathbf{y})$. Thus, we obtain an estimate of the marginal likelihood, denoted as $\widehat{p}_{Chen}$, which is given by

$$\log \widehat{p}_{Chen} = \log p(\mathbf{y}\,|\,\boldsymbol{\theta}^*) - \log\left[N^{-1} \sum_{t=1}^{N} \frac{g(\boldsymbol{\theta}^{(t)})}{p(\boldsymbol{\theta}^{(t)})} \frac{p(\mathbf{y}\,|\,\boldsymbol{\theta}^*)}{p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})} \right],$$

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

According to Chen (2005) the density $g(\boldsymbol{\theta})$ corresponds to the weighted conditional density which is used in the *Importance Weighted Marginal Density* estimation method introduced by the same author (Chen, 1994). Therefore, the guidelines for choosing a satisfactory function g can be found in Chen (1994). A usual approach is to facilitate a common distribution which mimics the *conditional marginal density*; in our context, the posterior distribution $p(\boldsymbol{\theta}\,|\,\mathbf{y})$. This can be achieved by fitting posterior moments so that g will have a shape roughly similar to that of the posterior distribution.

An alternative option discussed in Chen (2005) is to choose $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ if $p(\boldsymbol{\theta})$ is proper. This choice simplifies the estimate to

$$\log \widehat{p}_{Chen} = \log p(\mathbf{y}\,|\,\boldsymbol{\theta}^*) - \log\left[N^{-1} \sum_{t=1}^{N} \frac{p(\mathbf{y}\,|\,\boldsymbol{\theta}^*)}{p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(t)})} \right],$$

but this is not an optimal choice according to the author.

The optimal choice of g discussed in Chen (2005) is $g(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y})$. In this case we have that

$$\log p(\mathbf{y}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}^*) - \log E\left[\frac{p(\boldsymbol{\theta} \mid \mathbf{y})}{p(\boldsymbol{\theta})} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*)}{p(\mathbf{y} \mid \boldsymbol{\theta})} \right] \Leftrightarrow$$

$$\log p(\mathbf{y}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}^*) - \log E\left[\frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}) p(\mathbf{y})} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*)}{p(\mathbf{y} \mid \boldsymbol{\theta})} \right] \Leftrightarrow$$

$$\log p(\mathbf{y}) = \log p(\mathbf{y} \mid \boldsymbol{\theta}^*) - \log E\left[\frac{p(\mathbf{y} \mid \boldsymbol{\theta}^*)}{p(\mathbf{y})} \right].$$

However, we know that $p(\mathbf{y} \mid \boldsymbol{\theta}^*) = p(\boldsymbol{\theta}^* \mid \mathbf{y}) p(\mathbf{y}) \big/ p(\boldsymbol{\theta}^*)$, therefore it easy to see that the optimal choice of Chen results to the familiar identity

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

Thus, Chen's optimal choice of g leads to the use of the Candidate's estimator method of Chib (1995) or Chib and Jeliazkov (2001) if $p(\boldsymbol{\theta}^* \mid \mathbf{y})$ cannot be directly calculated.

Chapter 5: Illustration and Comparison of Methods in a Simple Regression Example

5.1 Models and prior selection

In this chapter we will evaluate the methods discussed in the previous chapter on four competing regression models. The data are the DC output and wind velocity observations presented in section 3.5. The four models we wish to compare are

$$\begin{aligned} M_0 &: \; y_i = a_0 + \varepsilon_{0i} \\ M_1 &: \; y_i = a_1 + b_1(x_i - \bar{x}) + \varepsilon_{1i} \\ M_2 &: \; y_i = a_2 + b_2(z_i - \bar{z}) + \varepsilon_{2i} \\ M_3 &: \; y_i = a_3 + b_3(x_i - \bar{x}) + c_3 x_i^2 + \varepsilon_{3i}, \end{aligned}$$

where y_i is the DC output, x_i is the wind velocity, z_i is the logarithm of wind velocity and $\varepsilon_{ji} \overset{iid}{\sim} N(0, \sigma_j^2)$ with $i = 1,...,25$ and $j = 0,1,2,3$. The parameter vector $\boldsymbol{\theta}_j$ of each model is $\boldsymbol{\theta}_0 = (a_0, \sigma_0^2)$, $\boldsymbol{\theta}_1 = (a_1, b_1, \sigma_1^2)$, $\boldsymbol{\theta}_2 = (a_2, b_2, \sigma_2^2)$ and $\boldsymbol{\theta}_3 = (a_3, b_3, c_3, \sigma_3^2)$. The selected prior densities are

$$\begin{aligned} M_0 &: \; a_0 \sim N(0, \mathbf{D}_0\sigma_0^2), \;\; \sigma_0^2 \sim IG(10^{-3}, 10^{-3}) \\ M_1 &: \; (a_1, b_1)^T \sim N_2(\mathbf{0}, \mathbf{D}_1\sigma_1^2), \;\; \sigma_1^2 \sim IG(10^{-3}, 10^{-3}) \\ M_2 &: \; (a_2, b_2)^T \sim N_2(\mathbf{0}, \mathbf{D}_2\sigma_2^2), \;\; \sigma_2^2 \sim IG(10^{-3}, 10^{-3}) \\ M_3 &: \; (a_3, b_3, c_3)^T \sim N_3(\mathbf{0}, \mathbf{D}_3\sigma_3^2), \;\; \sigma_3^2 \sim IG(10^{-3}, 10^{-3}), \end{aligned}$$

where $\mathbf{D}_j = n^2\left(\mathbf{X}_j\mathbf{X}_j^T\right)^{-1}$, for $j = 0,1,2,3$. With $\mathbf{X}_j$ we denote the corresponding design matrix of each model while *n* stands for sample size.

A more common choice for $\mathbf{D}_j$ would be $n(\mathbf{X}_j^T\mathbf{X}_j)^{-1}\sigma_j^2$; see Fernandez et al. (2001). Yet, preliminary MCMC runs revealed that the prior corresponding to the latter selection was informative; in the sense that provided posterior distributions away from the corresponding ones when an improper flat prior was adopted.

5.2 Simulating from the posterior

Gibbs sampling and M-H simulation are required in order to implement the methods which were reviewed in Chapter 4. Therefore, we use each method in order to obtain samples from the distributions $p(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j)$, for $j = 0,1,2,3$.

5.2.1 Details of the Gibbs sampler implementation

The full conditional distributions under each model are all of known form. For simplicity we denote $x_{1i} = x_i - \bar{x}$, $x_{2i} = x_i^2$, $x_{3i} = z_i - \bar{z}$. In addition, d_{mn}^j is considered as the $(m \times n)$ element of the matrix $\mathbf{D}_j$. Starting from the simple model we have

Model 0:

$$a_0 \mid \sigma_0^2, \mathbf{y} \sim N\left(w_{a_o} \bar{\mathbf{y}}, w_{a_o} \frac{\sigma_0^2}{n} \right),$$

$$\sigma_0^2 \mid a_0, \mathbf{y} \sim IG\left(A_0, B_0 \right),$$

$$\text{where } w_{a_o} = n / C_{a_0},\ C_{a_0} = n + \left(d_{11}^0 \right)^{-1},$$

$$A_0 = 10^{-3} + \frac{n+1}{2} \text{ and } B_0 = 10^{-3} + 0.5 \left\{ \sum_{i=1}^{n} y_i^2 - 2a_0 \sum_{i=1}^{n} y_i + {a_0}^2 C_{a_0} \right\}.$$

Model 1:

$$a_1 \mid b_1, \sigma_0^2, \mathbf{y} \sim N\left(w_{a_1} \left(n\bar{\mathbf{y}} - b_1 C_1 \right) n^{-1}, w_{a_1} \sigma_1^2 n^{-1} \right),$$

$$b_1 \mid a_1, \sigma_0^2, \mathbf{y} \sim N\left(w_{b_1} \left(\sum_{i=1}^{n} y_i x_{1i} - a_1 C_1 \right) \left(\sum_{i=1}^{n} x_{1i}^2 \right)^{-1}, w_{b_1} \sigma_1^2 \left(\sum_{i=1}^{n} x_{1i}^2 \right)^{-1} \right),$$

$$\sigma_1^2 \mid a_1, b_1, \mathbf{y} \sim IG\left(A_1, B_1 \right),$$

$$\text{where } w_{a_1} = n / C_{a_1},\ w_{b_1} = \sum_{i=1}^{n} x_{1i}^2 \Big/ C_{b_1},\ C_{a_1} = n + \left(d_{11}^1 (1 - \rho_{a_1 b_1}^2) \right)^{-1},$$

$$C_{b_1} = \sum_{i=1}^{n} x_{1i}^2 + \left(d_{22}^1 (1 - \rho_{a_1 b_1}^2) \right)^{-1},\ C_1 = \sum_{i=1}^{n} x_{1i} - \frac{d_{12}^1}{d_{11}^1 d_{22}^1 (1 - \rho_{a_1 b_1}^2)},\ A_1 = 10^{-3} + \frac{n+2}{2} \text{ and}$$

$$B_1 = 10^{-3} + 0.5 \left\{ \sum_{i=1}^{n} y_i^2 + {a_1}^2 C_{a_1} + {b_1}^2 C_{b_1} + 2a_1 b_1 C_1 - 2a_1 \sum_{i=1}^{n} y_i - 2b_1 \sum_{i=1}^{n} y_i x_{1i} \right\}$$

Model 2:

$$a_2 \mid b_2, \sigma_2^2, \mathbf{y} \sim N\left(w_{a_2}\left(n\bar{\mathbf{y}} - b_2 C_2\right)n^{-1}, w_{a_2}\sigma_2^2 n^{-1}\right),$$

$$b_2 \mid a_2, \sigma_2^2, \mathbf{y} \sim N\left(w_{b_2}\left(\sum_{i=1}^{n} y_i x_{3i} - a_2 C_2\right)\left(\sum_{i=1}^{n} x_{3i}^2\right)^{-1}, w_{b_2}\sigma_2^2\left(\sum_{i=1}^{n} x_{3i}^2\right)^{-1}\right),$$

$$\sigma_2^2 \mid a_2, b_2, \mathbf{y} \sim IG\left(A_2, B_2\right),$$

$$\text{where } w_{a_2} = n/C_{a_2},\ w_{b_2} = \sum_{i=1}^{n} x_{3i}^2 \Big/ C_{b_2},$$

$$C_{a_2} = n + \left(d_{11}^2(1-\rho_{a_2b_2}^2)\right)^{-1},\ C_{b_2} = \sum_{i=1}^{n} x_{3i}^2 + \left(d_{22}^2(1-\rho_{a_2b_2}^2)\right)^{-1},\ C_2 = \sum_{i=1}^{n} x_{3i} - \frac{d_{12}^2}{d_{11}^2 d_{22}^2(1-\rho_{a_2b_2}^2)},$$

$$A_2 = 10^{-3} + \frac{n+2}{2} \text{ and}$$

$$B_2 = 10^{-3} + 0.5\left\{\sum_{i=1}^{n} y_i^2 + {a_2}^2 C_{a_2} + {b_2}^2 C_{b_2} + 2a_2 b_2 C_2 - 2a_2\sum_{i=1}^{n} y_i - 2b_2\sum_{i=1}^{n} y_i x_{3i}\right\}.$$

Model 3:

$$a_3 \mid b_3, c_3, \sigma_3^2, \mathbf{y} \sim N\left(w_{a_3}\left(n\bar{\mathbf{y}} - b_3 C_{ab} - c_3 C_{ac}\right)n^{-1}, w_{a_3}\sigma_3^2 n^{-1}\right),$$

$$b_3 \mid a_3, c_3, \sigma_3^2, \mathbf{y} \sim N\left(w_{b_3}\left(\sum_{i=1}^{n} y_i x_{1i} - a_3 C_{ab} - c_3 C_{bc}\right)\left(\sum_{i=1}^{n} x_{1i}^2\right)^{-1}, w_{b_3}\sigma_3^2\left(\sum_{i=1}^{n} x_{1i}^2\right)^{-1}\right),$$

$$c_3 \mid a_3, b_3, \sigma_3^2, \mathbf{y} \sim N\left(w_{c_3}\left(\sum_{i=1}^{n} y_i x_{2i} - a_3 C_{ac} - b_3 C_{bc}\right)\left(\sum_{i=1}^{n} x_{2i}^2\right)^{-1}, w_{c_3}\sigma_3^2\left(\sum_{i=1}^{n} x_{2i}^2\right)^{-1}\right),$$

$$\sigma_3^2 \mid a_3, b_3, c_3, \mathbf{y} \sim IG\left(A_3, B_3\right),$$

$$\text{where } k = 1 - \rho_{a_3b_3}^2 - \rho_{a_3c_3}^2 - \rho_{b_3c_3}^2 + 2\rho_{a_3b_3}\rho_{a_3c_3}\rho_{b_3c_3},$$

$$w_{a3} = n/C_{a_3},\ w_{b_3} = \sum_{i=1}^{n} x_{1i}^2 \Big/ C_{b_3},\ w_{c_3} = \sum_{i=1}^{n} x_{2i}^2 \Big/ C_{c_3},$$

$$C_{a_3} = n + \left(kd_{11}^3/(1-\rho_{b_3c_3}^2)\right)^{-1}, C_{b_3} = \sum_{i=1}^{n} x_{1i}^2 + \left(kd_{22}^3/(1-\rho_{a_3c_3}^2)\right)^{-1},$$

$$C_{c_3} = \sum_{i=1}^{n} x_{2i}^2 + \left(kd_{33}^3/(1-\rho_{a_3b_3}^2)\right)^{-1},$$

$$C_{ab} = \sum_{i=1}^{n} x_{1i} - \frac{\rho_{a_3b_3} - \rho_{a_3c_3}\rho_{c_3b_3}}{k\sqrt{d_{11}^3 d_{22}^3}},\ C_{ac} = \sum_{i=1}^{n} x_{2i} - \frac{\rho_{a_3c_3} - \rho_{a_3b_3}\rho_{b_3c_3}}{k\sqrt{d_{11}^3 d_{33}^3}},$$

$$C_{bc} = \sum_{i=1}^{n} x_{1i}x_{2i} - \frac{\rho_{b_3c_3} - \rho_{a_3b_3}\rho_{a_3c_3}}{k\sqrt{d_{22}^3 d_{33}^3}},$$

$$A_3 = 10^{-3} + \frac{n+3}{2} \text{ and}$$

$$B_3 = 10^{-3} + 0.5\left\{\begin{array}{l} \sum_{i=1}^{n} y_i^2 + a_3{}^2 C_{a_3} + b_3{}^2 C_{b_3} + c_3{}^2 C_{c_3} + 2a_3b_3C_{ab} + 2a_3c_3C_{ac} \\ +2b_3c_3C_{bc} - 2a_3\sum_{i=1}^{n} y_i - 2b_3\sum_{i=1}^{n} y_i x_{1i} - 2c_3\sum_{i=1}^{n} y_i x_{2i} \end{array}\right\}.$$

We utilize five parallel chains of size 11.000 and discard the 1000 first iterations of each chain for the 'burn-in' period. Point estimates, posterior quantiles and the R-root measure for each model are summarized in Tables 5.1 to 5.4 below.

Model 0								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_0	1.608	0.134	0.919	1.520	1.608	1.695	2.250	1.0001
σ_0	0.663	0.098	0.408	0.593	0.652	0.719	1.356	0.9999

Table 5.1 *Model 0 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Gibbs sample of 50000 draws.*

Model 1								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_1	1.607	0.049	1.381	1.574	1.607	1.639	1.833	0.9999
b_1	0.241	0.019	0.152	0.228	0.241	0.254	0.341	1.0000
σ_1	0.244	0.036	0.149	0.218	0.240	0.265	0.496	0.9999

Table 5.2 *Model 1 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Gibbs sample of 50000 draws.*

Model 2								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_2	1.607	0.031	1.432	1.586	1.607	1.627	1.782	0.9999
b_2	1.415	0.070	1.051	1.368	1.415	1.461	1.816	1.0001
σ_2	0.153	0.023	0.092	0.138	0.151	0.167	0.373	0.9999

Table 5.3 *Model 2 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Gibbs sample of 50000 draws.*

Model 3								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_3	1.841	0.043	1.649	1.812	1.841	1.869	2.028	1.0001
b_3	0.255	0.011	0.176	0.247	0.255	0.263	0.337	0.9999
c_3	-0.038	0.005	-0.063	-0.042	-0.038	-0.035	-0.013	1.0000
σ_3	0.139	0.021	0.084	0.124	0.136	0.151	0.287	1.0000

Table 5.4 *Model 3 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Gibbs sample of 50000 draws.*

5.2.2 Details of the Metropolis-Hastings implementation

The joint posterior distribution for each model on logarithmic scale is

$$\log p(\boldsymbol{\theta}_0 \mid \mathbf{y}, M_0) = -2(10^{-3} + \frac{n+3}{2}) \log \sigma_0 - \frac{1}{\sigma_0^2} B_0 + \text{Constant}$$

$$\log p(\boldsymbol{\theta}_1 \mid \mathbf{y}, M_1) = -2(10^{-3} + \frac{n+4}{2}) \log \sigma_1 - \frac{1}{\sigma_1^2} B_1 + \text{Constant}$$

$$\log p(\boldsymbol{\theta}_2 \mid \mathbf{y}, M_2) = -2(10^{-3} + n + 2) \log \sigma_2 - \frac{1}{\sigma_2^2} B_2 + \text{Constant}$$

$$\log p(\boldsymbol{\theta}_3 \mid \mathbf{y}, M_3) = -2(10^{-3} + n + 2.5) \log \sigma_3 - \frac{1}{\sigma_3^2} B_3 + \text{Constant}$$

An independence chain which is allowed to depend on the data $q(\boldsymbol{\theta}_j^{t-1}, \boldsymbol{\theta}_j^t \mid \mathbf{y}) = q(\boldsymbol{\theta}_j^t \mid \mathbf{y})$ for $j = 0,1,2,3$ is utilized for each M-H simulation. The proposal distribution under each model is

$$q(\boldsymbol{\theta}_0 \mid \mathbf{y}) = q(a_0, \log\sigma_0 \mid \mathbf{y}) \sim N_2\left(\left(a_0^{(M)}, \log\sigma_0^{(M)}\right)^T, \mathbf{V}_0\right)$$

$$q(\boldsymbol{\theta}_1 \mid \mathbf{y}) = q\left(\left(a_1, b_1, \log\sigma_1\right)^T \mid \mathbf{y}\right) \sim N_3\left(\left(a_1^{(M)}, b_1^{(M)}, \log\sigma_1^{(M)}\right)^T, \mathbf{V}_1\right)$$

$$q(\boldsymbol{\theta}_2 \mid \mathbf{y}) = q\left(\left(a_2, b_2, \log\sigma_2\right)^T \mid \mathbf{y}\right) \sim N_3\left(\left(a_2^{(M)}, b_2^{(M)}, \log\sigma_2^{(M)}\right)^T, \mathbf{V}_2\right)$$

$$q(\boldsymbol{\theta}_3 \mid \mathbf{y}) = q\left(\left(a_3, b_3, c_3, \log\sigma_3\right)^T \mid \mathbf{y}\right) \sim N_4\left(\left(a_3^{(M)}, b_3^{(M)}, c_3^{(M)}, \log\sigma_3^{(M)}\right)^T, \mathbf{V}_3\right).$$

The upper script (M) stands for the mode of the log-target density while $\mathbf{V}_j$ is the inverse of the negative Hessian of the log- target evaluated at the mode. The covariance matrices $\mathbf{V}_j$ are

$$\mathbf{V}_0 = \sigma_0^{-2(M)} \begin{pmatrix} C_{a_0} & 2\left(\sum_{i=1}^n y_i - a_0^{(M)} C_{a_0}\right) \\ \bullet & 4B_0 \end{pmatrix},$$

$$\mathbf{V}_1 = \sigma_1^{-2(M)} \begin{pmatrix} C_{a_1} & C_1 & 2\left(\sum_{i=1}^n y_i - a_1^{(M)} C_{a_1} - b_1^{(M)} C_{b_1}\right) \\ \bullet & C_{b_1} & 2\left(\sum_{i=1}^n y_i x_{1i} - a_1^{(M)} C_{a_1} - b_1^{(M)} C_{b_1}\right) \\ \bullet & \bullet & 4B_1 \end{pmatrix},$$

$$\mathbf{V}_2 = \sigma_2^{-2(M)} \begin{pmatrix} C_{a_2} & C_2 & 2\left(\sum_{i=1}^n y_i - a_2^{(M)} C_{a_2} - b_2^{(M)} C_{b_2}\right) \\ \bullet & C_{b_2} & 2\left(\sum_{i=1}^n y_i x_{3i} - a_2^{(M)} C_{a_2} - b_2^{(M)} C_{b_2}\right) \\ \bullet & \bullet & 4B_2 \end{pmatrix},$$

$$\mathbf{V}_3 = \sigma_3^{-2(M)} \begin{pmatrix} C_{a_3} & C_{ab} & C_{ac} & 2\left(\sum_{i=1}^{n} y_i - a_3^{(M)} C_{a_3} - b_3^{(M)} C_{ab} - c_3^{(M)} C_{ac}\right) \\ \bullet & C_{b_3} & C_{bc} & 2\left(\sum_{i=1}^{n} y_i x_{1i} - b_3^{(M)} C_{b_3} - a_3^{(M)} C_{ab} - c_3^{(M)} C_{ac}\right) \\ \bullet & \bullet & C_{c_3} & 2\left(\sum_{i=1}^{n} y_i x_{2i} - c_3^{(M)} C_{c_3} - a_3^{(M)} C_{ac} - b_3^{(M)} C_{bc}\right) \\ \bullet & \bullet & \bullet & 4B_3 \end{pmatrix}.$$

The posterior means from the Gibbs sampling are used as estimates of the mode. Although, the posterior mean is a rough estimate of the mode, preliminary M-H simulations revealed that the use of the posterior median or of the ML estimate instead results to lower acceptance ratios. The M-H probability of transition for each model is

$$\log a_{MH}(\boldsymbol{\theta}_j^{(t-1)}, \boldsymbol{\theta}_j^{(t)}) = \log p\left(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}, M_j\right) + \log q\left(\boldsymbol{\theta}_j^{(t-1)} \mid \mathbf{y}\right) - \log \sigma_j^{(t-1)} - \\ - \log p\left(\boldsymbol{\theta}_j^{(t-1)} \mid \mathbf{y}, M_j\right) - \log q\left(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}\right) + \log \sigma_j^{(t)}.$$

Five parallel chains of size 11.000 are used again from starting points that are located within the posterior space of each model. The first 1000 iterations of each chain are discarded. Acceptance ratios average to approximately 82% for model 0, 80% for models 1 and 2 and 56% for model 3. Parameters point estimates, posterior quantiles and the R-root measure are presented in Tables 5.5 to 5.8 below.

Model 0								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_0	1.608	0.130	1.116	1.522	1.607	1.694	2.188	0.9999
σ_0	0.648	0.094	0.395	0.522	0.638	0.703	1.218	0.9999

Table 5.5 *Model 0 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Metropolis-Hastings sample of 50000 draws.*

Model 1								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_1	1.607	0.048	1.423	1.575	1.607	1.638	1.801	0.9999
b_1	0.241	0.019	0.156	0.228	0.241	0.254	0.318	1.0001
σ_1	0.239	0.035	0.137	0.215	0.235	0.260	0.434	0.9999

Table 5.6 *Model 1 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Metropolis-Hastings sample of 50000 draws.*

Model 2								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_2	1.607	0.030	1.483	1.587	1.607	1.627	1.731	1.0000
b_2	1.414	0.069	1.111	1.369	1.414	1.460	1.696	1.0001
σ_2	0.151	0.022	0.091	0.135	0.148	0.163	0.282	1.0001

Table 5.7 *Model 2 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Metropolis-Hastings sample of 50000 draws.*

Model 3								
			Posterior Quantiles					
Parameter	**Mean**	**St.dev**	**0%**	**25%**	**Median**	**75%**	**100%**	**R-root**
a_3	1.840	0.043	1.649	1.812	1.840	1.868	2.039	1.0001
b_3	0.255	0.011	0.205	0.248	0.255	0.262	0.301	1.0001
c_3	-0.038	0.005	-0.060	-0.042	-0.038	-0.034	-0.016	0.9999
σ_3	0.135	0.019	0.084	0.121	0.133	0.146	0.268	1.0001

Table 5.8 *Model 3 parameters posterior estimates of mean, standard deviation, quantiles and the R root reduction measure resulting from a Metropolis-Hastings sample of 50000 draws.*

5.3 Implementation of the methods

In this section we demonstrate the implementation of the marginal likelihood estimation methods presented in Chapter 4. The G draws from the Gibbs sampler are denoted as $\left\{\boldsymbol{\theta}^{(g)}\right\}$, with $g = 1,...,G$, while the M Metropolis-Hastings draws are denoted as $\left\{\boldsymbol{\theta}^{(m)}\right\}$, with $m = 1,...,M$.

5.3.1 Harmonic Mean Estimator

The Harmonic Mean estimator is given by $\widehat{p}_H = \left\{ M^{-1} \sum_{m=1}^{M} p(\mathbf{y} \mid \boldsymbol{\theta}^{(m)})^{-1} \right\}^{-1}$. This method is relatively easy to implement. We simply calculate the likelihood of the data under model M_j for all Gibbs iterations

$$p(\mathbf{y} \mid \boldsymbol{\theta}_0^{(g)}, M_0) = \left(\sqrt{2\pi\sigma_0^{2(g)}}\right)^{-n} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - a_0^{(g)}}{\sigma_0^{(g)}} \right)^2 \right\}$$

$$p(\mathbf{y} \mid \boldsymbol{\theta}_1^{(g)}, M_1) = \left(\sqrt{2\pi\sigma_1^{2(g)}}\right)^{-n} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - a_1^{(g)} - b_1^{(g)} x_{1i}}{\sigma_1^{(g)}} \right)^2 \right\}$$

$$p(\mathbf{y} \mid \boldsymbol{\theta}_2^{(g)}, M_2) = \left(\sqrt{2\pi\sigma_2^{(g)}}\right)^{-n} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - a_2^{(g)} - b_2^{(g)} x_{3i}}{\sigma_2^{(g)}} \right)^2 \right\}$$

$$p(\mathbf{y} \mid \boldsymbol{\theta}_3^{(g)}, M_3) = \left(\sqrt{2\pi\sigma_3^{2(g)}}\right)^{-n} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - a_3^{(g)} - b_3^{(g)} x_{1i} - c_3^{(g)} x_{2i}}{\sigma_3^{(g)}} \right)^2 \right\},$$

where $i = 1,...,n$ and $g = 1,...,G$.

From the likelihood values we obtain the marginal likelihood estimates by summing over the Gibbs draws under each separate model

$$\widehat{p}(\mathbf{y} \mid M_j)_H = \frac{G}{\sum_{g=1}^{G} p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(g)}, M_j)^{-1}}.$$

5.3.2 Laplace-Metropolis Estimator

As discussed in section 4.3 the Laplace-Metropolis estimator is given by

$$\widehat{p}_{LM} = (2\pi)^{d/2} \left|\mathbf{S}\right|^{1/2} p(\mathbf{y} \mid \boldsymbol{\theta}^{\max}) p(\boldsymbol{\theta}^{\max}),$$

where $\boldsymbol{\theta}^{\max}$ is the point that maximizes $p(\mathbf{y} \mid \boldsymbol{\theta}^{(g)}) p(\boldsymbol{\theta}^{(g)})$ among the G posterior draws and $\mathbf{S}$ is the sample covariance matrix of the output. The prior densities $p(\boldsymbol{\theta}_j \mid M_j)$ have the form

$$p(\boldsymbol{\theta}_0 \mid M_0) = \left(2\pi \left|\mathbf{P}_0\right|\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(a_0{}^T \mathbf{P}_0{}^{-1} a_0\right)\right\} \times$$
$$\times \left(10^{-3}\right)^{10^{-3}} \left(\sigma_0^2\right)^{-(10^{-3}+1)} \exp\left(-\frac{10^{-3}}{\sigma_0^2}\right) \Big/ \Gamma(10^{-3})$$
$$p(\boldsymbol{\theta}_1 \mid M_1) = 2\pi^{-1} \left|\mathbf{P}_1\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\left(a_1 \; b_1\right)^T \mathbf{P}_1^{-1} \left(a_1 \; b_1\right)\right)\right\} \times$$
$$\times \left(10^{-3}\right)^{10^{-3}} \left(\sigma_1^2\right)^{-(10^{-3}+1)} \exp\left(-\frac{10^{-3}}{\sigma_1^2}\right) \Big/ \Gamma(10^{-3})$$
$$p(\boldsymbol{\theta}_2 \mid M_2) = 2\pi^{-1} \left|\mathbf{P}_2\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\left(a_2 \; b_2\right)^T \mathbf{P}_2^{-1} \left(a_2 \; b_2\right)\right)\right\} \times$$
$$\times \left(10^{-3}\right)^{10^{-3}} \left(\sigma_2^2\right)^{-(10^{-3}+1)} \exp\left(-\frac{10^{-3}}{\sigma_2^2}\right) \Big/ \Gamma(10^{-3})$$
$$p(\boldsymbol{\theta}_3 \mid M_3) = 2\pi^{-3/2} \left|\mathbf{P}_2\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\left(a_3 \; b_3 \; c_3\right)^T \mathbf{P}_3^{-1} \left(a_3 \; b_3 \; c_3\right)\right)\right\} \times$$
$$\times \left(10^{-3}\right)^{10^{-3}} \left(\sigma_3^2\right)^{-(10^{-3}+1)} \exp\left(-\frac{10^{-3}}{\sigma_3^2}\right) \Big/ \Gamma(10^{-3}),$$

where $\mathbf{P}_j = \mathbf{D}_j \sigma_j^2$.

We calculate $p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(g)}, M_j) p(\boldsymbol{\theta}_j^{(g)} \mid M_j)$ for $g = 1, \ldots, G$ and locate the point $\boldsymbol{\theta}_j^{\max}$ which maximizes this quantity. The marginal likelihoods are then obtained from

$$\widehat{p}(\mathbf{y} \mid M_j)_{LM} = (2\pi)^{d_j/2} \left|\mathbf{S}_j\right|^{1/2} p(\mathbf{y} \mid \boldsymbol{\theta}_j^{\max}, M_j) p(\boldsymbol{\theta}_j^{\max} \mid M_j),$$

where d_j is the dimension of the parameter vector $\boldsymbol{\theta}_j$.

We also evaluate the marginal likelihoods for alternative choices of $\boldsymbol{\theta}_j^{\max}$ such as the posterior medians $\boldsymbol{\theta}_j^{Med}$ and the posterior means $\bar{\boldsymbol{\theta}}_j$. These choices do not require further calculations, we simply compute the quantities

$$\widehat{p}(\mathbf{y} \mid M_j)_{LM_{MEDIAN}} = (2\pi)^{d_j/2} \left|\mathbf{S}_j\right|^{1/2} p(\mathbf{y} \mid \boldsymbol{\theta}_j^{Med}, M_j) p(\boldsymbol{\theta}_j^{Med} \mid M_j)$$

and

$$\widehat{p}(\mathbf{y} \mid M_j)_{LM_{MEAN}} = (2\pi)^{d_j/2} \left|\mathbf{S}_j\right|^{1/2} p(\mathbf{y} \mid \bar{\boldsymbol{\theta}}_j, M_j) p(\bar{\boldsymbol{\theta}}_j \mid M_j).$$

5.3.3 Newton and Raftery Estimator

We implement the first method mentioned in section 4.4. The marginal likelihood estimates are given by

$$\widehat{p}(\mathbf{y} \mid M_j)_{NR} = \frac{\sum_{l=1}^{G} p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(l)}, M_j) \left\{ \delta \widehat{p}(\mathbf{y} \mid M_j)_{NR} + (1-\delta) p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(l)}, M_j) \right\}^{-1}}{\sum_{l=1}^{G} \left\{ \delta \widehat{p}(\mathbf{y} \mid M_j)_{NR} + (1-\delta) p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(l)}, M_j) \right\}^{-1}},$$

with $\delta \in (0,1)$. The $\boldsymbol{\theta}_j^{(l)}$ are draws from the mixture distribution

$$f(\boldsymbol{\theta}_j \mid M_j) = \delta p(\boldsymbol{\theta}_j \mid M_j) + (1-\delta) p(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j),$$

for $l = 1, ..., G$ and $j = 0, 1, 2, 3$. In order to acquire the draws $\boldsymbol{\theta}_j^{(l)}$ we first generate $\delta \times G$ independent draws from the prior distributions $p(\boldsymbol{\theta}_j \mid M_j)$. Then, we randomly replace $\delta \times G$ values of the posterior Gibbs samples with these draws. A problem that arises is that some draws from the distributions $p(\sigma_j^2 \mid M_j) \sim IG(10^{-3}, 10^{-3})$ take an infinite value. Therefore, we substitute these values with the maximum value observed among the draws $\sigma_j^{2(k)} < \infty$, with $k < l$.

We use as initial values the marginal likelihood estimates of the Harmonic Mean method and the Laplace-Metropolis method evaluated at the posterior mean and compare the resulting estimates after 1000, 5000 and 10000 iterations for a choice of δ equal to 0.05.

5.3.4 Bridge Sampling Estimators

In order to implement the bridge sampling techniques we must initially generate draws $\left\{\boldsymbol{\theta}_j^{Apr}\right\}$ from a density $g(\boldsymbol{\theta}_j^{Apr} \mid M_j)$, which must be an accurate approximation of the corresponding posterior distribution $p(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j)$. This is achieved by appropriately fitting posterior moments.

We generate the parameters $\sigma_j^{2\,Apr}$ from an Inverse Gamma distribution, that is $\sigma_j^{2\,Apr} \sim IG(e_j, f_j)$. The mean and variance of this distribution are

$$E(\sigma_j^{2\,Apr}) = \frac{f_j}{e_j - 1} \quad \text{and} \quad Var(\sigma_j^{2\,Apr}) = \frac{f_j^{\,2}}{(e_j - 1)^2 (e_j + 1)}.$$

Equating these to the estimates $\overline{\sigma_j^2} = \frac{1}{G}\sum_{g=1}^{G} \sigma_j^{2(g)}$ and $S_{\sigma_j^2}^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\sigma_j^{2(g)} - \overline{\sigma_j^2}\right)^2$ obtained by the posterior Gibbs samples and solving with respect to e_j and f_j we have that

$$e_j = \frac{\left(\overline{\sigma_j^2}\right)^2}{S_{\sigma_j^2}^2} + 2 \quad \text{and} \quad f_j = \overline{\sigma_j^2}(e_j - 1).$$

The rest of the parameters are sampled similarly from normal distributions:

- $a_0^{\,Apr} \sim N\left(\overline{a_0}, S_{a_0}^2\right)$, where $\overline{a_0} = \frac{1}{G}\sum_{g=1}^{G} a_0^{(g)}$ and $S_{a_0}^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(a_0^{(g)} - \overline{a_0}\right)^2$
- $\left(a_1^{\,Apr}\ b_1^{\,Apr}\right)^T \sim N_2\left(\left(\overline{a_1}\ \overline{b_1}\right)^T, \mathbf{S}_1\right)$, where $\overline{a_1} = \frac{1}{G}\sum_{g=1}^{G} a_1^{(g)}$, $\overline{b_1} = \frac{1}{G}\sum_{g=1}^{G} b_1^{(g)}$ and

$$\mathbf{S}_1 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\left(a_1^{(g)}\ b_1^{(g)}\right) - \left(\overline{a_1}\ \overline{b_1}\right)\right)\left(\left(a_1^{(g)}\ b_1^{(g)}\right) - \left(\overline{a_1}\ \overline{b_1}\right)\right)^T$$

- $\left(a_2^{Apr}\ b_2^{Apr}\right)^T \sim N_2\left(\left(\overline{a_2}\ \overline{b_2}\right)^T, \mathbf{S}_2\right)$, where $\overline{a_2} = \frac{1}{G}\sum_{g=1}^{G} a_2^{(g)}$, $\overline{b_2} = \frac{1}{G}\sum_{g=1}^{G} b_2^{(g)}$ and

  $$\mathbf{S}_2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\left(a_2^{(g)}\ b_2^{(g)}\right)-\left(\overline{a_2}\ \overline{b_2}\right)\right)\left(\left(a_2^{(g)}\ b_2^{(g)}\right)-\left(\overline{a_2}\ \overline{b_2}\right)\right)^T$$

- $\left(a_3^{Apr}\ b_3^{Apr}\ c_3^{Apr}\right)^T \sim N_3\left(\left(\overline{a_3}\ \overline{b_3}\ \overline{c_3}\right)^T, \mathbf{S}_3\right)$, where $\overline{a_3} = \frac{1}{G}\sum_{g=1}^{G} a_3^{(g)}$, $\overline{b_3} = \frac{1}{G}\sum_{g=1}^{G} b_3^{(g)}$, $\overline{c_3} = \frac{1}{G}\sum_{g=1}^{G} c_3^{(g)}$ and

  $$\mathbf{S}_3 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\left(a_3^{(g)}\ b_3^{(g)}\ \mathrm{c}_3^{(g)}\right)-\left(\overline{a_3}\ \overline{b_3}\ \overline{c_3}\right)\right)\left(\left(a_3^{(g)}\ b_3^{(g)}\ \mathrm{c}_3^{(g)}\right)-\left(\overline{a_3}\ \overline{b_3}\ \overline{c_3}\right)\right)^T.$$

In total we sample $L = 50000$ draws of $\left\{\boldsymbol{\theta}_j^{Apr}\right\}$, for $j = 0,1,2,3$. Thus, the densities $g(\cdot \mid M_j)$ are

$$g(\cdot \mid M_0) = N(\overline{a_0}, S_{a_0}^2) \times IG(e_0, f_0)$$
$$g(\cdot \mid M_1) = N_2\left(\left(\overline{a_1}\ \overline{b_1}\right)^T, \mathbf{S}_1\right) \times IG(e_1, f_1)$$
$$g(\cdot \mid M_2) = N_2\left(\left(\overline{a_2}\ \overline{b_2}\right)^T, \mathbf{S}_2\right) \times IG(e_2, f_2)$$
$$g(\cdot \mid M_3) = N_3\left(\left(\overline{a_3}\ \overline{b_3}\ \overline{c_3}\right)^T, \mathbf{S}_3\right) \times IG(e_3, f_3).$$

We then calculate the marginal likelihoods for the three types of bridge estimators reviewed in section 4.5 from

$$\widehat{p}(\mathbf{y} \mid M_j)_{HB} = \frac{\sum_{l=1}^{L} g(\boldsymbol{\theta}_j^{Apr(l)} \mid M_j)^{-1}}{\sum_{g=1}^{G}\left\{p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(g)}, M_j)\, p(\boldsymbol{\theta}_j^{(g)} \mid M_j)\right\}^{-1}},$$

$$\widehat{p}(\mathbf{y}\,|\,M_j)_{GB}=\frac{\sum_{l=1}^{L}\left\{p(\mathbf{y}\,|\,\boldsymbol{\theta}_j^{Apr(l)},M_j)\,p(\boldsymbol{\theta}_j^{Apr(l)}\,|\,M_j)\big/g(\boldsymbol{\theta}_j^{Apr(l)}\,|\,M_j)\right\}^{1/2}}{\sum_{g=1}^{G}\left\{g(\boldsymbol{\theta}_j^{(g)}\,|\,M_j)\big/p(\mathbf{y}\,|\,\boldsymbol{\theta}_j^{(g)},M_j)\,p(\boldsymbol{\theta}_j^{(g)}\,|\,M_j)\right\}^{1/2}},$$

and

$$\widehat{p}(\mathbf{y}\,|\,M_j)_{OptB}=\frac{\sum_{l=1}^{L}W_{2j}^{(l)}\Big/\left(s_1W_{2j}^{(l)}+s_2\,\widehat{p}(\mathbf{y}\,|\,M_j)_{OptB}\right)}{\sum_{g=1}^{G}1\Big/\left(s_1W_{1j}^{(g)}+s_2\,\widehat{p}(\mathbf{y}\,|\,M_j)_{OptB}\right)}.$$

For the optimum estimator we have that

$$s_1=s_2=0.5,$$

$$W_{2j}^{(l)}=p(\mathbf{y}\,|\,\boldsymbol{\theta}_j^{Apr(l)},M_j)\,p(\boldsymbol{\theta}_j^{Apr(l)}\,|\,M_j)\big/g(\boldsymbol{\theta}_j^{Apr(l)}\,|\,M_j)\ \text{ and}$$

$$W_{1j}^{(g)}=p(\mathbf{y}\,|\,\boldsymbol{\theta}_j^{(g)},M_j)\,p(\boldsymbol{\theta}_j^{(g)}\,|\,M_j)\big/g(\boldsymbol{\theta}_j^{(g)}\,|\,M_j).$$

We set as initial values the estimates $\widehat{p}(\mathbf{y}\,|\,M_j)_{GB}$ and iterate the equation 1000 times.

5.3.5 Candidate's estimators

As discussed in section 4.5 the Candidate's estimators are based on the identity

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

which holds for any point $\boldsymbol{\theta}^*$. The goal is to estimate the posterior ordinate $p(\boldsymbol{\theta}^*\,|\,\mathbf{y})$ at point $\boldsymbol{\theta}^*$ which is usually taken to be a high density point with respect to the posterior density. We select $\boldsymbol{\theta}_j^*$, for $j=0,1,2,3$ as the point that maximizes the corresponding log target density of each model M_j.

5.3.5.1 The Chib Estimator

The Gibbs sample is utilized again in order to implement the method of Chib. The calculation of the posterior ordinate depends on the dimension of $\boldsymbol{\theta}$, so we will demonstrate the use of this method for each model separately.

Starting from model 0 we have that $p(\boldsymbol{\theta}_0^* \mid \mathbf{y}) = p(a_0^*, \sigma_0^{2*} \mid \mathbf{y}) = p(\sigma_0^{2*} \mid a_0^*, \mathbf{y}) p(a_0^* \mid \mathbf{y})$. The quantity $p(\sigma_0^{2*} \mid a_0^*, \mathbf{y})$ can be calculated directly since the full conditional densities are of known form as we have seen earlier in this chapter. In addition, we have that

$$p(a_0^* \mid \mathbf{y}) = \int p(a_0^* \mid \sigma_0^2, \mathbf{y}) p(\sigma_0^2 \mid \mathbf{y}) d\sigma_0^2,$$

so we obtain the estimate

$$\widehat{p}(a_0^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(a_0^* \mid \sigma_0^{2(g)}, \mathbf{y}).$$

Thus, we can estimate the marginal likelihood of model 0 on logarithmic scale from

$$\log \widehat{p}(\mathbf{y} \mid M_0)_{Chib} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_0^*) + \log p(\boldsymbol{\theta}_0^*) - \log p(\sigma_0^{2*} \mid a_0^*, \mathbf{y}) - \log\left(G^{-1} \sum_{g=1}^{G} p(a_0^* \mid \sigma_0^{2(g)}, \mathbf{y}) \right)$$

Notice that a reduced Gibbs run was not necessary in order to estimate the marginal likelihood of model 0.

For model 1 the posterior ordinate can be decomposed similarly

$$p(\boldsymbol{\theta}_1^* \mid \mathbf{y}) = p(a_1^*, b_1^*, \sigma_1^{2*} \mid \mathbf{y}) = p(\sigma_1^{2*} \mid a_1^*, b_1^*, \mathbf{y}) p(a_1^* \mid b_1^*, \mathbf{y}) p(b_1^* \mid \mathbf{y}).$$

The ordinate $p(\sigma_1^{2*} \mid a_1^*, b_1^*, \mathbf{y})$ can be calculated directly. Thus, the reduced ordinates $p(a_1^* \mid b_1^*, \mathbf{y})$ and $p(b_1^* \mid \mathbf{y})$ remain to be estimated; we have that

$$p(b_1^* \mid \mathbf{y}) = \int p(b_1^* \mid a_1, \sigma_1^2, \mathbf{y}) p(a_1, \sigma_1^2 \mid \mathbf{y}) da_1 d\sigma_1^2,$$

which yields the estimator

$$\widehat{p}(b_1^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(b_1^* \mid a_1^{(g)}, \sigma_1^{2(g)}, \mathbf{y}).$$

In addition, we have that

$$p(a_1^* \mid b_1^*, \mathbf{y}) = \int p(a_1^* \mid b_1^*, \sigma_1^2, \mathbf{y}) p(\sigma_1^2 \mid b_1^*, \mathbf{y}) d\sigma_1^2 .$$

In order to estimate this ordinate we continue sampling from the distributions $p(\sigma_1^2 \mid a_1, b_1^*, \mathbf{y})$ and $p(a_1 \mid b_1^*, \sigma_1^2, \mathbf{y})$ keeping b_1 fixed at point b_1^*. We iterate this Gibbs sampler 51.000 times and discard the first 1000 iterations acquiring in total a sample L=50.000. The draws $\left\{a_1^{(l)}, \sigma_1^{2(l)}\right\}$ for $l = 1,...,L$, follow the distribution of $a_1, \sigma_1^2 \mid b_1^*, \mathbf{y}$. So, we acquire the estimate

$$\hat{p}(a_1^* \mid b_1^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(a_1^* \mid b_1^*, \sigma_1^{2(l)}, \mathbf{y}) .$$

The marginal likelihood on logarithmic scale is then estimated from

$$\log \hat{p}(\mathbf{y} \mid M_1)_{Chib} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_1^*) + \log p(\boldsymbol{\theta}_1^*) - \log p(\sigma_1^{2*} \mid a_1^*, b_1^*, \mathbf{y}) - \log\left(L^{-1} \sum_{l=1}^{L} p(a_1^* \mid b_1^*, \sigma_1^{2(l)}, \mathbf{y}) \right)$$
$$- \log\left(G^{-1} \sum_{g=1}^{G} p(b_1^* \mid a_1^{(g)}, \sigma_1^{2(g)}, \mathbf{y}) \right).$$

The estimation of $p(\mathbf{y} \mid M_2)$ for model 2 is identical since the two models have equal number of parameters. The reduced Gibbs run is applied on the distributions $p(\sigma_2^2 \mid a_2, b_2^*, \mathbf{y})$ and $p(a_2 \mid b_2^*, \sigma_2^2, \mathbf{y})$, resulting to L draws of $\left\{a_2^{(l)}, \sigma_2^{2(l)}\right\}$ from $a_2, \sigma_2^2 \mid b_2^*, \mathbf{y}$. Thus, we have

$$\log \hat{p}(\mathbf{y} \mid M_2)_{Chib} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_2^*) + \log p(\boldsymbol{\theta}_2^*) - \log p(\sigma_2^{2*} \mid a_2^*, b_2^*, \mathbf{y}) - \log\left(L^{-1} \sum_{l=1}^{L} p(a_2^* \mid b_2^*, \sigma_2^{2(l)}, \mathbf{y}) \right)$$
$$- \log\left(G^{-1} \sum_{g=1}^{G} p(b_2^* \mid a_2^{(g)}, \sigma_2^{2(g)}, \mathbf{y}) \right).$$

Estimating the marginal likelihood of model 3 is more complicated. The posterior ordinate is decomposed as shown below

$$p(\boldsymbol{\theta}_3^* \mid \mathbf{y}) = p(a_3^*, b_3^*, c_3^*, \sigma_3^{2*} \mid \mathbf{y}) = p(\sigma_3^{2*} \mid a_3^*, b_3^*, c_3^*, \mathbf{y}) p(a_3^* \mid b_3^*, c_3^*, \mathbf{y}) p(b_3^* \mid c_3^*, \mathbf{y}) p(c_3^* \mid \mathbf{y}).$$

The first reduced ordinate can be again calculated directly. The last reduced ordinate can be expressed as $p(c_3^* \mid \mathbf{y}) = \int p(c_3^* \mid a_3, b_3, \sigma_3^2, \mathbf{y}) p(a_3, b_3, \sigma_3^2 \mid \mathbf{y}) da_3 db_3 d\sigma_3^2$ and therefore, estimated from

$$\widehat{p}(c_3^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(c_3^* \mid a_3^{(g)}, b_3^{(g)}, \sigma_3^{2(g)}, \mathbf{y}).$$

In addition, we have that $p(b_3^* \mid c_3^*, \mathbf{y}) = \int p(b_3^* \mid a_3, c_3^*, \sigma_3^2, \mathbf{y}) p(a_3, \sigma_3^2 \mid c_3^*, \mathbf{y}) da_3 d\sigma_3^2$. So, draws from the distribution $a_3, \sigma_3^2 \mid c_3^*, \mathbf{y}$ are required in order to estimate this ordinate. Therefore, we sample from the distributions

$$p(a_3 \mid b_3, c_3^*, \sigma_3^2, \mathbf{y}),$$
$$p(b_3 \mid a_3, c_3^*, \sigma_3^2, \mathbf{y}),$$
$$p(\sigma_3^2 \mid a_3, b_3, c_3^*, \mathbf{y}),$$

with c_3 remaining constant at c_3^*. The draws $\left\{a_3^{(l)}, b_3^{(l)}, \sigma_3^{2(l)}\right\}$ for $l = 1, ..., L$, follow the distribution of $a_3, b_{3,} \sigma_3^2 \mid c_3^*, \mathbf{y}$. Thus, we obtain the estimate

$$\widehat{p}(b_3^* \mid c_3^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(b_3^* \mid a_3^{(l)}, c_3^*, \sigma_3^{2(l)}, \mathbf{y}),$$

again with L equal to 50.000.

Finally, we have that $p(a_3^* \mid b_3^*, c_3^*, \mathbf{y}) = \int p(a_3^* \mid b_3^*, c_3^*, \sigma_3^2, \mathbf{y}) p(\sigma_3^2 \mid b_3^*, c_3^*, \mathbf{y}) d\sigma_3^2$. The estimation of this integral requires draws from the distribution $\sigma_3^2 \mid b_3^*, c_3^*, \mathbf{y}$. Therefore, we run a second reduced Gibbs run with the distributions $p(a_3 \mid b_3^*, c_3^*, \sigma_3^2, \mathbf{y})$ and $p(\sigma_3^2 \mid b_3^*, c_3^*, a_3, \mathbf{y})$ keeping b_3 and c_3 fixed at points b_3^*, c_3^*. This reduced run provides draws of $\left\{a_3^{(m)}, \sigma_3^{2(m)}\right\}$, for $m = 1, ..., M$, from the distribution $a_3, \sigma_3^2 \mid b_3^*, c_3^*, \mathbf{y}$ and yields the estimator

$$\widehat{p}(a_3^* \mid b_3^*, c_3^*, \mathbf{y}) = M^{-1} \sum_{m=1}^{M} p(a_3^* \mid b_3^*, c_3^*, \sigma_3^{2(m)}, \mathbf{y}),$$

with $M = 50.000$. Thus, we can estimate the marginal likelihood of model 3 from

$$\log \widehat{p}(\mathbf{y} \mid M_3)_{Chib} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_3^*) + \log p(\boldsymbol{\theta}_3^*) - \log p(\sigma_3^{2*} \mid a_3^*, b_3^*, c_3^*, \mathbf{y}) - \log\left(M^{-1} \sum_{m=1}^{M} p(a_3^* \mid b_3^*, c_3^*, \sigma_3^{2(m)}, \mathbf{y}) \right)$$
$$- \log\left(L^{-1} \sum_{l=1}^{L} p(b_3^* \mid a_3^{(l)}, c_3^*, \sigma_3^{2(l)}, \mathbf{y}) \right) - \log\left(G^{-1} \sum_{g=1}^{G} p(c_3^* \mid a_3^{(g)}, b_3^{(g)}, \sigma_3^{2(g)}, \mathbf{y}) \right).$$

In brief, the calculation of the posterior ordinate for each model required the following steps:

Model 0:

- Decompose the posterior ordinate as $p(a_0^*, \sigma_0^{2*} \mid \mathbf{y}) = p(\sigma_0^{2*} \mid a_0^*, \mathbf{y}) p(a_0^* \mid \mathbf{y})$
- Compute $p(\sigma_0^{2*} \mid a_0^*, \mathbf{y})$ (known)
- Estimate $p(a_0^* \mid \mathbf{y})$ from $\widehat{p}(a_0^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(a_0^* \mid \sigma_0^{2(g)}, \mathbf{y})$

Model 1:

- Decompose the posterior ordinate as $p(a_1^*, b_1^*, \sigma_1^{2*} \mid \mathbf{y}) = p(\sigma_1^{2*} \mid a_1^*, b_1^*, \mathbf{y}) p(a_1^* \mid b_1^*, \mathbf{y}) p(b_1^* \mid \mathbf{y})$
- Compute $p(\sigma_1^{2*} \mid a_1^*, b_1^*, \mathbf{y})$ (known)
- Estimate $p(b_1^* \mid \mathbf{y})$ from $\widehat{p}(b_1^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(b_1^* \mid a_1^{(g)}, \sigma_1^{2(g)}, \mathbf{y})$
- Acquire a sample L from $a_1, \sigma_1^2 \mid b_1^*, \mathbf{y}$ through Gibbs sampling with the distributions $p(\sigma_1^2 \mid a_1, b_1^*, \mathbf{y})$ and $p(a_1 \mid b_1^*, \sigma_1^2, \mathbf{y})$
- Estimate $p(a_1^* \mid b_1^*, \mathbf{y})$ from $\widehat{p}(a_1^* \mid b_1^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(a_1^* \mid b_1^*, \sigma_1^{2(l)}, \mathbf{y})$

Model 2:

- Decompose the posterior ordinate as $p(a_2^*, b_2^*, \sigma_2^{2*} \mid \mathbf{y}) = p(\sigma_2^{2*} \mid a_2^*, b_2^*, \mathbf{y}) p(a_2^* \mid b_2^*, \mathbf{y}) p(b_2^* \mid \mathbf{y})$
- Compute $p(\sigma_2^{2*} \mid a_2^*, b_2^*, \mathbf{y})$ (known)
- Estimate $p(b_2^* \mid \mathbf{y})$ from $\widehat{p}(b_2^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(b_2^* \mid a_2^{(g)}, \sigma_2^{2(g)}, \mathbf{y})$

- Acquire a sample L from $a_2, \sigma_2^2 \mid b_2^*, \mathbf{y}$ through Gibbs sampling with the distributions $p(\sigma_2^2 \mid a_2, b_2^*, \mathbf{y})$ and $p(a_2 \mid b_2^*, \sigma_2^2, \mathbf{y})$
- Estimate $p(a_2^* \mid b_2^*, \mathbf{y})$ from $\hat{p}(a_2^* \mid b_2^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(a_2^* \mid b_2^*, \sigma_2^{2(l)}, \mathbf{y})$

Model 3:

- Decompose the posterior ordinate as $p(a_3^*, b_3^*, c_3^*, \sigma_3^{2*} \mid \mathbf{y}) = p(\sigma_3^{2*} \mid a_3^*, b_3^*, c_3^*, \mathbf{y}) p(a_3^* \mid b_3^*, c_3^*, \mathbf{y}) p(b_3^* \mid c_3^*, \mathbf{y}) p(c_3^* \mid \mathbf{y})$
- Compute $p(\sigma_3^{2*} \mid a_3^*, b_3^*, c_3^*, \mathbf{y})$ (known)
- Estimate $p(c_3^* \mid \mathbf{y})$ from $\hat{p}(c_3^* \mid \mathbf{y}) = G^{-1} \sum_{g=1}^{G} p(c_3^* \mid a_3^{(g)}, b_3^{(g)}, \sigma_3^{2(g)}, \mathbf{y})$
- Acquire a sample L from $a_3, b_3, \sigma_3^2 \mid c_3^*, \mathbf{y}$ through Gibbs sampling with the distributions $p(a_3 \mid b_3, c_3^*, \sigma_3^2, \mathbf{y})$, $p(b_3 \mid a_3, c_3^*, \sigma_3^2, \mathbf{y})$, $p(a_2 \mid b_2^*, \sigma_2^2, \mathbf{y})$ and $p(\sigma_3^2 \mid a_3, b_3, c_3^*, \mathbf{y})$
- Estimate $p(b_3^* \mid c_3^*, \mathbf{y})$ from $\hat{p}(b_3^* \mid c_3^*, \mathbf{y}) = L^{-1} \sum_{l=1}^{L} p(b_3^* \mid a_3^{(l)}, c_3^*, \sigma_3^{2(l)}, \mathbf{y})$
- Acquire a sample M from $a_3, \sigma_3^2 \mid b_3^*, c_3^*, \mathbf{y}$ through Gibbs sampling with the distributions $p(a_3 \mid b_3^*, c_3^*, \sigma_3^2, \mathbf{y})$ and $p(\sigma_3^2 \mid b_3^*, c_3^*, a_3, \mathbf{y})$
- Estimate $p(a_3^* \mid b_3^*, c_3^*, \mathbf{y})$ from $\hat{p}(a_3^* \mid b_3^*, c_3^*, \mathbf{y}) = M^{-1} \sum_{m=1}^{M} p(a_3^* \mid b_3^*, c_3^*, \sigma_3^{2(m)}, \mathbf{y})$

5.3.5.2 The Chib and Jeliazkov Estimator

The posterior ordinate estimator of Chib and Jeliazkov reviewed in section 4.5.2 is based on the equation

$$p(\boldsymbol{\theta}^* \mid \mathbf{y}) = \frac{\int a_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) q(\boldsymbol{\theta}, \boldsymbol{\theta}^* \mid \mathbf{y}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}}{\int a_{MH}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}^*, \boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}}.$$

As discussed earlier in this chapter we use an independence M-H chain which is

allowed to depend on the data $\mathbf{y}$, that is $q(\boldsymbol{\theta}_j^{t-1}, \boldsymbol{\theta}_j^t \mid \mathbf{y}) = q(\boldsymbol{\theta}_j^t \mid \mathbf{y})$; see section 5.2.2.

Thus, the posterior ordinate under each model M_j, for $j = 0,1,2,3$ is given by

$$p(\boldsymbol{\theta}_j^* \mid \mathbf{y}) = \frac{\int a_{MH}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*) q(\boldsymbol{\theta}_j^* \mid \mathbf{y}) p(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j) d\boldsymbol{\theta}}{\int a_{MH}(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j) q(\boldsymbol{\theta}_j \mid \mathbf{y}) d\boldsymbol{\theta}}.$$

The probabilities of transition in the numerator and denominator are given by

$$a_{MH}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j^*) = \min\left[1, \frac{p\left(\boldsymbol{\theta}_j^* \mid \mathbf{y}, M_j\right) q\left(\boldsymbol{\theta}_j \mid \mathbf{y}\right) \frac{1}{\sigma_j}}{p\left(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j\right) q\left(\boldsymbol{\theta}_j^* \mid \mathbf{y}\right) \frac{1}{\sigma_j^*}}\right]$$

and

$$a_{MH}(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j) = \min\left[1, \frac{p\left(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j\right) q\left(\boldsymbol{\theta}_j^* \mid \mathbf{y}\right) \frac{1}{\sigma_j^*}}{p\left(\boldsymbol{\theta}_j^* \mid \mathbf{y}, M_j\right) q\left(\boldsymbol{\theta}_j \mid \mathbf{y}\right) \frac{1}{\sigma_j}}\right].$$

So, we estimate the posterior ordinates $p(\boldsymbol{\theta}_j^* \mid \mathbf{y})$, for $j = 0,1,2,3$ from

$$\widehat{p}(\boldsymbol{\theta}_j^* \mid \mathbf{y}) = \frac{\sum_{m=1}^{M} a_{MH}(\boldsymbol{\theta}_j^{(m)}, \boldsymbol{\theta}_j^*) q(\boldsymbol{\theta}_j^* \mid \mathbf{y})}{\sum_{k=1}^{K} a_{MH}(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j^{(k)})}.$$

where $\boldsymbol{\theta}_j^{(m)}$ are draws from the M-H outputs and $\boldsymbol{\theta}_j^{(k)}$ are draws from the distributions $q(\boldsymbol{\theta}_j \mid \mathbf{y})$ presented in section 5.1.2. Thus, the marginal likelihood estimates on logarithmic scale are given by

$$\log \widehat{p}(\mathbf{y} \mid M_j)_{C-J} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_j^*, M_j) + \log p(\boldsymbol{\theta}_j^* \mid M_j) - \log\left(\frac{\sum_{m=1}^{M} a_{MH}(\boldsymbol{\theta}_j^{(m)}, \boldsymbol{\theta}_j^*) q(\boldsymbol{\theta}_j^* \mid \mathbf{y})}{\sum_{k=1}^{K} a_{MH}(\boldsymbol{\theta}_j^*, \boldsymbol{\theta}_j^{(k)})}\right).$$

An important remark is that new sampling from the distributions $q(\boldsymbol{\theta}_j \mid \mathbf{y})$ was not necessary for the implementation of the method. Due to the use of an

independence chain, we simply kept all the proposed draws from $q(\boldsymbol{\theta}_j \mid \mathbf{y})$ during the M-H runs. This also means that the sample sizes cancel out in the estimation of the posterior ordinates since $K = M = 50.000$. In addition, the quantities $q(\boldsymbol{\theta}_j^* \mid \mathbf{y})$ are in fact constant and need to be calculated only once.

5.3.6 Chen estimator

According to this method we estimate the marginal likelihoods $p(\mathbf{y} \mid M_j)$ from

$$\log \hat{p}(\mathbf{y} \mid M_j)_{Chen} = \log p(\mathbf{y} \mid \boldsymbol{\theta}_j^*, M_j) - \log\left[G^{-1} \sum_{g=1}^{G} \frac{g(\boldsymbol{\theta}_j^{(g)} \mid M_j)}{p(\boldsymbol{\theta}_j^{(g)} \mid M_j)} \frac{p(\mathbf{y} \mid \boldsymbol{\theta}_j^*, M_j)}{p(\mathbf{y} \mid \boldsymbol{\theta}_j^{(g)}, M_j)} \right],$$

where $\boldsymbol{\theta}_j^{(g)}$ are the draws from the Gibbs simulations and $\boldsymbol{\theta}_j^*$ is again the point which maximizes the log target density under each model M_j, for $j = 0,1,2,3$.

The distributions $g(\cdot \mid M_j)$ are approximating densities of the posterior distributions $p(\boldsymbol{\theta}_j \mid \mathbf{y}, M_j)$. We actually use the same approximating densities which were utilized for the Bridge sampling estimators; see section 5.3.4.

5.4 Comparing the models

In this example, we have used conjugate priors in order to be able to calculate the marginal likelihoods under each model $p(\mathbf{y} \mid M_j)$ analytically and compare them with the estimates obtained from each method. In Table 5.9 we present the true marginal likelihood values on physical scale and on natural logarithmic scale along with the posterior probabilities under the assumption that each model is equally probable a-priori, that is $p(M_j) = 0.25$ for $j = 0,1,2,3$. Model 2 and model 3 seem to be clearly preferable from model 0 and model 1. Between the former, model 2 yields the highest marginal likelihood value and posterior probability.

	Model 0	Model 1	Model 2	Model 3
Marginal	7.111×10^{-16}	1.959×10^{-6}	0.2028	0.1078
Log-Marginal	-34.8797	-13.1429	-1.5953	-2.2270
Posterior Probability	2.288×10^{-15}	6.306×10^{-6}	0.6528	0.3471

Table 5.9 *Marginal likelihood of the data under each regression model calculated in physical and logarithmic scale and the resulting posterior probabilities under the assumption that the competing models are equally probable a-priori. Model 2 yields the highest marginal likelihood and posterior probability.*

In Table 5.10 we compare the models considering twice the natural logarithm of the Bayes factor.

$\mathbf{2lnBF_{ij}}$				
Model (*i*)	**Model (*j*)**			
	$j=0$	$j=1$	$j=2$	$j=3$
$i=1$	43.47	1	-23.1	-21.83
$i=2$	66.57	23.1	1	1.26
$i=3$	65.31	21.83	-1.26	1

Table 5.10 *Comparison of the models based on twice the natural logarithm of Bayes Factors. Evidence against model 0 is very strong when compared to models 1, 2 and 3. Model 2 is also very strongly supported in comparison to model 1 but not in comparison to model 3; between model 2 and model 3, evidence in favor of the former is "not worth than a bare mention".*

Based on the interpretations discussed in section 2.7 we can say that evidence against the simple model is very strong when compared to models 1, 2 and 3. In addition, the comparisons of model 2 and model 3 versus model 1, respectively, indicate very strong evidence against the latter. This is not the case for models 2 and 3; the value 1.26 is not adequate in order to make decisive statements against model 3 or in favor of model 2.

From the posterior sample acquired by the Gibbs sampler we also evaluate the AIC, BIC and DIC criteria. The variation of AIC and BIC evaluated at the posterior mean of the deviance is also computed. Results are summarized in Table 5.11. In contrast to the marginal likelihoods and Bayes factors which tend to support model 2, all information criteria conclude to model 3 as the best fitted model.

	$AIC_{\min(D)}$	$AIC_{\bar{D}}$	$BIC_{\min(D)}$	$BIC_{\bar{D}}$	DIC
Model 0	52.56	52.63	55	55.06	52.54
Model 1	2.68	2.95	6.34	6.61	2.65
Model 2	-24.32	-23.24	-20.66	-19.59	-23.97
Model 3	-29.15	-27.59	-24.27	-22.71	-28.46

Table 5.11 *The calculated AIC, BIC and DIC values for each model. AIC and BIC are evaluated at the minimum value and at the posterior mean of the deviance. All information criteria support Model 3 which yields the smallest values.*

5.5 Comparing results

In this section we will compare results obtained from each method with respect to the true marginal likelihood values. All estimates except that of Chib and Jeliazkov were obtained from the Gibbs sampler. The Gibbs and M-H posterior samples are of size 50.000. The marginal likelihood estimates, calculated on logarithmic scale and rounded up to four decimal places, are summarized in Table 5.12.

As we can see the Harmonic Mean (HM) point estimates are more distant from the true marginal values than any other estimates. The Laplace-Metropolis method produces in general better results. Among the Laplace-Metropolis estimators the ones evaluated at the posterior mean (LM_{Mean}) are closer to the true marginal likelihoods. Newton and Raftery's estimators are strongly affected from the selection of starting values and from the number of iterations. For the first estimator (NR1) we used the HM results as initial values and iterated the equation 5000 times. These estimates are substantially better than the HM estimates but still distanced from the true marginal likelihood values. Using the LM_{Mean} estimates as initial values resulted to the estimators NR2 and NR3 after 1000 and 10000 iterations respectively; we can notice that the increase in the number of iterations resulted to point estimates which are a little closer to the target values, yet convergence is extremely slow. The Bridge sampling estimators are among the most accurate; the Geometric (GB) and Optimum (OptB) point estimates converge to the true marginal values with an accuracy of two decimal places. Chib's estimator also produces satisfactory results; its point estimates for models 1 and 2 are closest to the true marginal likelihood values

than any other estimates. This is not the case for the Chib-Jeliazkov's (C-J) estimates which are actually similar to the Laplace-Metropolis estimates. Finally, Chen's estimates are also very close to the target values, especially the point estimates for models 0 and 3 which are more accurate than any other estimates.

		Model			
Method	**Estimators**	$j=0$	$j=1$	$j=2$	$j=3$
Harmonic Mean Estimator	$\log \widehat{p}(\mathbf{y} \mid M_j)_{HM}$	-27.0795	-2.3291	10.3757	12.3474
Laplace-Metropolis Estimators	$\log \widehat{p}(\mathbf{y} \mid M_j)_{LM}$	-34.9255	-12.0834	-0.0748	-0.4867
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{LM_{MEDIAN}}$	-35.0587	-12.2900	-0.2847	-0.7894
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{LM_{MEAN}}$	-35.1299	-12.3835	-0.3738	-0.8947
Newton and Raftery's Estimator	$\log \widehat{p}(\mathbf{y} \mid M_j)_{NR1}$	-35.4089	-10.8035	1.8898	3.8843
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{NR2}$	-35.4137	-12.4280	-0.3968	-0.8966
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{NR3}$	-34.1560	-12.7516	-0.5790	-0.9136
Bridge Sampling Estimators	$\log \widehat{p}(\mathbf{y} \mid M_j)_{HB}$	-34.7110	-13.8097	-1.4798	-2.6405
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{GB}$	-34.8789	-13.1437	-1.5941	-2.2278
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{OptB}$	-34.8788	-13.1436	-1.5940	-2.2275
Candidate's Estimators	$\log \widehat{p}(\mathbf{y} \mid M_j)_{Chib}$	-34.8792	-13.1429	-1.5950	-2.2305
	$\log \widehat{p}(\mathbf{y} \mid M_j)_{C-J}$	-32.8172	-12.4058	-0.3959	-0.7456
Chen's Estimator	$\log \widehat{p}(\mathbf{y} \mid M_j)_{Chen}$	-34.8799	-13.1448	-1.5931	-2.2273
TARGET		**-34.8797**	**-13.1429**	**-1.5953**	**-2.2270**

Table 5.12 *Marginal likelihood estimates for the four competing regression models resulting from simulated posterior samples of 50000 draws. Gibbs sampling has been used for all estimators except that of Chib and Jeliazkov (C-J) which was calculated through the use of Metropolis-Hastings simulation.*

In Table 5.13 we present the posterior probabilities estimates for model 2 and model 3 obtained from the corresponding marginal likelihood estimates - the

posterior probabilities of models 0 and 1 are not shown, since they are very close to zero. Concerning the Laplace-Metropolis and Newton and Raftery's estimators, we restrict attention to the estimates LM_{Mean} and NR3 respectively. As we can see, the HM method provides clearly the most inaccurate estimates, since it assigns greater posterior probability in model 3.

		Model	
Method	**Estimators**	$j=2$	$j=3$
Harmonic Mean Estimator	$\hat{p}(M_j \mid \mathbf{y})_{HM}$	0.1222	0.8778
L-M Estimator	$\hat{p}(M_j \mid \mathbf{y})_{LM_{MEAN}}$	0.6274	0.3726
N & R Estimator	$\hat{p}(M_j \mid \mathbf{y})_{NR3}$	0.5829	0.4171
Bridge Sampling Estimators	$\hat{p}(M_j \mid \mathbf{y})_{HB}$	0.7615	0.2385
	$\hat{p}(M_j \mid \mathbf{y})_{GB}$	0.6533	0.3467
	$\hat{p}(M_j \mid \mathbf{y})_{OptB}$	0.6533	0.3467
Candidate's Estimators	$\hat{p}(M_j \mid \mathbf{y})_{Chib}$	0.6537	0.3463
	$\hat{p}(M_j \mid \mathbf{y})_{C-J}$	0.5866	0.4134
Chen's Estimator	$\hat{p}(M_j \mid \mathbf{y})_{Chen}$	0.6534	0.3466
TARGET		**0.6528**	**0.3471**

Table 5.13 *Posterior probabilities estimates for model 2 and model 3 (the corresponding estimates for models 0 and 4 are not presented since their posterior probabilities are very close to zero).*

In order to estimate the MC error of the marginal likelihood estimates we use the batch mean method utilizing 50 batches of size 1000. The batched marginal likelihood estimates and the corresponding batched MC error (batched standard deviations) estimates are presented in Table 5.14. The Geometric Bridge (GB) and Optimum Bridge (OptB) estimates have the smallest batched standard deviations among all estimators. Chib's and Chen's estimates follow; the former results to lower batched standard deviations for models 0, 1 and 2 while the latter

has a lower standard deviation for model 3. Among the rest, Chib-Jeliazkov's (C-J) estimates have the smallest standard deviations. The methods that result to the highest deviations are the Harmonic Bridge (HB) and the Harmonic Mean (HM) methods.

		Model			
Method	**Estimators**	$j=0$	$j=1$	$j=2$	$j=3$
Harmonic Mean Estimator	$\log \hat{p}(\mathbf{y} \mid M_j)_{HM}$	-26.7117 (0.1011)	-1.8917 (0.1077)	10.7979 (0.1083)	12.9799 (0.1338)
Laplace-Metropolis Estimator (at mean)	$\log \hat{p}(\mathbf{y} \mid M_j)_{LM_{MEAN}}$	-35.1311 (0.0032)	-12.3849 (0.0042)	-0.3751 (0.0035)	-0.8976 (0.0048)
Newton & Raftery's Estimator	$\log \hat{p}(\mathbf{y} \mid M_j)_{NR3}$	-36.1431 (0.1885)	-12.6974 (0.0439)	-0.5580 (0.0290)	-0.9161 (0.0066)
Bridge Sampling Estimators	$\log \hat{p}(\mathbf{y} \mid M_j)_{HB}$	-35.0194 (0.1719)	-13.5857 (0.1489)	-1.9113 (0.2001)	-2.7294 (0.2022)
	$\log \hat{p}(\mathbf{y} \mid M_j)_{GB}$	-34.8789 (0.0006)	-13.1437 (0.0010)	-1.5941 (0.0010)	-2.2278 (0.0010)
	$\log \hat{p}(\mathbf{y} \mid M_j)_{OptB}$	-34.8788 (0.0006)	-13.1436 (0.0010)	-1.5940 (0.0010)	-2.2275 (0.0009)
Candidate's Estimators	$\log \hat{p}(\mathbf{y} \mid M_j)_{Chib}$	-34.8792 (0.0007)	-13.1429 (0.0010)	-1.5950 (0.0009)	-2.2302 (0.0036)
	$\log \hat{p}(\mathbf{y} \mid M_j)_{C-J}$	-32.8172 (0.0015)	-12.4058 (0.0017)	-0.3959 (0.0024)	-0.7456 (0.0037)
Chen's Estimator	$\log \hat{p}(\mathbf{y} \mid M_j)_{Chen}$	-34.8799 (0.0013)	-13.1447 (0.0021)	-1.5930 (0.0016)	-2.2272 (0.0017)
TARGET		**-34.8797**	**-13.1429**	**-1.5953**	**-2.2270**

Table 5.14 *Batched marginal likelihood estimates and MC error estimates (in brackets) resulting from 50 batches of size 1000.*

In conclusion, we could say that the Geometric and the Optimum Bridge sampling estimators, Chib's candidate's estimator and Chen's estimator are the methods which perform substantially better according to the aforementioned results. The ergodic means plots of these estimates for all four competing models are presented in Figures 5.1 to 5.4. One can notice the slightly higher batched

MC error resulting from Chen's method for models 0, 1 and 2 and from Chib's method for model 3.

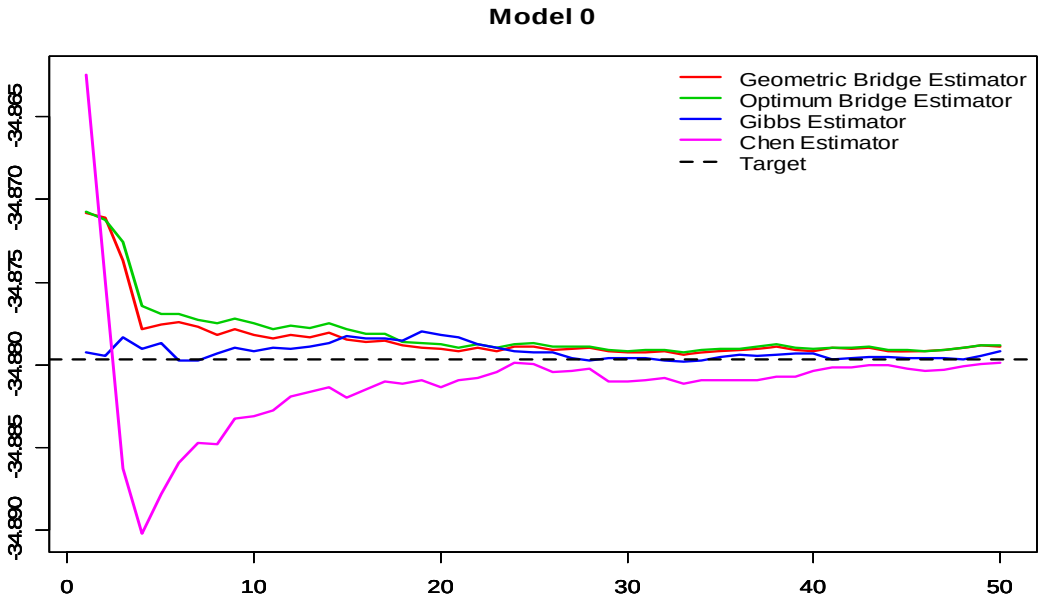

Figure 5.1 *Ergodic mean plots for the four marginal likelihood estimates of model 0 which perform better; the Geometric and Optimum Bridge sampling estimators and Chib's and Chen's estimators. The black dashed line represents the true marginal density of model 0.*

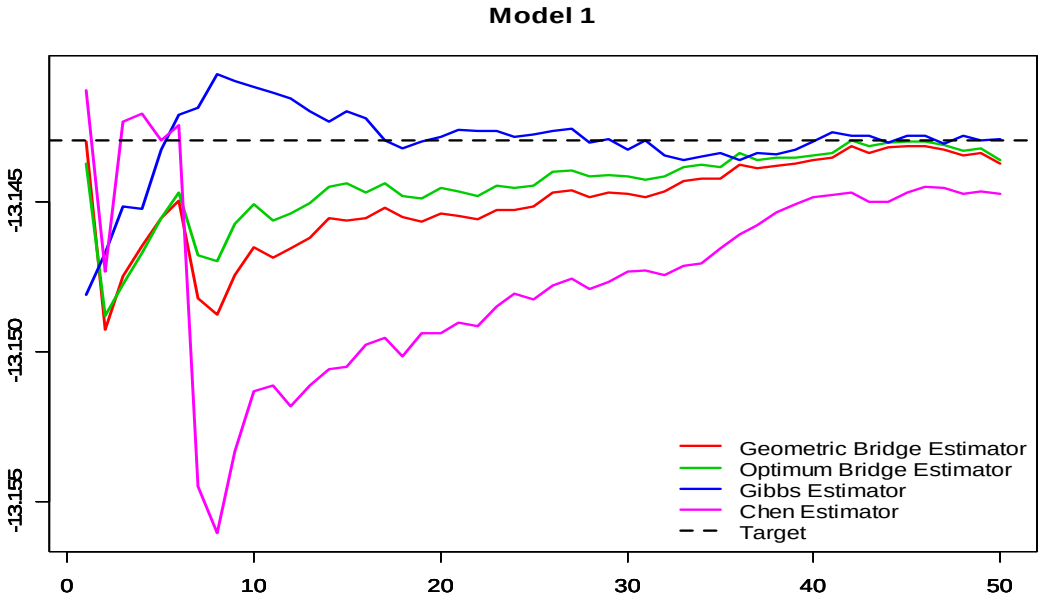

Figure 5.2 *Ergodic mean plots for the four marginal likelihood estimates of model 1 which perform better; the Geometric and Optimum Bridge sampling estimators and Chib's and Chen's estimators. The black dashed line represents the true marginal density of model 1.*

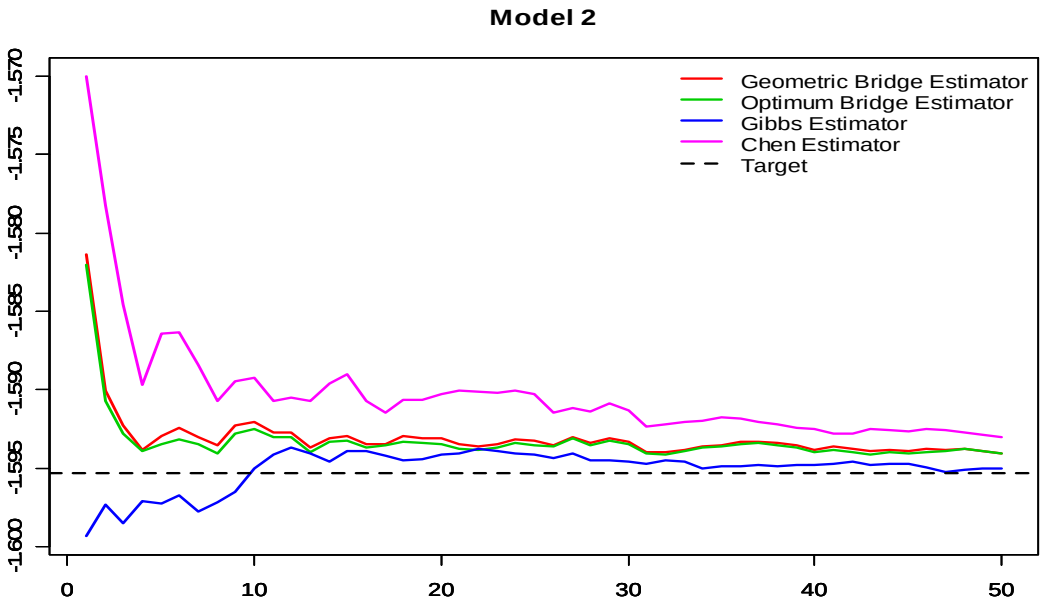

Figure 5.3 *Ergodic mean plots for the four marginal likelihood estimates of model 2 which perform better; the Geometric and Optimum Bridge sampling estimators and Chib's and Chen's estimators. The black dashed line represents the true marginal density of model 2.*

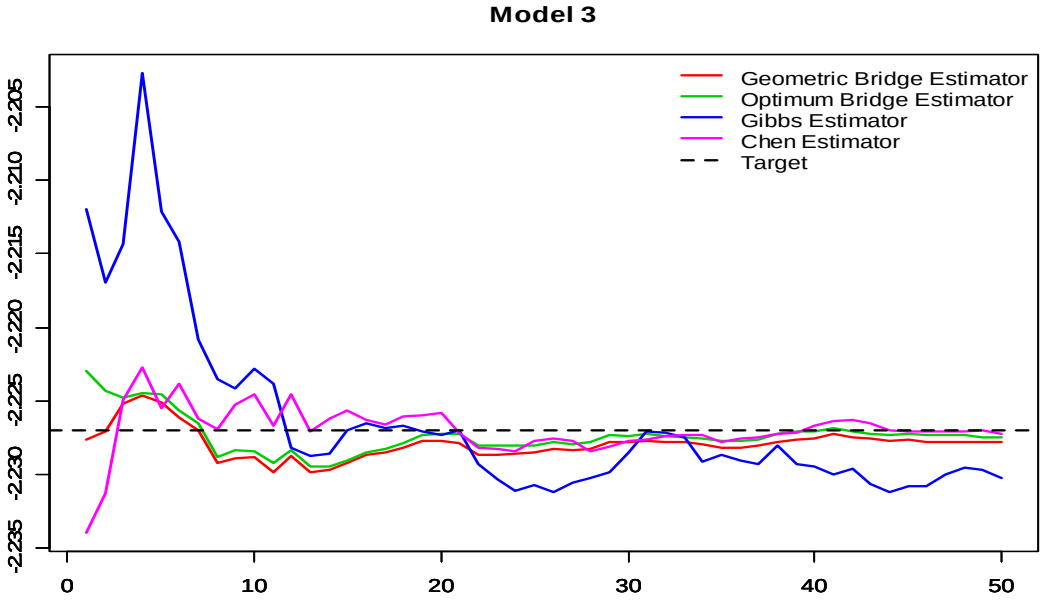

Figure 5.4 *Ergodic mean plots for the four marginal likelihood estimates of model 3 which perform better; the Geometric and Optimum Bridge sampling estimators and Chib's and Chen's estimators. The black dashed line represents the true marginal density of model 3.*

Despite that not all of the methods result to accurate marginal likelihood estimates some may still produce results that are accurate enough for interpretation of Bayes factors on logarithmic scale. Point estimates of twice the natural logarithm of the Bayes Factor of model 2 versus model 3 are shown in Table 5.15.

Method	**Point Estimates of** $2\ln BF_{23}$	
Harmonic Mean Estimator	$2\ln\widehat{BF}_{23(HM)}$	-3.9432
Laplace-Metropolis Estimator (at mean)	$2\ln\widehat{BF}_{23(LM_{MEAN})}$	1.0418
Newton & Raftery's Estimator	$2\ln\widehat{BF}_{23(NR3)}$	0.6669
Bridge Sampling Estimators	$2\ln\widehat{BF}_{23(HB)}$	2.3214
	$2\ln\widehat{BF}_{23(GB)}$	1.2675
	$2\ln\widehat{BF}_{23(OptB)}$	1.2670
Candidate's Estimators	$2\ln\widehat{BF}_{23(Chib)}$	1.2711
	$2\ln\widehat{BF}_{23(C-J)}$	0.6442
Chen's Estimator	$2\ln\widehat{BF}_{23(Chen)}$	1.2684
TARGET		**1.2635**

Table 5.15 *Point estimates of twice the natural logarithm of the Bayes Factor of model 2 versus model 3; the Optimum and Geometric bridge sampling estimates (OptB and GB) along with Chib's and Chen's estimates are closer to the true value.*

The methods which perform well in the estimation of the marginal likelihoods produce of course satisfactory points estimates; the Optimum Bridge (OptB) estimate is closer to the true target value than any other estimate while the Geometric Bridge (GB), Chen's and Chib's estimates follow. The Laplace-Metropolis approximation also results to a fairly satisfactory estimate (LM_{Mean}). Chib-Jeliazkov's (C-J) and Newton and Raftery's (NR3) estimates are more distant from the target value, yet these estimates do not affect the interpretation

of $2\ln BF_{23}$. Unlikely, the Harmonic Bridge estimate (HB) indicates stronger evidence in favor of model 2 and therefore affects the interpretation. The Harmonic Mean (HM) estimate changes the interpretation totally from "weak evidence in favor of model 2" to "positive evidence against model 2".

Batched estimates of $2\ln BF_{23}$ and the corresponding MC errors along with 95% confidence intervals resulting from the percentiles of the 50 batches are presented in Table 5.16.

Method	**Estimator**	**Batched point estimates**	**95% Percentiles C.I.**
Harmonic Mean Estimator	$2\ln \widehat{BF}_{23(HM)}$	-4.3640 (0.3409)	(-9.1272,-0.1488)
Laplace-Metropolis Estimator (at mean)	$2\ln \widehat{BF}_{23(LM_{MEAN})}$	1.0449 (0.0121)	(0.8850,1.2286)
Newton & Raftery's Estimator	$2\ln \widehat{BF}_{23(NR3)}$	0.7194 (0.0618)	(-0.0922 ,1.0168)
Bridge Sampling Estimators	$2\ln \widehat{BF}_{23(HB)}$	1.6361 (0.5543)	(-6.3735, 8.3093)
	$2\ln \widehat{BF}_{23(GB)}$	1.2675 (0.0026)	(1.2388,1.2983)
	$2\ln \widehat{BF}_{23(OptB)}$	1.2669 (0.0026)	(1.2340,1.2977)
Candidate's Estimators	$2\ln \widehat{BF}_{23(Chib)}$	1.2705 (0.0072)	(1.1683,1.3715)
	$2\ln \widehat{BF}_{23(C-J)}$	0.6445 (0.0069)	(0.5525,0.7282)
Chen's Estimator	$2\ln \widehat{BF}_{23(Chen)}$	1.2684 (0.0046)	(1.2096,1.3276)
TARGET		**1.2635**	

Table 5.16 *Batched estimates of twice the natural logarithm of the Bayes factor of model 2 versus model 3 and the corresponding MC errors (in brackets), derived from 50 batches of size 1000. The 95% confidence intervals resulting from the percentiles of the 50 batches are shown in the last column to the right .The Optimum (OptB) and the Geometric (GB) bridge sampling estimators produce the smallest batched standard deviation.*

The Geometric (GB) and Optimum Bridge (OptB) sampling methods result again to accurate point estimation and to the lowest batched MC errors. Chen's and Chib's estimates follow with the latter resulting to a slightly higher batched standard deviation. Chib-Jeliazkov's (C-J) estimate is far from the target value

but results to a small batched standard deviation, while the Laplace-Metropolis estimate (LM_{Mean}) is closer to the target value but yields a higher batched standard deviation. Nevertheless, none of the latter two estimates affect the interpretation of the Bayes Factor. This is not the case for the rest of the estimates especially for the Harmonic Bridge (HB) and Harmonic Mean (HM) estimates which result to the highest batched MC errors.

5.6 Summary

In this chapter we implemented the marginal likelihood estimation methods reviewed in chapter 4 for four linear regression models. Metropolis-Hastings simulation and Gibbs sampling were utilized in order to acquire posterior samples from the four competing models. Posterior summaries for the corresponding parameters of each model resulting from both simulation methods were presented.

We then described the implementation process for each estimation method and finally compared the resulting marginal likelihood, Bayes Factor and posterior probability estimates with respect to the true values of the corresponding quantities. The main conclusions derived from the preceding analysis are the following:

- The Optimum and Geometric Bridge sampling estimators performed overall better than the rest of the methods. They produced accurate point estimates and the lowest batched standard deviations for all models.
- Chib's and Chen's estimates were also accurate with relatively low batched standard deviations but without displaying the same stability among the different models.
- The Harmonic Bridge and Harmonic mean estimators - especially the latter - proved to be unstable; the resulting estimates were not accurate even for interpretations of Bayes Factors on logarithmic scale.
- The rest of the methods produced results satisfactory enough for interpretations of Bayes Factors on logarithmic scale.

Chapter 6: Conclusions and Further Discussion

6.1 Conclusions concerning the marginal likelihood estimators

Based on the implementation process and on the results obtained and presented in the previous chapter certain final remarks can be made for each marginal likelihood estimation method reviewed in this thesis.

The Harmonic Mean method is easy to implement but it is clearly unstable and fails to estimate accurately the marginal likelihood. The resulting estimates were more distant from the true values than any other estimate. In addition, its results proved to be unsafe for interpretation of Bayes Factors on the logarithmic scale.

The Laplace-Metropolis method is based on an approximating result; hence the estimates derived from this method were not accurate. Despite this fact it is easily implemented and can it can produce results that are accurate enough for interpretation on logarithmic scale.

The Newton and Raftery's method provided estimates which were closer to the true marginal likelihood values but still not accurate enough. In addition, its iterative process required substantial computer time since convergence proved to be slow. As discussed in section 4.4, this method requires sampling from the prior distribution therefore the results presented in this thesis may be affected from the extremely vague prior selection for the linear regression nuisance parameter σ_j^2, for models $j = 0,1,2,3$.

The Harmonic Bridge sampling estimates were in general close to the true marginal likelihood values but shared the instability of the Harmonic Mean estimates. These estimates resulted to the highest batched standard deviations and have proven to be unsafe for interpretation of Bayes Factors between models with similar marginal likelihoods. Contrary, the Geometric and Optimum estimators have proven to be more accurate and stable than all other estimates; the results were accurate to the second decimal place for all models and had the lowest standard errors observed. In addition, these methods were easier to

implement and less time consuming than other methods. As discussed in section 4.5 Bridge sampling techniques are based on the existence of an approximating density. So, as long as a satisfactory approximating density can be found, Bridge sampling seems to be the most trustworthy method of marginal likelihood estimation. This implies that Bridge sampling may be difficult to implement in high dimensional problems; nevertheless, increased dimensionality is a setback for nearly all methods.

The Candidate's method produced diverse results. The Gibbs sampling based method of Chib resulted to satisfactory estimates which were accurate and had relatively small standard errors. More specific, Chib's method provided the most accurate marginal likelihood point estimates for two out of the four competing models. However, the method can be time consuming in terms of computer programming, especially for high dimensional cases, although radically new programming is not required and clever blocking of the parameter vector may facilitate the implementation process. Chib's method has also received criticism; Neal (1999) argues that this method produces bias for marginal likelihood estimates of mixture models, hidden Markov models and other models with similar symmetries; see also Frühwirth-Schnatter (2004). The Metropolis-Hastings based method of Chib and Jeliazkov did not perform as well as the method of Chib. Despite the small batched standard deviations, the resulting estimates were not as accurate as the estimates derived from the method of Chib. Nevertheless, they proved to be safe for interpretation of Bayes Factors on logarithmic scale. The method is rather computationally intense; as we have seen in section 4.6.2 it is based on the reversibility of the sub-kernel of the Metropolis-Hastings algorithm and therefore requires draws from the proposal distributions and calculation of probabilities from these densities. So, clever blocking is recommended before implementing this method in high dimensional problems. In addition, experimenting with alternative choices such as the multivariate t distribution instead of the multivariate normal distribution may result to more accurate estimates than those presented in this thesis.

Finally, the method of Chen also resulted in satisfactory estimates and low standard errors. Two of the marginal likelihood point estimates were more accurate than the corresponding Optimal estimates of Bridge sampling but overall batched standard deviations were higher. Viewed as a generalization of the Candidate's method, the method of Chen has certain advantages with respect to the former; first it does not depend on the specific form of the MCMC sampling process and second it is less time consuming and much easier to implement. As with Bridge sampling the method of Chen requires an approximating density, therefore, performance actually depends on the precision of the approximation.

6.2 Further discussion

As we have seen, all methods investigated in this thesis are "direct" methods which use existing MCMC posterior samples in order to estimate marginal likelihoods of separate models and then evaluate the corresponding Bayes Factors; this by itself implies much effort when the number of competing models is large. Alternatives to these "direct" methods are existing MCMC algorithms like the RJMCMC algorithm (Green, 1995), the Carlin and Chib algorithm (Carlin and Chib, 1995) and the Metropolised Carlin and Chib algorithm (Dellaportas et al., 2002). These methods sample simultaneously over parameter and model space and deliver posterior model probabilities. Yet, neither the implementation of such methods is effortless since they require careful specification of all competing models along with certain tuning constants in order to ensure successful mixing in model space. In addition, they do not perform always better; Han and Carlin (2001) compare some of the aforementioned methods and the method of Chib in linear regression models and conclude that Chib's point estimate is more accurate than those of five of the other six methods.

Finally, Bayes Factors themselves have received some criticism for not always being the safest solution regarding model selection. The main argument

against Bayes Factors is their apparent sensitivity to the specification of the prior distribution. The sensitivity of Bayes Factors to prior selection can be observed for both proper and improper prior distributions, yet the latter case is more problematic since Bayes Factors are not interpretable under improper priors; see Kass and Raftery (1995) and Draper (1995) and the associated discussion.

Several authors proposed alternative versions of Bayes Factors in order to cope with this problem. Among them, Aitkin (1991) proposed the use of *Posterior Bayes Factors* which remain interpretable under improper priors, while O'Hagan (1995) introduced the *Fractional Bayes Factors* based in the use of a training sample in order to acquire prior information.

Alternative options to Bayes Factors regarding model selection are also available. As already mentioned, the simplest option is the use of Information Criteria. More elaborate approaches are mainly based on predictive schemes; Geisser and Eddy (1979) were among the first to propose the use of conditional predictive densities for model selection, Gelfand et al. (1992) argue that model selection remains closely linked to model assessment and proposed several cross-validatory analyses of predicted residuals, Laud and Ibrahim (1995) introduced three model selection criteria based on the predictive density while Waller et al. (1997) extended their methods, Greenberg and Parks (1997) recommended the examination of changes in predicted means and general variance ratios. Finally, Gelfand and Ghosh (1998) introduced a more general approach, based on loss functions, which aims to minimize the posterior predictive loss.

REFERENCES

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

Akaike, H. (1974). A new look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.

Berger, J. and Bernando, J.M. (1992). On the Development of Reference Points (with discussion), in *Bayesian Statistics*, 4, Oxford University Press, 35-60.

Bernando, J.M. (1979). Reference Posterior Distributions for Bayesian Inference (with discussion), *Journal of the Royal Statistical Society*, B, 41, 113-147.

Besag, J. (1989). A Candidate's Formula: A Curious Result in Bayesian Prediction, *Biometrika*, 76, 183.

Besag, J. (2001). Markov Chain Monte Carlo for Statistical Inference, Working Paper, No. 9, Center for Statistics and the Social Sciences, University of Washington, USA.

Brooks, S.P. (1998). Markov Chain Monte Carlo Methods and its Application, *The Statistician*, 47**,** 69-100.

Brooks, S.P. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde, *Journal of the Royal Statistical Society*, B, 64, 616-618.

Brooks, S.P., Dellaportas, P. and Roberts, G.O. (1997). An Approach to Diagnosing Total Variation Convergence of MCMC Algorithms, *Journal of Computational and Graphical Statistics*, 6, 251-265.

Brooks, S.P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations, *Journal of Computational and Graphical Statistics*, 7, 434-455.

Brooks, S.P. and Roberts, G.O. (1998). Convergence Assessment Techniques for Markov Chain Monte Carlo, *Statistics and Computing*, 8, 319-335.

Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society*, B, 57, 473-484.

Carlin, B.P. and Louis, T. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, first edition, Chapman and Hall, London.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler, *The American Statistician*, 46, 167-174.

Chen, M.-H. (1994). Importance-Weighted Marginal Bayesian Posterior Density Estimation, *Journal of the American Statistical Association*, 89, 818-824.

Chen, M.-H. (2005). Computing Marginal Likelihoods from a Single MCMC Output, *Statistica Neerlandica*, 59, 16-29.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association*, 90, 1313-1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, 327-335.

Chib, S. and Jeliazkov, I. (2001). Marginal Likelihood from the Metropolis-Hastings Output, *Journal of the American Statistical Association*, 96, 270-281.

Congdon, P. (2004). Bayesian Model Choice Based on Monte Carlo Estimates of Posterior Model Probabilities, *Computational Statistics & Data Analysis*, 50, 346-357.

Consonni, G. and Veronese, P. (1992). Conjugate Priors for Exponential Families Having Quadratic Variance Function, *Journal of the American Statistical Association*, 87, 1123-1127.

Cowles, M.K. and Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association*, 91, 883-904.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, Chapman and Hall, London.

Dalal, S.R. and Hall, W.J. (1983). Approximating Priors by Mixtures of Natural Conjugate Priors, *Journal of the Royal Statistical Society*, B, 45, 278-286.

Datta, G.S. and Ghosh, M. (1995). Some Remarks on Noninformative Priors, *Journal of the American Statistical Association*, 90, 1357-1363.

Datta, G.S. and Ghosh, M. (1996). On the Invariance of Noninformative Priors, *The Annals of Statistics*, 24, 141-159.

Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2002). On Bayesian Model and Variable Selection Using MCMC**,** *Statistics and Computing*, 12, 27-36.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models*, Chapman and Hall, London.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discussion), *Journal of the Royal Statistical Society*, B, 57, 45-97.

Fernandez, C., Ley, E. and Steel, M.F.J. (2001). Benchmark Priors for Bayesian Model Averaging, *Journal of Econometrics*, 100, 381-427.

Frühwirth-Shnatter, S. (2004). Estimating Marginal Likelihoods for Mixture and Markov Switching Models Using Bridge Sampling Techniques, *Econometrics Journal*, 7, 143-167.

Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, second edition, Chapman and Hall, New York and London.

Geisser. S. and Eddy, W.F. (1979). A Predictive Approach to Model Selection, *Journal of the American Statistical Association*, 75, 153-160.

Gelfand, A.E. (2000). Gibbs Sampling, *Journal of the American Statistical Association*, 96, 1300-1304.

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods (with discussion), in *Bayesian Statistics*, 4, Oxford University Press, 147-167.

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

Gelman, A., Carlin, J.B, Stern, H.S. and Rubin D.B. (1995). *Bayesian Data Analysis*, first edition, Chapman and Hall, London.

Gelman, A., Meng, X.-L. and Stern, H. (1993). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies (with discussion), *Statistica Sinica*, 6, 733-807.

Gelman, A. and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 7, 457-511.

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

Ghosh, J.K. and Mukerjee, R. (1992). Non-Informative Priors (with discussion), in *Bayesian Statistics*, 4, Oxford University Press, 195-210.

Ghosh, J.K. and Mukerjee, R. (1995). On Perturbed Ellipsoidal and Highest Posterior Density Regions with Approximate Frequentist Validity, *Journal of the Royal Statistical Society*, B, 57, 761-769.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York.

Green, P.J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, *Biometrika*, 82, 711-732.

Greenberg, E. and Parks, R.P. (1997). A Predictive Approach to Model Selection and Multicollinearity, *Journal of Applied Econometrics*, 12, 67-75.

Gutierrez-Pena, E. and Smith, A.F.M. (1995). Conjugate Parameterizations for Natural Exponential Families, *Journal of the American Statistical Association*, 90, 1347-1356.

Han, C. and Carlin, B.P. (2001). Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review, *Journal of the American Statistical Association*, 96, 1122-1132.

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

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial, *Statistical Science*, 14, 382-401.

Hyndman, R.J. (1996). Computing and Graphing Highest Posterior Density Regions, *The American Statistician*, 50, 120-126.

Ibrahim, J.G. and Laud, P.W. (1991). On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior, *Journal of the American Statistical Association*, 86, 981-986.

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

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

Kass, R.E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules, *Journal of the American Statistical Association*, 91, 1343-1370.

Laud, P. and Ibrahim, J. (1995). Predictive Model Selection, *Journal of the Royal Statistical Society*, B, 57, 247-262.

Lewis, S.M. and Raftery, A.E. (1997). Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator, *Journal of the American Statistical Association*, 92, 648-655.

Lopes, H.B. (2002). Bayesian Model Selection, available at http://eagle.ufrj.br/~hedibert/modelselection.pdf, Universidade Federal do Rio de Jainero, Brazil.

Meng, X.-L. (1994). Posterior Predictive p-Values, *The Annals of Statistic*, 22, 1142-1160.

Meng, X.-L. and Wong, W.H. (1996). Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration, *Statistica Sinica*, 6, 831-860.

Mengersen, K.L., Robert, C.P. and Guihenneuc-Jouyaux, C. (1998). MCMC Convergence Diagnostics: A "REVIEWWW", available at http://citeseer.ist.psu.edu/78250.html, Universite Paris, France.

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

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*, third edition, Wiley, New York.

Morris, C.N. (1983). Natural Exponential Families with Quadratic Variance Functions: Statistical Theory, *The Annals of Statistics*, 11, 515-529.

Neal, R.M. (1999). Erroneous Results in "Marginal Likelihood from the Gibbs Output", available at http://www.cs.toronto.edu/~radford/ftp/chib-letter.pdf, University of Toronto, Canada.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap, *Journal of the Royal Statistical Society*, B, 56, 3-48.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison (with discussion), *Journal of the Royal Statistical Society*, B, 57, 99-138.

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

Rubin, D.B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician, *The Annals of Statistic*, 12, 1151-1172.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion), *Journal of the Royal Statistical Society*, B, 64, 583-639.

Tierney, L. and Kadane J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, 81, 82-86.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997). Hierarchical Spatio-Temporal Mapping of Disease Rates, *Journal of the American Statistical Association*, 92, 607-617.

Wright, D.E. (1986). A Note on the Construction of Highest Posterior Density Intervals, *Applied Statistics*, 35, 49-53.