National and Kapodistrian University of Athens
Medical School, Department of Mathematics

University of Ioannina
Deparment of Mathematics

"BAYESIAN INFERENCE AND VARIABLE SELECTION IN NORMAL AND BINOMIAL REGRESSION MODELS WITH APPLICATIONS IN MEDICAL RESEARCH"

by
Konstantinos Pateras
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Medical Statistics

Athens, 2013

Thesis Supervisor:
Prof. Ioannis Ntzoufras

The current Master thesis was conducted as part of the courses for obtaining the Master's Degree in

BIOSTATISTICS

conferred by the Medical School and the Department of Mathematics, National & Kapodistrian University of Athens and the department of Mathematics, University of Ioannina.

Approved the ..... / ..... / 2013 by the examining committee :

NAME	**GRADE**	**SIGNATURE**
I. NTZOUFRAS	Associate Professor	.....
D. KARLIS	Associate Professor	.....
L. MELIGOTSIDOU	Lecturer	.....

Abstract

Subjectivity in any science means
that we unreveal universal truths
by combining subjective and
objective methods.
– Unknown –

The most important stage of inference in statistics lies in model selection. The same principle applies in Bayesian perspective which is going to be examined, in this thesis. Even though the origin of the philosophy of Bayesian statistics lies way back in the early 18th century, not much steps forward were made, due to the cumbersome high dimensional integrals, which were difficult to be computed based on the primary use of techniques that helped mostly on one dimensional integrals. As years passed, technology made possible the calculation of challenging computational problems. The usage of Markov Chain Monte Carlo techniques, which were recently been re-discovered, produced a easy to implement methodology to deal with those kind of -until then prohibitive- computations.

In this Master Thesis we will deal with full Bayesian inference in both normal and binomial regression models when there is doubt about the structure of the linear combination and the parameters of the model. There will be a review of the relative methodology and emphasis will be placed in possible alternatives a priori distributions which can be used in such type of problems. The usage of advanced MCMC algorithms for a priori estimation of parameters, variable selection. In the frame of this Master thesis synopsis, application and comparison of existing code and programs in R and WinBUGS environments for variable selection and model averaging will be presented. The methodology will be applied in both simulated and medical data with emphasis on those derived from the European Health Interview Survey (EHIS) 2009 held in Greece.

In memory of my beloved grandparents.

Acknowledgments

Firstly, I should express my sheer gratitude to Prof. Ioannis Ntzoufras. Without his contribution this MSc Thesis would never have been fulfilled.

I would also like to thank Prof. Karlis D. for his constant initiative to cover in detail every fresh knowledge I was acquiring. Prof. Meligotsidou L. for her comments and the possibility that she gave me to gain experience in teaching courses.

During the last two years I was part of the Hellenic Center of Disease and Control Prevention, which provided the EHIS 2009 data used in parts of this thesis. I should thank my Head of Department's Lia Tzala for all the assistance, the flexibility provided and her guidance throughout those years and mostly during the last months of this thesis' fulfillment. The presence of Fivos Anastasakis, my colleague, was of great importance. He inspired me to continue, at times when I was seriously thinking of changing the subject of this MSc Thesis.

Last but not least, I am mostly grateful to my family and Maria for their constant support and encouragement.

The fancy way, in which this Msc Thesis is organized, is due to the use of the free and rather powerful computer typesetting markup language named LaTeX. Acknowledge to the people involved in these license free project. Information about LaTeX as presented in Oetiker et al. (2010) is quoted. A small tribute to those that spent time putting things together :

> **LaTeX** *(Lamport 1994) is based on Donald E. Knuth's TeX typesetting language or certain extensions. TeX is pronounced "Tech," with a "ch: [...] The "ch" originates from the Greek alphabet where X is the letter "ch" or "chi" [...] LaTeX was first developed in 1985 by Leslie Lamport, and is now being maintained and developed by the LaTeX3 Project. LaTeX is a document preparation system for the TeX typesetting program. It offers programmable desktop publishing features and extensive facilities for automating most aspects of typesetting and desktop publishing, including numbering and cross-referencing, tables and figures, page layout, bibliographies, and much more. LaTeX was originally written in 1984 by Leslie Lamport at SRI International and has become the dominant method for using TeX; few people write in plain TeX anymore. The current version is LaTeX2e (styled LaTeX 2$_\varepsilon$). As it is distributed under the terms of the LaTeX Project Public License (LPPL), LaTeX is free software.*
>
> `http://www.latex-project.org/`

This thesis as a whole is based on free licensed software. We should not be depended on franchised and expensive programs that more often offer less than they cost. The **R statistical programming language** in cooperation with LaTeX provide an optimal, quite fast and flexible way to create reports. The whole practical section of chapters 2 - 5 - 6 - 7 is programmed in R code and by using the package **xtable** for exporting tables to LaTeX created an almost automatic way of producing reports. Moreover an automatic way of saving graphs from R to the Hard Disk Drive was created and then loading them directly in LaTeX as part of the document's code. Considering the above, the installation of last minutes changes in core parts of the MCMC and Variable selection sections and rerun of the whole analysis was easily feasible.

> R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R (...)
>
> R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.
>
> `http://www.r-project.org/`

The main software used for the implementation of the fifth chapter's example is OpenBUGS. A short overview of its history, current state and future development is presented here. More information can be easily be found on the program's website.

> BUGS is a software package for performing Bayesian inference Using Gibbs Sampling. The user specifies a statistical model, of (almost) arbitrary complexity, by simply stating the relationships between related variables. The software includes an 'expert system', which determines an appropriate MCMC (Markov chain Monte Carlo) scheme (based on the Gibbs sampler) for analyzing the specified model. The user then controls the execution of the scheme and is free to choose from a wide range of output types. (...)
>
> There are two main versions of BUGS, namely WinBUGS and OpenBUGS. Note that software exists to run OpenBUGS (and analyze its output) from within both R and SAS, amongst others[1]. (...)
>
> Initially, BUGS only used fairly specialized algorithms. In 1996, however, the project moved to Imperial College, London (headed by Nicky Best, who had already been involved for some years in Cambridge) and work began on expanding the software's capabilities. In particular, Jon Wakefield and Dave Lunn joined the project at this stage to work on implementing non-linear models, and development of a standalone Windows version of the software gained momentum. (...)
>
> (...)Now that OpenBUGS has progressed from being somewhat experimental to a stable and reliable package, we are now focusing all development efforts on it.
>
> `http://www.openbug.info/w/`

[1] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/remote14.shtml

Declaration

I declare that this thesis was composed by myself, that the work contained in here is my own except where clearly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Athens, 2013 _____

Konstantinos Pateras

Chapters' Description

Chapter 1 focuses on some of the main aspects of Bayesian theory. We attempt a brief introduction to basic notions concerning the prior and the posterior distribution which play an important role on Bayesian theory. In the end a short comment on Graph Theory is provided, while references for further reading are attached.

In Chapter 2 the reader is given a first idea of the history and theory of Markov Chains and how they are combined with Monte Carlo integration to create the Markov Chains Monte Carlo (MCMC). References for further reading are provided, analytical reviews of the methods can be found in several articles attached. Advantages and disadvantages of each method along with the algorithm for simulation are provided. In particular cases figures were created and attached for visual understanding of the steps of the procedures.

A practical example will be found in chapter 3, using data from the EHIS 2009 study, on a simple logistic regression problem, with the use of three of those algorithms (Independent Metropolis Hastings, random walk Metropolis and the Slice sampler). The example finishes with comparisons between the convergence, properties of the algorithms, posterior summaries and differences are pointed out between each MCMC method and each sampling scheme with the use of diagnostic plots and statistical tests. The last, though, are crudely used as no known theoretical implementation to Generalized Linear Models was found.

Chapter 4 contains a brief review of alternative classical and Bayesian methods for model choice, comparison and checking. We present a review of classical ways of Model selection by implementation of Stepwise procedures and the use of Information Criteria concerning non nested models, most of them based on computation and comparison of the models likelihoods. Simple model comparison with the use of Bayes Factor, short description of Lindley's Paradox and how the Marginal Likelihood is of great importance for latest proposed techniques to compare competitive models. While only a brief summary with appropriate references is provided concerning direct methods for estimating the marginal likelihood, a larger overview of variable

/ model variable selection techniques will be found in chapter 5 with examples comparing three of the techniques mentioned using real and simulated data.

In chapter 6 extra attention is given to the reversible jump Markov chain Monte Carlo (rjMCMC). Each chapter includes a quick note on latest techniques in model selection using MCMC with appropriate references for further reading. The chapter concludes with a short presentation on open research fields of this area of Bayesian Statistics. A practical implementation of reversible jump with Gibbs sampling on Bayesian Model Selection on Linear Models with application on data derived from Clyde et al. (2011) for later convenience and comparison is provided with the use of jumping interface in WinBUGS (Lunn et al. 2009), together with a comparison to the most flexible of the previously applied methods (GVS).

The last chapter (7) of this MSc thesis provides a short comparison of three well known and established R packages used for model comparison inside the Bayesian framework (BMA - BMS - BAS). An example using simulated data already been analyzed in the previous chapter will be presented, concerning linear regression model selection, while comparative tables both for replicability and efficiency will lead us towards the end of this Master thesis.

Computations, plots and tables presented in this thesis were carried out in the R programming language, version >3.0.0 (Masked Marvel) (R Development Core Team 2008) and OpenBUGS >3.2 (Lunn et al. 2000). The packages used and the summary of the code developed are briefly provided in the Appendix and will be included in the final thesis digital version.

Contents

1 Basic concepts of Bayesian Theory **1**

1.1 Historical Aspects 1

1.2 Introduction 2

1.2.1 Do they agree to disagree? 2

1.2.2 Bayes Theorem 4

1.2.3 Inference 4

1.3 Priors 5

1.3.1 Conjugate Priors 6

1.3.2 Low-Information Priors 8

1.3.3 Jeffreys' prior 9

1.4 Bayesian Inference 9

1.4.1 Point estimation - Credible sets 10

1.4.2 Predictive Distribution 11

1.5 Probabilistic Graphical Models- Bayesian Networks 12

1.6 Closing remarks 13

2 Introduction to MCMC **14**

2.1 Random Variable Generation 15

2.2 Markov Chains 16

2.3 Rejection & Importance sampling 17

2.4 Markov Chain Monte Carlo 18

2.4.1 Metropolis Hasting Algorithm 19

2.4.2 The Gibbs Sampler 20

2.4.3 The Metropolis within Gibbs Algorithm 22

2.4.4 Slice Sampler . . . 23
2.5 Convergence Diagnostics . . . 25
2.5.1 Monte Carlo Error . . . 25
2.5.2 Multiple Chains . . . 26
2.6 Conclusion . . . 26

3 Europe Health Interview Survey study – Greece 2009 27
3.1 Generalized Linear Models . . . 28
3.2 Self-assessed health Vs. Education . . . 30
3.2.1 Diagnostic Plots . . . 35
3.2.1.1 Example 1 - Independent Metropolis Hastings . . . 35
3.2.1.2 Example 2 - random walk Metropolis . . . 36
3.2.1.3 Example 3 - Slice sampler . . . 36
3.2.2 Diagnostic Tests . . . 40
3.2.3 Samplers Comparison for Simple Logistic Regression . . . 40
3.2.4 Other ways of diagnosing convergence . . . 42
3.2.4.1 Graphical analysis of contour plots . . . 42
3.2.5 Posterior Summaries - Interpretation . . . 42
3.3 Model Checking . . . 45
3.4 Closing Remarks . . . 47

4 Introduction to Model Selection 48
4.1 Classical Model Comparison . . . 48
4.1.1 Stepwise Procedures . . . 49
4.1.2 Information Criteria . . . 50
4.2 Bayesian Model Comparison . . . 50
4.2.1 Bayes Factor . . . 50
4.2.2 Marginal Likelihood . . . 53
4.3 Bayesian variable selection with direct methods . . . 54
4.3.1 Conclusion - Further reading . . . 55

5 Bayesian Variable/Model Selection Using MCMC 56

5.1 Bayesian Model Averaging . . . 57
5.2 Variable Selection Initial notions . . . 58
5.3 Zellner's g-prior and extensions . . . 59
5.4 Indicator variable selection algorithms . . . 61
5.4.1 Stochastic search variable selection (SSVS) . . . 61
5.4.2 Unconditional priors Gibbs sampler (KM) . . . 63
5.4.3 Gibbs Variable Selection (GVS) . . . 64
5.5 Model space search algorithms . . . 65
5.6 Latest Variable / Model selection algorithms . . . 67
5.7 Posterior model/variable selection inference . . . 67
5.8 GVS, SSVS and KM implementation in BUGS . . . 68
5.8.1 Simulated data . . . 69
5.8.2 EHIS 2009 data . . . 71
5.9 Closing remarks . . . 74

6 Reversible Jump MCMC 75
6.1 Introductory notions . . . 75
6.1.1 Comparison notes Reversible Jump Vs. Carlin and Chib . . . 78
6.1.2 Population-Based Reversible-Jump MCMC . . . 78
6.2 DAG for probabilistic models . . . 79
6.3 Jump Interface rjMCMC . . . 81
6.3.1 Clyde's Simulated Data Scheme . . . 82
6.4 Conclusion . . . 84

7 Bayesian Variable Selection in R 85
7.1 Raftery's et al. "Bayesian Model Averaging" . . . 85
7.2 Feldkircher's and Zeugner's "Bayesian Model Averaging Library" . . . 86
7.3 Clyde's Bayesian "Model Averaging using Bayesian Adaptive Sampling" . . . 86
7.4 General Comparisons . . . 88

8 Conclusion - Further Research 92

9 Appendix 93

9.1 Appendix A . . . 93
9.1.1 Notations . . . 93
9.2 Appendix B . . . 96
9.2.1 Abbreviations . . . 96
9.3 Appendix C . . . 99
9.3.1 R Packages . . . 99
9.3.1.1 MCMCpack package . . . 99
9.3.1.2 MCMCmnl . . . 99
9.3.2 CODA package . . . 101
9.3.2.1 CODA diagnostics . . . 101
9.3.2.2 CODA plots . . . 102
9.3.2.3 summary.mcmc . . . 103
9.3.2.4 plot.diagnostics . . . 104
9.3.3 xtable package . . . 104
9.3.3.1 xtableMCMCsummaries . . . 105
9.3.3.2 xtableMCMCdiagnostics . . . 105
9.3.4 Other R functions . . . 105
9.3.4.1 ContourPlots . . . 105
9.3.4.2 writeDatafileR . . . 106
9.3.4.3 erg.mean . . . 106
9.3.4.4 Extra tables/graphs . . . 107
9.3.5 Convergence Tests . . . 109
9.3.5.1 Geweke Diagnostic . . . 109
9.3.5.2 Heidelberger-Welch Diagnostic . . . 111
9.3.5.3 Raftery-Lewis Diagnostic . . . 112
9.3.5.4 The Effective Sample Size Diagnostic . . . 114
9.3.5.5 Autocorrelation Diagnostic . . . 115
9.3.5.6 Diagnostics synopsis . . . 116
9.4 Variable Selection Functions . . . 116
9.4.1 SSVS, KM & GVS . . . 116
9.4.2 reversible jump via OpenBUGS jump add-on . . . 120

9.4.3 BMA - BMS - BAS 121

Bibliography **122**

List of Figures

2.1 Example of Gibbs sampler parameter space exploration. 22
2.2 Example of Slice sampler with auxiliary variable (u) 24

3.1 Logit Regression Contour plot of Log Likelihood values of β_0 Vs. β_1, with confidence intervals of {0, 0.5, 0.75, 0.9, 0.95, 0.99} 33
3.2 Indepedent Metropolis Hastings - MCMC Diagnostic Plots 37
3.3 Random Walk Metropolis - MCMC Diagnostic Plots 38
3.4 Slice Sampler - MCMC Diagnostic Plots 39
3.5 Comparison of Indepedent Metropolis Hastings, Random Walk Metropolis and the Slice Sampler. Lines : {Black : IndMH - Red : RWM - Green : Slice} . . . 41
3.6 IndMH - Logit Regression Contour plot of the joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations 43
3.7 RWM - Logit Regression Contour plot of joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations 43
3.8 Slice sampler - Logit Regression Contour plot of joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations 44

6.1 Directed Acyclic Graph of model presented in equation 6.3 80
6.2 Basic hierarchical model for Jump's Interface reversible jump as a DAG. (Lunn et al. 2006) 80
6.3 Jump Interface model mixing representation, for the two chains initiated for the linear simulated regression predictors using Jump Interface in WinBUGS. . . . 83

9.1 Geweke-Brooks plot for the 3^{rd} sampling scheme of the RWM algorithm, showing how the Geweke diagnostic chooses and testes window A with window B and then widening window $A \rightarrow A^{'}$ to repeat the test. 110

9.2 Geweke-Brooks plot for the 3^{rd} sampling scheme of the RWM algorithm, showing what happens to Geweke's Z-score when repeatedly bigger number of iterations are being discarded from the beginning of the chain, the plot never discards more than half of the chain. 111

List of Tables

1.1 Conjugate priors and some of their characteristics for some common used distributions. 7

3.1 Table containing the most common Binomial Link Functions 30
3.2 Number of answers in each category of the "Self-reported Health Status" according to the education level of the responder (ISCED 0 - ISCED 6). 31
3.3 Summary of the changes in various quantities of examples 1.1 to 3.3 *Except for the Slice Sampler which acceptance rate is by definition equal to 1. 34
3.4 Posterior Summaries of the IndMH according to the scheme of iteration (*details in table 3.3). Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006). 45
3.5 Posterior Summaries of the random walk MH according to the scheme of iteration (details in table 3.3. Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006). 45
3.6 Posterior Summaries of the Slice Sampler according to the scheme of iteration (details in table 3.3. Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006). 46

4.1 Bayes Factor and its logarithm interpretation (Kass & Raftery 1995) 52
4.2 Bayes factor and twice its natural logarithm interpretation (Kass & Raftery 1995) 52

5.1 Model selection results of simulated data for SSVS, KM, GVS and BMA.# Occam's Window OR=500000, * UIP Empirical Bayes independent prior, **Laplace Approximation. Iterations=15000, Burnin Period = 5000 71

5.2 SSVS - KM - GVS - BMA variable selection indicators for binomial simulated data. #Occam's Window OR=500000. Iterations=15000, Burnin Period = 5000 72

5.3 SSVS - KM - GVS - BMA model selection results of EHIS 2009 dataset. Age = Age(in Years), Sex = Sex(Male/Female), Education=Please refer to table 9.4, Long-illness = Long Illness(Yes/No), Urban=Urban(Yes/No). #Occam's Window OR=500000, * Empirical Bayes independent prior, **Laplace Approximation. Iterations=15000, Burnin Period = 5000. 73

5.4 SSVS - KM - GVS - BMA variable selection indicators, Standard Deviations and MC errors for EHIS 2009 Data. #Occam's Window OR=500000. Iterations=15000, Burnin Period = 5000 73

6.1 Posterior inclusion probabilities of the linear simulated dataset predictors using the Jump Interface and GVS in WinBUGS . Iterations=15000, Burnin Period = 5000 83

6.2 Posterior model probabilities and corresponding Bayes Factors for the models with high probability versus the most probable model for the simulated data using Jump Interface and GVS in WinBUGS. Iterations=15000, Burnin Period = 5000 84

7.1 Characteristics of the packages and code to be used. BD= Birth - Death, RJ=Reversible Jump,HM=Hierarhical Mixture Model, LP= Leaps and Bounds Algorithm, UIP=Unit Information Prior, EUIP=UIP Empirical Bayes Independent Prior, BIC=Bayesian Information Criterion, AMCMC= Adaptive MCMC, EUIP*= considering constant prior variance between covariates. The number in parenthesis refer to the number of the programs alternative choices available. 88

7.2 BMA - BMS - BAS Posterior Inclusion Probabilities for linear simulated data, Param. Prior = Empirical or UIP (g-prior, with g=n), Model Prior = Uniform, considering full enumeration of the model space. PIP=Posterior inclusion Probabilities, P.Mean=Posterior Mean. *Predictors chosen from Stepwise method with AIC step. " x9 is correlated with x2 $\rho = 0.99$. 89

7.3 Posterior Inclusion Probabilities of the linear simulated dataset regressors under the set of summarized programs in R / WinBUGS. C. p-value corresponds to the p-value returned from the classical regression of the full model. *Predictors chosen from Stepwise method with AIC step in R, ** Classical p-value, " X9 is correlated with x2 $\rho = 0.99$. BMA -BMS - BAS consider full enumeration of the model space. Iterations = 20000, Burnin Period=20000. For more information see table 7.1 91

9.1 Running (System, user, elapsed) times of MCMCmnl of MCMCpack under all sampling schemes for 8000 iterations in seconds 107

9.2 Running (System, user, elapsed) times for WinBUGS (SSVS - KM - GVS - rjMCMC) and R (BMA - BMS - BAS) linear regression variable selection for Clyde's simulated dataset in seconds. R programs performed full enumeration, WinBUGS programs were measured for 20000 iterations. 107

9.3 Information on five frequently used Information Criteria placed according to date of development. (3.1.1) part corresponds to $-2logL(\hat{\theta}|y)$ 107

9.4 Number of answers in each category of the variable " How is your Health in general? "" according to sex of the responder (Male/Female) and the marginal distribution of each of the two variables. 108

9.5 Comparative table of Geweke Diagnostic for three algorithms used given a particular sampling scheme, see table 3.3, for windows sizes of $A = \frac{n}{5}, B = \frac{n}{2}$ 110

9.6 Comparative table of Heidelberger Diagnostic ($\beta(0)$)for the three algorithms used given a particular sampling scheme (see table 3.3) 112

9.7 Comparative table of Heidelberger Diagnostic ($\beta(1)$)for the three algorithms used given a particular sampling scheme (see table 3.3) 112

9.8 Comparative table of Raftery - Lewis Diagnostic ($\beta(0)$) for the three algorithms used given a particular sampling scheme (see table 3.3) 113
9.9 Comparative table of Raftery - Lewis Diagnostic ($\beta(1)$) for the three algorithms used given a particular sampling scheme (see table 3.3) 113
9.10 Comparative table of Effective Size Diagnostic for the three algorithms used given a particular sampling scheme (see table 3.3) 114
9.11 Comparative table of Autocorrelation Diagnostic ($\beta(0)$) for the three algorithms used given a particular sampling scheme (see table 3.3) 115
9.12 Comparative table of Autocorrelation Diagnostic ($\beta(1)$) for the three algorithms used given a particular sampling scheme (see table 3.3) , the brackets show the interval in which the autocorrelation has decreased in non significant values. . . 115

Chapter 1

Basic concepts of Bayesian Theory

1.1 Historical Aspects

Every allowed extension
of Aristotelian logic to plausibility theory
is isomorphic to Bayesian probability theory.
- *Unknown* -

During the 18th century a clerk and amateur mathematician named Thomas Bayes lived in England. It is estimated that Thomas Bayes was born around 1702 and died in 7th April 1761. Thomas Bayes is mostly known for having formulated a simple case of the Bayes theorem with computing a distribution for the probability parameter of a Binomial distribution.

Bayes work was presented as a solution to :

> Given the number of times in which an unknown event has happened and failed [...] the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named (Bayes & Price 1763).

There is great mystery associated with Thomas Bayes. The year of his birth is not exactly known, his portrait as presented across the scientific field is not that clear and last but more importantly, is he really the writer of the scientific publication attributed to him? According to

Stigler's historical research (Stigler 1986*b*) the writer of the essay, which lead the theorem to be named "Bayes Theorem", seems to be a Professor of Mathematics named Nicholas Saunderson born in January 1682, who lived in Cambridge and considered according to Stigler's research as the earliest discoverer of Bayes Theorem. The truth of that interpretation was questioned by Edwards (1986).

The general form of Bayes Theorem was presented by Laplace, a French mathematician and astronomer, introducing a considerably pioneer way of dealing with inference later called Bayesian Probability. According to Stigler's research (Stigler 1986*b*), Laplace is thought not to be familiar with the work of Bayes, due to the fact that they worked and lived in different locations and that in France the work of Thomas Bayes was not known.

In recent history, it was Harold Jeffreys, a mathematician, statistician and astronomer, who revived the Bayesian view of probability with his work "Theory of Probability". To conclude this short historical introduction, Jacob Bernoulli besides the introduction of the "Law of large numbers" theorem, in his book "Ars Conjectandi" also stated the issue of the reverse probabilities, almost fifty years earlier than Thomas Bayes, without being able though, to conclude to an equation (Bernoulli 1713).

1.2 Introduction

1.2.1 Do they agree to disagree?

Inside every non-Bayesian
there is a Bayesian
struggling to get out.
- Denis V. Lindley -

The principal difference between the two major statistical approaches - Frequentic and Bayesian - is that the one, which we are going to examine more thoroughly, consider parameters not as constants but as random variables characterized by a prior distribution. Therefore making inference while taking into account not only one value of the parameter (e.g. the one that

maximizes the likelihood). There lies one of the biggest advantage/disadvantage of the youngest brother of those two siblings.

Frequentists statisticians used to treat those in favor of the alternative approach as a minority until the late 80s. That was a general truth. The Bayesian approach, as examined later in this thesis, needed computer power to reach its real potential. Trouble appeared while trying to calculate the posterior distribution, which is needed for assumptions to be reached. Therefore, data analysis was out of reach, except for special cases where "Conjugate Priors" created a bypass for the analyst to reach inference, without the need of calculating the normalizing constant.

It was only after the introduction of MCMC (Markov Chain Monte Carlo) techniques, and the rapid progress of informatics and therefore personal computers, that Bayesian statistics started getting close to what they could truly offer to the statistical society, becoming a valuable tool to every research's hand.

Ideas derived from the Frequentic approach :

- The parameters of the population are unknown fixed constants.
- Statistical procedures have a long-term meaning, like an infinite repetition of the same experiment.
- Probabilities are interpreted as a frequency after a long number of experiments.

Ideas derived from the Bayesian approach :

- The parameters are now considered random variables, as we are not certain of the real values.
- The way to make inference is just the use of the rules of probabilities.
- Each person has his own way of thinking, so the prior beliefs naturally vary across people.
- There can be a continuous revision of our beliefs as data come to our hand.

That last two points of Bayesian statistics makes them even more related to real life situations and as a result a more sensible and natural way of quantifying problems.

1.2.2 Bayes Theorem

Two possible outcomes of a given situation are considered, A and B. Then, assume that $A = A_1 \cup ...\cup$An for which Ai $\bigcap Aj = \emptyset$ for every i $\neq$ j. Therefore, Bayes' theorem provides an expression for the conditional probability of A_i given B, which is equal to

$$P(Ai|B) = \frac{P(B|Ai)P(Ai)}{P(B)} = \frac{P(B|Ai)P(AI)}{\sum_{i=1}^{n} P(B|Ai)P(Ai)} \tag{1.1}$$

$$P(Ai|B) \propto P(B|Ai)P(Ai) \tag{1.2}$$

The last equation (1.2) is also called the Bayes' rule introduced by Piere Laplace, stating the proportionality of the two parts, considering $P(B)$ from equation 1.1 to be constant (Hoffmann-Jørgensen 1994).

One can rather easily notice that Bayes theorem is just a simple result, derived by the use of probability theory. Its introduction in statistical inference, created a juxtaposition, giving birth to two separated field in statistics, Bayesian statistics and frequentist statistics (Efron 1986, Berger 2000).

1.2.3 Inference

If we had to summarize in two points what is known as the Bayes solution : (1) We should specify the "prior" distribution $\pi(\theta)$, (2) and then use the formula to compute $\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}$ Let us consider a random sample $y = (y_1, \ldots, y_n)$, with $f(y_i|\theta)$ as the distribution function which describes the random variables $y_1, \ldots, y_n$. Therefore, the likelihood function is given by $f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$ which sets the probability of observing y_i under different values of the θ parameter. The Bayes Theorem incorporates the information already gathered, our prior beliefs for the parameter(s), represented by one or more prior distributions, then takes into account the observed data and make inference. For the case where θ is continuous, the following equation is obtained representing the posterior distribution

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)\, d\theta} \tag{1.3}$$

where $\int f(y|\theta)\pi(\theta)\,d\theta = f(y)$ is the marginal likelihood of the data, $\pi(\theta)$ the prior of parameter θ and $f(y|\theta)$ the likelihood of the data given θ.

For the discrete case the equation changes to $\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\sum f(\theta)f(y|\theta)\,d\theta}$, where $\sum \pi(\theta)f(y|\theta)\,d\theta = p(y)$ is the marginal likelihood of the data.

According to the Bayes theorem, the posterior distribution of the parameter given the data can be written as $\pi(\theta|y) \propto \pi(\theta)f(y|\theta)$. Given the above one can, with much ease, point out that the posterior distribution contains information both from our prior beliefs -Prior distribution $f(\theta)$- and the data being observed -Likelihood $f(y|\theta)$-.

Controlling the prior distribution can express the fact that we are very certain of our beliefs by setting the variance of the distribution to a low value, while we can express our prior ignorance by placing a large variance. This process is called *elicitation* of prior knowledge. We must point that in most of the cases in real life prior information is not available. There are various techniques for specifying our prior ignorance. Either way, the researcher has to take into account the cases in which the placing of an uninformative prior is improper - does not integrate to 1 - and therefore will cause computation troubles, especially if the concluding posterior is simultaneously an improper one, making proper inference rather impossible.

Another point that should make a researcher balance towards Bayesian statistics is that, even if the observed data are collected one after the other, he can infer at various points using the updating rule. The current posterior will play the role of the future prior, when new data is collected and processed. This is one of the most useful and important aspects of Bayesian thinking. The continuous update of the distribution when new information is at hand with the use of simple rules of probability.

The general form of the above statement can be expressed as follows for data $\pi(\theta|y^1, \ldots, y^t) \propto f(y^t|\theta)\pi(\theta|y^1, \ldots, y^{t-1})$ where y^t is the data collected in time t.

1.3 Priors

Within the Bayesian field, a prior distribution - prior deriving from the Latin adjective meaning earlier, first, e.t.c."- of an uncertain quantity, is the probability distribution that would

represent the beliefs of an individual, before he gets in contact or taking into account the collected data. Determination of the prior is the main point for criticism of Bayesian inference, an issue that will haunt us down until the very end of this thesis. The researcher should be extremely certain of why and when to choose a prior over another.

As already mentioned, one of the most important differences between Bayesian and Classical statistics is that in Bayesian statistics parameters are treated as random variables and for the inference to be made, a prior distribution has to be specified for these parameters. The parameters of the prior distribution itself are called *hyper-parameters*, so as to be easier to distinguish them from the others and they also can be given a prior called hyper-prior. A vast variety of prior "families" exists, the most important though, especially in the early stages of the Bayesian statistics history, were the conjugate priors due to special characteristics that made them easier to handle before the revolution in the computing industry. These different families of priors are briefly mentioned below : **Subjective priors** - the result of a person's opinion with expertise in the problem, **Informative priors** - a prior that represents the state of certain information you possess, **Conjugate priors** - an easily updateable prior, more often a member of the exponential family distribution (Morris 1983) (see section 1.3.1), **Non-informative (Low-Information) priors** - stating the fact that no prior knowledge exists. **Jeffrey's Prior** being a special case, that also has an important ability of *being invariant under reparameterization*, **Improper priors** - a distribution that does not have the properties of a proper distribution.

In the following subsections there will be a brief mention of Conjugate and Non-informative priors (Jeffrey's prior).

1.3.1 Conjugate Priors

One of the most intensive computational difficulties in Bayesian Inference arise when the normalizing constant in the denominator of the posterior density (eq : 1.3) has to be computed.

Even the simplest choice of priors $\pi(\theta)$ can lead to computational difficulties. That is why until the arise of the computational era, one should have been cautious in the definition of that prior distribution. A conjugate prior is a distribution that has identical algebraic form

as its posterior. As a result, these prior distributions return a posterior of the same family of distributions, without the need of any extreme analytical computation to take place. All distributions which are part of the exponential family have conjugate priors (Gelman et al. 2003). Some of those prior distributions are presented in the table 1.1.

The use of conjugate priors simplify the computations to reach to the posterior. Therefore, when a common distribution, $\pi(\theta|\phi)$, is used as a prior some choices are more advantageous for the calculation of the posterior than others. We can carefully choose that prior, being a member of a family of distributions which is conjugate to the likelihood, trying to achieve a posterior that belongs to the same distributional family with the prior, avoiding unnecessary inconvenient calculations.

Moreover, the use of a mixture of conjugate priors can be used if our beliefs cannot be expressed with a single conjugate prior, while maintaining the flexibility and the straightforward computation. A mixture of Normal priors will be used later on the parameters of a model, while dealing with variable selection.

For a more detailed version of table 1.1 (see Ntzoufras 2011, chap. 2). A review on conjugate priors can be found among others in Fink (1997).

Distribution	Likelihood	Prior distribution	Posterior parameters
Poisson	$Y_i \sim Poisson(\lambda)$	$\lambda \sim (\alpha, b)$	$\hat{\alpha} = n\bar{y} + \alpha,$ $\hat{b} = n + b$
Binomial	$Y_i \sim binomial(\lambda)$	$p \sim beta(\alpha, b)$	$\hat{\alpha} = \sum_{i=1}^{n} y_i + \alpha,$ $\hat{b} = \sum_{i=1}^{n} N_i + b$
Normal (known σ^2)	$Y_i \sim N(\lambda)$	$\mu\|\sigma^2 \sim N(\mu_0, \sigma_0^2)$	$\hat{\mu} = w\bar{y} + (1-w)\mu_0,$ $\hat{\sigma}^2 = w\sigma^2/n,$ $w = \sigma_0^2/(\sigma_0^2 + \sigma^2/n)$
Exponential family (ϕ known)	$Y_i \sim$ $expf(\theta, \phi, \alpha(), b(), c())$	$\alpha \sim Dirichlet(\alpha_0)$ $exp\{[\theta\theta_0 - \tau_0 b(\theta)]/\alpha(\phi)\}$	$\hat{\alpha} = \sum_{i=1}^{n} y_i + \alpha_0$ $\hat{\theta} = n\bar{y} + \theta_0,$ $\hat{\tau} = n + \tau_0$

Table 1.1: Conjugate priors and some of their characteristics for some common used distributions.

1.3.2 Low-Information Priors

In Bayesian analysis, there are times in which no prior information concerning the parameter is available to rely on or maybe the researcher will, to avoid wrongly applying subjective inference, let the data speak for themselves.

In such occasions the prior $\pi(\theta)$ has to contain no or more precisely low information about the parameter θ in the sense that no specific value should be promoted over the others. Such priors are called vague, low-informative or default priors.

In practice in some particular events we can create a vague low-information prior by making a distribution as flat as possible. This can easily be done by placing large values (e.g. to the variance of the Normal distribution), which will eventually help the distribution to expand and approximately be flat for the values we are interested for.

When we deal with a continuous parameter space , $\theta \in \Theta = [a, b]$ bounded by $-\infty < a < b < +\infty$ then a non-informative prior can be given by the uniform distribution (Bolstad 2007)

$$\pi(\theta) = \frac{1}{b-\alpha}, \quad \alpha < \theta < b \tag{1.4}$$

If the parameter is unbounded the form of a uniform prior is represented by the following form, $\pi(\theta) = c, \quad c > 0$.

The researcher should be rather careful when using such priors. If the prior distribution integrates to infinity and as a result is improper by definition ($\int \pi(\theta) d\theta = \infty$), awkward situations may arise and therefore, preventing the researcher to proceed with proper inference. Even in cases like the above, there is a way to end up with proper inference, that is by making the likelihood with respect to θ to integrate into a finite state.

The use of the uniform prior to state that no prior information is available and therefore can be considered a simple solution, the main disadvantage of such a choice is that the uniform distribution is variant to reparameterization, which means that $\pi(\theta)$ may give low or no information for θ but $\pi(\gamma)$ may contain information to γ, when $\gamma = g(\theta)$. The use of the Jeffrey's non-informative prior which is invariant to transformation is a general solution to these types of issues (Jeffreys 1946).

1.3.3 Jeffreys' prior

In general when using flat priors, we are unable to end up with posteriors in closed forms, therefore, the use of a set of techniques called Monte Carlo Markov Chains is inevitable.

Jeffreys' prior is a non-informative prior which is invariant to 1-1 transformations and has the form $\pi_o(\theta) = |I(\theta)|^{\frac{1}{2}}$, where I (θ) is the expected Fisher information matrix, with elements i, j. $I_{ij}(\theta) = -E_{X|\theta}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} log\ f(x|\theta)\right]$. Calculating $I(\theta)$ 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. It should also be noticed that Jeffreys' prior is not necessary a flat prior.

Moreover, Jeffreys' prior is consistent to the main principle in determining the prior, the fact that we should not take into account the data first, be being determined in such a way that makes use of the form of the likelihood and not the actual data. Even though Jeffreys' prior is objective as described above, for many parameterizations, it favors some values be giving them more weight than others, often ending up being informative.

1.4 Bayesian Inference

As we have already mentioned, all information about θ after analyzing the data can be represented by $\pi(\theta|y)$ the posterior distribution. The posterior density graph contains all the information about Bayesian statistical inference. The researcher can, each time, point out the location, dispersion and shape of the posterior and immediately conclude to what areas are more plausible than others.

However, due to the fact that such visualization becomes difficult under multi-parameter problems and also because a lot of researchers from other scientific fields, are familiar to the classical ways of results representation, we can easily, by using decision theory, estimate points and create credible regions, making the inference presentation more user friendly and easy to be understood. For models where no conjugate prior is available, modern computational techniques unfetter our hands and compute features like the mean, the mode, the standard deviation or the quantiles of the posterior distribution.

1.4.1 Point estimation - Credible sets

But how to make a choice under various certain conditions of uncertainty? This question is easily answered with the use of a loss function. Let Θ be a set of possible states of nature θ, and let $\alpha \in A$ be the set of actions available. Then define $l(\theta, \alpha)$ as the loss that the researcher has to pay by taking action α when the state of nature is θ (Berger, 1985). One of the rules to choose among A actions is to select an α that minimize that loss particular chosen function. As a consequence of the existence of many loss functions, each researcher will conclude in a different estimate according to the loss function that he selected.

Therefore, when a point estimation is required so as to reduce the dimension of the information to a single number, we can make use of Bayesian decision theory. However, a part of the statistical community believes that if we try to reduce the posterior to a number, we can easily be misguided concerning our final results. So other alternative measures should also be noted. The statistic will depend on the random sample, therefore it is a random variable, and its distribution is its sampling distribution.

> While it is easy to demonstrate examples for which there can be no satisfactory point estimate, yet the idea is very strong among people in general and some statistician in particular that there is a need for such a quantity [...] statements about uncertain quantities ought to be made in terms which reflect that uncertainty as nearly as possible (see Box n.d., pg.309-310).

Frequentist statistics emphasize on unbiased estimators because averaged over all possible random samples, an unbiased estimator gives the true value. In contrast, Bayesian statistics does not place any emphasis on being unbiased. *In fact, Bayesian estimators are usually biased* (Bolstad 2007).

From a Bayesian perspective, point estimation means that we would use a single statistic to summarize the posterior distribution. The most important number summarizing a distribution would be its location. The posterior mean or median would be two right candidates.

Measures derived directly from the posterior distributions provide information for both location and dispersion. As mentioned before adequate choices for location measures are either the

mean, the median or the mode, where appropriate quantities for measuring the dispersity are among others the variance, the standard deviation.

One can summarize beliefs over θ using a single statistic, providing a not that competent value for θ. The usage of such an approach is criticized negatively as it does not provide us with information concerning the uncertainty of θ. An even more complete analysis could be reached by presenting the uncertainty over θ. Therefore, is necessary alternative ways to represent the shape and dispersion of $\pi(\theta)$ to be used.

One of those alternative ways are the so called "Credible sets" which are somehow the Bayesian equivalent of the Frequentist confidence intervals. They are called credible sets or credible regions. By definition, a 100 x $(1-\alpha)\%$ credible set for θ is a subset C of Θ such that $1-\alpha \leq Pr(C|y) = \int_C \pi(\theta|y)d\theta$ where integration is replaced by the summation over discrete components of θ. In other words, represent direct probability statements of the likelihood of θ falling inside C while using credible sets.

"The probability that θ lies in C given the observed data y is at least $(1-\alpha)$."

this is in total contrast to the frequentist confidence interval statement, which clearly contain the necessity of repetition of the experiment a large number of times. During the end of only one experiment (which is the case, most of the times) we are in physical possession of only one group of data. Immediately the possibility of our computed C containing θ, using Frequentist credible intervals, will be either 1 or 0 (Carlin & Louis 2011).

It is most often preferable to create the Highest Posterior Density (HPD) credible set, defined as the set $C(x) = \theta \in \Theta : \pi(\theta|y) \geq \gamma$, where γ is the largest constant satisfying $P(C|y) \geq 1-\alpha$ which contains the "most likely" θ values (Carlin & Louis 2011).

1.4.2 Predictive Distribution

An other important ability of Bayesian statistical analysis is the fact that all inference about future observations comes easily only by using the posterior distribution $\pi(\theta|x) = \dfrac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$. The predictive distribution $f(y|x) = \int f(y|\theta)\pi(\theta|x)d\theta$ is the way to acquire information about new observations given the data observed, the proposed prior and the likelihood. The predictive

inference in Bayesian statistics comes from the posterior predictive distribution. Even though the most common and appropriate way to make decisions on model selection is the Bayes Factor, the predictive distribution can also play an important role for decisions on model selection, as we will notice later in this thesis.

The posterior density of θ also provides the information necessary to test hypotheses about θ. Hypothesis testing in Bayesian statistics highly depends on model selection. Posterior probabilities as well as Bayes Factor are crucial when one tries to perform model selection and hypothesis testing. Posterior predictive p-values are also used in Bayesian inference indicating the probability that replicated data is more extreme than the data observed (Rubin 1984). More details on Bayesian p-values in Gelman, Meng & Stern (1996), while more information on Bayes Factor can be found in Section 4.2.1.

1.5 Probabilistic Graphical Models- Bayesian Networks

The first paper in the graph theory's history is regarded to be "The Seven Bridges of Konigsberg", written by Leonhard Euler and published in 1736 (Euler 1736), proving, with the use of graph theory, that there can be no possible solution for a particular problem.

The way of denoting the structural conditional independence which took off only in the early 90's, except from the known conventional copulas, is graphical models. (see Pearl (1988) and Lauritzen & Spiegelhalter (1988)). A graphical model is a probabilistic model where the conditional independence structure between variables are denoted by a graph and as a result for better understanding and compact specification of full joint distributions. Those models are used generally in statistics, but more especially in Bayesian statistics and Machine Learning. When such a network structure is a directed acyclic graph (DAG), the graphical model indicates a factorization of the joint probability of the whole set of random variables. Belief Networks (Bayes' Networks) are one way to state the independence assumptions made in a distribution.

Probabilistic Graphical Models provide new insights into existing models, a new user-friendly framework for designing models making them easier to be efficiently implemented in existing

software. OpenBUGS a computer software for Bayesian analysis of complex statistical models using Gibbs sampling, a program that will be used for the examples presented in Chapter 5 and 6, uses that kind of graph based algorithms for computation using a platform called *D*oodle (Spiegelhalter et al. 2003). For a researcher with limited initial understanding of statistics is the best way for a quick introduction in more complex concepts.

Many examples (e.g. The famous Burglar - Alarm Network) and more insights can be found in Pearl (1988), while definitions and an introduction to Bayesian reasoning and more complex notions of this area of research in David Barber (2010)

1.6 Closing remarks

This chapter attempted a small introduction to the basic concepts of Bayesian Statistics, while a short comment with appropriate references were provided on Probabilistic Graphical Models. In the next chapter (2) we will provide a short introduction to the field of simulation Markov Chains and the use of Markov chain Monte Carlo, containing with an overview of basic techniques and notions. In chapter 3, the implementation of part of those techniques in data derived from the EHIS Greece 2009 study is provided in R. Further details on EHIS 2009 study can be found in Chapter 3.

Chapter 2

Introduction to MCMC

Where a calculator on the ENIAC is equipped
with 18,000 vacuum tubes and weighs 30 tons,
computers in the future may have only 1,000
vacuum tubes and perhaps weigh 1 to 2 tons
- Popular Mechanics, March 1949 -

The main reason for the widespread interest in MCMC methods is that these methods are extremely general and versatile and can be used to sample univariate and multivariate distributions when other methods (e.g. classical methods that produce independent and identically distributed draws) either fail or are difficult to be implemented. As we will notice, the fact that MCMC methods produce dependent draws causes no substantive complications in summarizing the target distribution. These methods have proved useful in all aspects of Bayesian inference, in the last two decades even on the computation of quantities used for comparing competing Bayesian models.

The posterior distribution is the only tool for the researcher to make inference using Bayesian Statistics. Having access to the posterior we can calculate any value and therefore produce and summarize results. Computation of moments, percentiles is always a process of summarizing posterior information with the use of the posterior distribution computed from complex integrals. The latter being a huge drawback, decelerating the application of Bayesian statistics, during the mids of the previous century.

During the first years of theoretical establishment of the Bayesian approach we were obliged to use conjugate priors for inference. Basic inference for generalized linear models was easy to be implemented, due to the fact that distributions of the exponential family, as we have seen before (table 1.1), always have a conjugate prior (Morris 1983). Before the computing evolution and the advance in the methods used for simulation, the large sample theory made the application of asymptotic methods available " Normal - Laplace approximation " (Tierney & Kadane 1986). Such methods were used to acquire analytic approximations of the posterior being subject to limitations concerning the sample size and the number of dimensions.

Recent developments in the area of computer hardware made the generation of random draws from the posterior feasible. The Monte Carlo integration is used for calculating any summary for which we are interested for. Monte-Carlo (MC) methods were introduced during World War II for nuclear physics calculations at Los Alamos National Labs, where the first computer (ENIAC[1]) was located. The first MCMC algorithm though, is associated with a second computer, called MANIAC[2], built in Los Alamos under the direction of Metropolis in early 1952 (Robert & Casella 2011).

The next sections present a brief overview of certain Markov chains and MCMC methods.

2.1 Random Variable Generation

Let assume that a way exist that our computer can provide as with IID random variables from a uniform distribution $U \sim (0, 1)$ and interpreting it as probability. The next methods describe how that IDD sample will be used to create random variables from any distribution under consideration (Normal, exponential, Binomial and others).

Let assume that a CDF exists of a distribution F that can describe a random variable X. We now want to generate values distributed according to F. By following the **inverse transformation method** one should attempt the next steps : (1) Generate a random number from $U \sim (0, 1)$, (2) Compute x while F(x)=u holds. Then x can be consider a random draw from distribution

[1]Electronic Numerical Integrator and Computer

[2]Mathematical Analyzer, Numerator, Integrator, and Computer

F. For details and proof of the above one can search (see Devroye 1986, Chapter 2) among others.

With the **accept reject method** we can generate Y distributed by f using this algorithm. Let assume that c is a constant such that $f(x) \leq cg(x)$. We then can (1) Generate $X \sim g$ and $U \sim U(0,1)$, (2) Then set Y=X if $U \leq \frac{f(X)}{cg(X)}$, (3) If not then repeat step 1. This method is also called Rejection Sampling and will be examined in section 2.3.

Details for another method called **Composition - Convolution** which is rather useful for generating from compound distributions can be found in L'Ecuyer (2012).

2.2 Markov Chains

A collection of random variables θ^t where $t \in T$ is called a stochastic process. T is called the parameter space. The values that such a process can generate are the state space. In a Markov Chain process the past and future states are independent given the present state, $P(\theta^{(t+1)}|\theta^{(0)}, \theta^{(1)}, ..., \theta^{(t)}) = P(\theta^{t+1}|\theta^t)$.

Such a process to be a Markov Chain should be (i) **Irreducible**, no matter where it starts has the ability to reach another state with positive probability, (ii) **Aperiodic**, if there are parts of the state space that can be visited at any spaced times. *P*eriodic if the above does not happens and parts of the state space can only be visited at certain spaced times and (iii) **Positive Recurrent**, if for all states I, if the process starts at i, it will return there with $\pi = 1$ and also the waiting time until it first returns is finite.

Moreover, a positive recurrent and aperiodic chain is called **ergodic**. While, if the chain is also irreducible then the **stationary distribution is unique**. By creating an ergodic and irreducible Markov Chain with stationary distribution you can, to some accuracy, calculate the mean, the standard deviation or the quantiles of π, simply by waiting until stationarity is established and then monitor for a long period applying MCMC diagnostics, then producing the needed posterior summaries.

2.3 Rejection & Importance sampling

Rejection sampling is a method for generating an independent and identically distributed (IID) sample from a distribution which is known up to a normalizing constant (Neumann 1963). Suppose that we desire to acquire a sample from $p(\theta|y)$ which is known up to a normalizing constant, $p^* \propto p$. Then if you can find a integrable envelope function $q(\theta|y)$ such that : (i) $Mq(\theta|y) \geq p(\theta|y) \, \forall \, \theta$ and (ii) $g(\theta|y)$ is chosen as a distribution easy to sample from.

Even though, considerable flexibility is permitted in the choice of envelope function, in cases where θ is a high dimensional vector, in practice is often difficult to figure out, how to make the Rejection sampling algorithm creates an IID sample. In that case, we can relax the assumptions of independence considering samples the form a Markov chain (Section 2.2).

A special case of rejection sampling is the Adaptive Rejection Sampling, a method that only works for log concave densities and was firstly introduced for application to Gibbs sampling for dealing with full conditionals that are algebraically messy but often log-concave (Gilks et al. 1995). The idea behind this technique is to create an upper envelope on $p(x)$ and use this in replacement of $Mq(x)$ in Rejection Sampling.

Let assume that we need to compute the expected value of $f(x)$ with respect to a probability distribution p(x). **Importance sampling** is the first useful algorithm which was introduced for computing the $\int f(x)p(x)dx$ due to the complexity of $p(x)$.

In importance sampling, to deal with this difficulty, another distribution $q(x)$ is introduced from which we will draw samples from. This distribution is called the sampling distribution. The basic notion of importance sampling is to draw from a distribution similar to $p(x)$ and then weight the result so as to correct the bias introduced by sampling from the $q(x)$ - the wrong distribution. Our actual p(x) or q(x) will often be unnormalized and we have to use the set of our samples from q(x) in order to estimate the normalization factor. The algorithm name rather misinform the reader of what it really does, importance approximation should be a more appropriate name. $q(x)$ should be picked such that the variance of $f(x)w(x)$ is minimum.

In importance sampling we can not acquire samples from p(x) directly, we can easily overcome this with a process called Sampling Importance Re-Sampling (Rubin 1987). Other newer techniques based on importance sampling, which firstly introduced weights, are the Adaptive Importance Sampling (AIS) (Cheng & Druzdzel 2000) and the Sequential Importance Resampling (SIR) (Gordon et al. 1993). The resampling stage of the first has no statistical advantage, and the latter can be problematic using a very small or very large step size. However, both of them are producing smaller errors and have vastly improve the speed given the precision we desire in comparison to the simple Importance Sampler.

2.4 Markov Chain Monte Carlo

The basis of the Monte Carlo approach, obtaining numerical approximations to posterior, is the Law of Large Numbers and therefore, the Central Limit Theorem. IID sampling creates estimates of true summaries of the posterior through Monte Carlo integration. As the number of iterations $m \leftarrow \infty$ these summaries become consistent.

The fundamental references for connecting the Markov chain theory and MCMC methods can be found among others in Nummelin (1984) and Meyn et al. (2009). So, if you can construct an ergodic and irreducible Markov Chain with the following three (3) desirable properties : (a) Should have the same state as θ, (b) Should be rather easy to simulate from and (c) Its stationary distribution is $p(\theta|y)$.

Due to the fact that $p(\theta|y)$ is usually known up to a normalizing constant, it is appropriate to appear in calculations only through rations of the following form $\frac{p(\theta|y)}{p(\theta^*|y)}$. Provided that you only do the monitoring of the Markov Chain after it has reached stationarity, the MCMC estimates are still consistent because of the ergodic theorem. Therefore the MCMC and the IID Monte Carlo approaches described earlier are equally valid.

Let assume that you have then obtained a sample from $p(\theta|y)$, then one can estimate the mean and the standard deviation by $\hat{\mu} = \bar{\theta} = \frac{\sum_{i=1}^{n} \theta_{(i)}}{n}, \quad SD = \sqrt{\frac{\sum_{i=1}^{n} (\theta_{(i)} - \bar{\theta})^2}{n-1}}$, compute quantiles and produce histograms that summarize posterior information.

2.4.1 Metropolis Hasting Algorithm

Metropolis Hastings (M-H) is a generalized version of the basic Metropolis Algorithm introduced by Hastings (1970). The M-H algorithm is based on the existence of a proposal/candidate distribution $q_t(\theta^{t-1}, \theta^t)$ which is part of a transition kernel. In the case of the metropolis algorithm is restricted in symmetric candidate distributions, $q(\theta^{'}|\theta^{(t)}) = q(\theta^{(t)}|\theta^{'})$. As we have already mentioned if the number of iterations is large (weak law of large numbers), iterations from the transition kernel converge to the equilibrium distribution. In Metropolis Algorithm we follow iteratively the next steps

- Initiate values for $\theta^{(0)}$.
  - Generate $\theta^{'}$ from a candidate distribution $q(\theta|\theta^{(t)})$.
  - Calculate $\alpha_M = min\left(1, \frac{p(\theta^{'}|y)q(\theta^{(t)}|\theta^{'})}{p(\theta^{(t)}|y)q(\theta^{'}|\theta^{(t)})}\right)$
  - Set $\theta^{(t+1)} = \theta^{'}$ with probability α_M otherwise set $\theta^{(t+1)} = \theta^{(t)}$.

A frequent proposal of candidate distribution is $q(\theta^{'}|\theta^{(t)}) = N(\bar{\theta}, S_\theta)$. Where S_θ is the covariance matrix which controls the convergence of the speed.

According to Gelman et al. (2003) a good candidate distribution should have the next properties : (a) Easy to sample from $q(\theta^{'}|\theta(t))$, $\forall\ \theta$, (b) Easy calculation of the probability of transition $\alpha = (\theta^{'}|\theta^{(t)})$, (c) Each move or jump should travel a considerable distance within the parameter space and (d) The jumps are not rejected frequently (in need of calibration of the tuning parameters).

One approach suggests to run an initiative chain, then obtain a crude estimation and then using it as the scale of the candidate distribution (Carlin & Louis 2011). In absence of general rules, the scale selection is a process to be calibrated. We change the scale so as to achieve an acceptance ration between (0.3 - 0.5) for univariate distributions (Gelman, Roberts & Gilks 1996).

Metropolis Hastings - Pros and Cons

(+) Can work with both discrete and continuous distributions.

(+) Able to perform in any dimensions $x \in \Re^p$.

(+) Easy to be implemented.

(–) Produce auto–correlated samples.

(–) The chain can be slow in mixing.

(–) The chain may get trapped in just a part of the distribution.

There are several extensions of the general Metropolis–Hasting (MH) algorithm. In the first Metropolis algorithm only symmetric proposals were considered $q(\theta^{'}|\theta) = q(\theta|\theta^{'})$ (Metropolis et al. 1953). The random-walk Metropolis (RWM) considers a even more special case where $q(\theta^{'}|\theta) = q(|\theta - \theta^{'}|)$, where every next step is dependent to the previous. Another special case of the MH algorithm is the independent sampler which proposal distribution does not depend on the previous state of the chain as in RWM.

The Independent Metropolis–Hastings (IndMH) and the RWM will be implemented in a chapter's 3 example and compared to the slice sampler introduced in section 2.4.4. In that way we will be seeing the difference between one algorithm that explores local information (RWM), one that ignores it (IndMH) and one introducing auxiliary variables to reach convergence.

2.4.2 The Gibbs Sampler

Geman and Geman introduced the Gibbs sampler for optimization in a discrete image processing problem without completion (Geman & Geman 1984). In this algorithm a component in each step is updated from the corresponding conditional posterior, so when θ has n dimensions we need n steps in each iteration. The acceptance ratio for performing a move in Gibbs sampler is equal to 1. As a result, every simulated quantity is accepted and thus the convergence assessment for the Gibbs sampler must be differently be treated from the MH techniques. From the conditional distributions, the researcher can with great ease, create the joint distribution, using the Hammersley - Clifford theorem, which indicate that with knowledge of conditional probabilities we are capable of reaching the joint probability (Hammersley 1971). Gibbs sampler works relatively well in missing data models and with convenience can perform really well with high-dimensional problems as well.

The following steps are used to simulate a Gibbs sample

- Initiate values for $\theta^{(0)}$.
- For $t = 1, ..., T$ the next actions take place.
  - Set $\theta = \theta^{(t-1)}$.
    - For $j = 1, ..., d$, update $\theta_j \sim f(\theta_j|\theta_{/j}, y)$
  - Set $\theta^{(t)}$ otherwise set $\theta^{(t+1)} = \theta^{(t)}$.

where $\theta_{/j}$ is the parameter vector θ without parameter j.

Then one can easily approximate the marginal distribution having samples from the whole set of full conditionals. Figure 2.1 represents who the algorithm explores the space in Gibbs sampling in two dimensions. One should notice that the steps are vertical and horizontal, which is anticipated as every parameter is individually updated. More information concerning the Gibbs sampler can be found in Gilks et al (1996), Robert and Casella (2004)

Gibbs Sampler - Pros and Cons

(+) easy to understand, easy to implement

(+) good trade off between acceptance and mixing $->$ acceptance ratio is always 1!

(+) open - source, black box implementations OpenBUGS - WinBUGS.

(–) the form of all univariate conditional distributions will not always be known. (see Section 2.4.3)

Implementation of Gibbs sampler - WinBUGS

One certain concern should be as to how to derive all the different univariate priors for any model under consideration with ease. The team at the MRC Biostatistics Unit (Cambridge), consisted be David Spiegelhalter, Gilks and others provided as with a free software package that implements the Gibbs sampler under a rather wide variety of situations. The software is called WinBUGS and stands for Bayesian inference using Gibbs Sampling. Using it, one simply

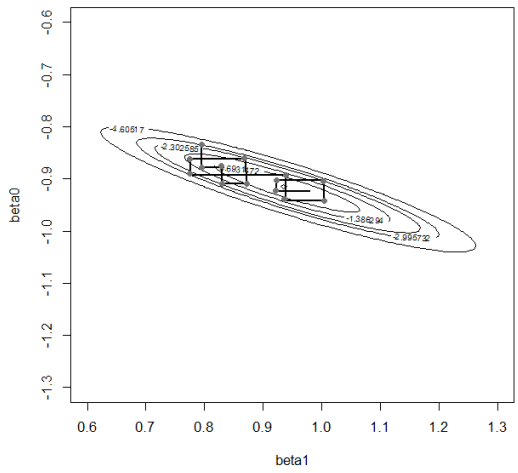


Figure 2.1: Example of Gibbs sampler parameter space exploration.

needs to make general considerations, while it computes all the required univariate marginal distributions (see e.g. (Thomas et al. 1992), (Spiegelhalter et al. 2003))

(Marginalize whenever you can!)

2.4.3 The Metropolis within Gibbs Algorithm

In the previous section we described how using the Gibbs sampler one can get the joint distribution of interest. The Gibbs sampler algorithm though, requires that the full conditionals of all unknown parameters are of know form and easy to generate from. In times that is not the case, as we do not know the exact form of part or the whole set of the full conditionals. It is often impossible to obtain the full conditionals of all the parameters. In such a case it is advantageous to make use of hybrid MCMC chains. The Metropolis-Hastings algorithm can be used to generate sample from the conditionals of a parameter of unobtainable form. In other words, we are making Metropolis-Hastings step within the framework of the Gibbs sampler.

The convergence of this algorithm is not very clear, as the MWG is more of a combination of algorithms, which each one of them alone would not converge. However, if the proposals used for each component are chosen in such a way that their property of irreducibility and

aperiodicity remains, each component will tend to its equilibrium and convergence will be reached (Carlin and Louis, 1996).

Metropolis Within Gibbs Sampler - Pros and Cons

(+) Can perform when no information on one / all full conditional distributions exists.

(+) In contrast to RWM is more efficient converging faster during the iterations.

(–) Autocorrelation again may be strong and as a result the mixing becomes really slow.

(–) While the number of parameters increases and the model becomes more complex the run time per iteration decreases.

The advantage of MWG over RWM is that it is more efficient with information per iteration, so convergence is faster in iterations. The disadvantages of MWG are that covariance is not included in proposals, and it is more time-consuming due to the evaluation of the model specification function for each parameter per iteration. As the number of parameters increases, and especially as model complexity increases, the run-time per iteration decreases for a given time interval. Since fewer iterations are completed in a given time-interval, the possible amount of thinning is also at a disadvantage.

2.4.4 Slice Sampler

All methods described above have some limitations when it comes to automatically construct a Markov chain sampler from ones model specification. Until now we have clearly stated that, to implement Gibbs sampling, one may need to devise methods for sampling from non-standard univariate distributions and to use the Metropolis based algorithms must find an appropriate "candidate" distribution that will eventually lead to efficient sampling. An alternate way to sample from $f(x)$, $p(x)$ was introduced by Neal (2003), overcoming the above difficulties.

The general approach is to sample from $f(x)$ from which sample is generally difficult to be generated. For that, one can specify u auxiliary variables and the conditional $f(u|x)$ to form the joint distribution $f(u, x) = f(u|x)f(x)$. Then with the use of a MCMC algorithm sample from (x, u) and marginalizing with computations over u to obtain samples from $f(x)$.

Even though the choosing of the auxiliary variables are usually a hard question to answer and mostly depends on each problem and its physical meaning, a usual choice of $f(u|x)$ is the uniform distribution $U(0, f(x))$.

The steps of the algorithm can be summarized as follows

- Generate $u \backsim f(u|x) = U(0, f(x))$
- Generate an interval $(x_l, x_r), x \in (x_l, x_r)$
  - Generate $x \backsim f(u|x)f(x) = U(x_l, x_r)$
  - If $f(x) > u$ stop the inner loop
  - If not make the interval (x_l, x_r) shorter.

A graphical representation of how the slice sampler operates can be found in figure 2.2. The gray lines depict the intervals that will be considered for each of the uniform univariate draws.

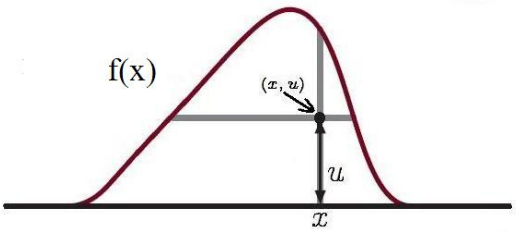


Figure 2.2: Example of Slice sampler with auxiliary variable (u)

Slice Sampler - Pros and Cons

(+) almost fully automated procedure - there's only one tuning parameter, w, that can be fixed in an adaptive tuning phase.

(+) easy to adapt if the variable has a bounded support.

(+) it usually can outperform the Metropolis - Hastings algorithm.

(+) avoids random walking.

(–) multimodality can be problematic - modes can be missed if the support is unbounded.

(–) univariate slice sampling can perform poorly when variables are highly correlated.

(–) the chain is often highly correlated due to expansion.

More details on the Slice sampling can be found in Neal (2003), extensions of the algorithm includes the Hyperrectangle slice sampling (Neal 2003, Thompson 2011) and the Reflective slice sampling, for sampling from multidimensional slices with the use of ballistic dynamics with specular reflections off the interior boundaries of the slice (Downs et al. 2000).

2.5 Convergence Diagnostics

There are numerous graphical and numerical proposals for measuring whether or not our chain has reached its equilibrium. A graphical way to monitor the convergence of the algorithm is to take into consideration plots of the iterations versus the generated values (Trace Plots). If all the values are within an area without strong periodicities and tendencies, then convergence may be assumed. The Running (Ergodic) Mean plots provide us with a way to check how well our chains are mixing. Another commonly used graph is the Autocorrelation plot. As it is widely known the MCMC output has positive autocorrelations which can be dealt with by keeping the first generated values in every batch of k iterations (sampling lag), resulting in the need to produce a larger output.

2.5.1 Monte Carlo Error

A way to monitor the variability of each estimate due to the simulation is the Monte Carlo Error (most commonly using the Batch Mean Method). For any $k(\theta)$ quantity that we are interested for, we partition the sample $\{\theta_1, ..., \theta_N\}$ into K batches B_b , $b = 1, ..., K$. The estimate of the Monte Carlo error of $\mathrm{K}(\hat{\theta})$ is given by the standard deviation of the batch means estimates $MCE = \sqrt{\frac{1}{K(K-1)} \sum_{b=1}^{K} [k(\bar{\theta})_b - k(\bar{\theta})]}$, where $k(\bar{\theta})$, the sample mean is given by $k(\bar{\theta}) = \frac{1}{K} \sum_{b=1}^{K} k(\bar{\theta}_b)$

The advantage of the Batch Means method is that is easy to be implemented. On the other hand, we require a large enough number of batches for the Central Limit Theorem to be

reasonable and also the size of each batch should be large enough so as the batch mean estimates are somehow independent. Taking the previous into account N might some times have to be very large. More details on the Batch Means method can be found in Carlin and Louis (2000, p172) among others.

Another method of estimating the standard deviation of the MCMC chain is called Window Estimator and estimates the variance directly. The main advantage of Windows estimate is that it does not rely on approximation theory to the same extent as the Batch Means method. However, more time and work intensiveness is required to compute the latter. More information on convergence diagnostics, used for the binomial case, are given in the example of the next chapter (see sections 9.3.5.1 to 9.3.5.5)

2.5.2 Multiple Chains

Lastly, without providing much details, we have to mention that one can run multiple chains with different starting points to reach convergence more securely, a detail review of similar methods and arguments concerning these techniques can be found in Brookst (1996). When the lines of those different chains cross in trace or in ergodic mean plots (see section 3.2.1), convergence is highly ensured. This technique is very common to graphically inspect whether or not our chain has reached convergence.

2.6 Conclusion

In this chapter we visited the early origins of MCMC Theory, dedicated a section for describing how the Markov Chains work and some simple but important for the time they were introduced algorithms. A connection with MCMC with notice to basic sampling scenes (Metropolis Hastings Algorithm, Gibbs Sampler and the Slice sampler) in detail with references for further reading. The next chapter present a practical implementation of IndMH, RWM and the Slice sampler while considering a simple binomial regression model using the MCMCpack package (Martin et al. 2011) in R statistical programming language.

Chapter 3

Europe Health Interview Survey study – Greece 2009

The European Health Interview Survey consists of health interviews which try to offer a comprehensive health status and other health-related factors of the countries in the European Union. For that to be accomplished a series of personal home-based interviews are being held. The European Health Interview Survey (EHIS), is organized and managed by Eurostat, is conducted and as a result contains information from all European Union (EU) Member Countries.

One of the primary objectives of these surveys is for the European Commission to strengthen its disease alert system. Every five years the surveys are to be repeated. Each of the Member States have the possibility to add information to be surveyed into their questionnaire. However, the main statistical methodology and the key variables to be measured have to be remained unchanged, mainly for comparison reasons. A health interview survey should cover the following topics among other

- Self-reported health status.
- Physical related data (Height / Weight).
- The decrease in activities that a person performed due to health issues.
- General health problems and chronic illnesses

- Smoke and alcohol consumption and other habits.

The EHIS 2009 - Greece survey covered around 6200 private households from the entire country, leading to an equal number of responders, given the condition that the surveyed persons are over 15 year old. A dataset coming from the health interview survey, which was conducted in Greece during 2009, will be used in this example of this chapter and for chapter's 5 real data example. The usage of those example will not only try to help readers understanding the theory presented throughout this thesis, but also report part of the results of the current survey, while performing comparative analysis. For an extensive descriptive analysis of the results please follow the Hellenic Statistical Authority which was mainly responsible to carry out the study in Greece [1].

The real data collected from the above study were extracted from the National Health Map Database for use in this MSc Thesis. Extra information concerning the dataset was provided after formal requests from the Health Map's Office based in the Hellenic Center of Disease and Control Prevention of Greece[2].

The National Health Map is a project seeking to collect all available health information including data from a variety of sources. Apart from raw information it collects data from both Public and Private Health entities in Greece. More information can be found on the Health Map's Website[3].

3.1 Generalized Linear Models

Generalized linear models are used for the analysis of both continuous and discrete response variables (McCullagh & Nelder 1989). In that way they may be considered as a generalization of normal linear regression models. The most common distributions can be used with these type of models as they are theoretically based on the exponential family of distributions (Normal, Binomial, Poisson, Gamma, e.t.c.). For details on some properties of the most common distributions of the exponential family see Table 1.1 in Chapter 1.

[1] http://www.statistics.gr/portal/page/portal/ESYE/BUCKET/A2103/Other/A2103_SHE22_MT_5Y_00_2009_00_2009_01_F_EN.pdf

[2] http://www.keelpno.gr/

[3] http://www.ygeianet.gov.gr/

The Generalized linear models (GLMs) have become popular because of their generality which leads to a vast range of application. They have provided a general way of dealing with the formulation of statistical models. They consist of certain components. (1) The stochastic component y_i, (2) the systematic component also called the linear predictor which consist of a linear function of explanatory variables and (3) the link function, which is the function that connects the response parameters with the explanatory variables.

The density function of the exponential family distributions is given by

$$f(y|\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{\phi + c(y,\phi)}\right) \tag{3.1}$$

where ϕ is the dispersion parameter and θ is the canonical parameter. The mean and the variance are given by

$$E(Y) = \frac{db(\theta)}{d\theta} = b^{'}(\theta) \tag{3.2}$$

$$V(Y) = \frac{d^2b(\theta)}{d\theta^2}\alpha(\phi) = b^{''}(\theta)\alpha(\phi) \tag{3.3}$$

The link function is of crucial meaning for GLMs. It is the way that the parameters of the response, are matched with the linear predictor and the covariates included in the model. The link function must every time be used wisely as it should map the range of values in which the parameter of interest lies with the set of real numbers in which the linear predictor takes values (Ntzoufras 2009). In our case the binomial, the link function should map the probability of success from [0,1] to the set of real numbers.

The most widely used link function of the binomial models is logit. In the next table, the most common link functions for the binomial model are presented.

Within the framework of Bayesian inference, βs in GLMs take zero values for prior mean and most usually large values for prior variance. In this way the researcher expresses his prior ignorance for the size of each effect.

Common Binomial Link Functions

Title	Link Function
$Logit\ (Canonical)$	$g(p) = (\frac{p}{1-p})$
$Probit$	$g(p) = \Phi^{-1}(p)$
$Complementary\ log-log$	$g(p) = log[-log(1-p)]$

Table 3.1: Table containing the most common Binomial Link Functions

More information concerning details and points of GLMs can be found in a plethora of books. (e.g. McCullagh and Nelder, 1989). Details concerning Bayesian inference on GLMs and recent developments can be found in Congdon (2006), while, for ways of analytical implementation using WinBUGS software one should look in Ntzoufras (2009).

3.2 Self-assessed health Vs. Education

Derived from the above study, one of the key questions is how a person judge his/her general health status. Many of the answers reported were above fair. In particular 67% of our sample claims to have good or very good health status.

Several studies have reported that many socioeconomic factors and especially the educational levels are related to the SRHS. Kunst et al. (2003) reported inequalities in "diseased weighted" self-reported health according to the level of education in 12 European countries of larger or smaller magnitude but still of significant substantial size. The findings in Montazeri (2008) indicated an inverse relationship between educational level and self reported health status. Following the above claims we will search if something similar can be reported in Greece's EHIS 2009 data.

In this chapter's naive example we are investigating if the education level of the responder plays an important role on how he/she judge his/her general health. Of course, we are not capable of pointing out a causal relation between those two variables due to various restrains. (a) Our study is a cross sectional one, presenting only a "photo" of the current population and (b) they may of course be other covariates that are simultaneously correlated with both

"Education status" and "Self-Reported health status" which if taken into account will diminish any relation found during this naive test (Secondary associations). A more complex model taking into account simultaneously other measured quantities is presented later on in this thesis in Chapter 5 while taking into account model uncertainty.

We assume that the levels of the variable "How is your health in general?", from now on called as "Self-Reported Health Status (SRHS)", are ordinal and one category differs from the one right before and right after it by a constant value of 1. So the values of the levels are Very Bad=1, Bad=2, Fair=3, Good=4, Very Good=5. For this example will only use the ones reporting bad health status=1 : (bad, very bad) and good health status=0 : (good, very good), creating a binary response variable. The analysis could be performed without reducing the levels of the response variable to 2 categories with the use of ordinal multinomial regression. For more information see e.g. Congdon (see 2007, Chapter 7).

Therefore, our sample size reduces to 4855 from 6172. The descriptive relation between the Education level and the "Self reported Health Status" can be noticed in table 3.2

What is your highest education degree?

ISCED Level	ISCED 0	ISCED 1	ISCED 2	ISCED 3	ISCED 4	ISCED 5	ISCED 6
Very Good - Good	159	828	565	1313	248	1017	10
Very Bad - Bad	214	323	71	72	7	28	0
Total	373	1151	636	1385	255	1045	10

Table 3.2: Number of answers in each category of the "Self-reported Health Status" according to the education level of the responder (ISCED 0 - ISCED 6).

The logit model will be used to model these data as it most commonly used in such type of data, while other frequently used link functions are the probit and the cloglog as we have already seen in table 3.1. The levels of SRHS will be represented by y and will take the place of our dependent variable, while the levels of education will be the explanatory variable x.

If the probability of those having reported a {good, very good} health status is p and n is the total number of persons, while the subscript i=1,...,7 represent the levels of self-observed Self-Reported Health, then $1 - p_i$ is the probability of having reported {bad, very bad} within

the i^{th} level and n_i the number of persons that have answered the i^{th} level as their level of education, $i \in \{1, 7\}$, then

$$logit(p_i) = b_0 + b_1 x_i$$
$$\text{and } y_i|p_i \sim Bin(n_i, p_i) \text{ , for i = 1, ... ,7.}$$

The logit model as we can see has as a dependent variable a transformation of p_i that represents the odds of the unknown probabilities. This element can be expanded as the probability of a person having reported {good, very good} health status p_i, over the probability an individual reporting {bad, very bad} health status $(1 - p_i)$.

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) \text{ for i = 1, ... ,7.}$$

We calculated the Maximum Likelihood Estimates of b_0 and b_1 along with their corresponding standard deviations, $\hat{b_0} = 0.9402(\frac{0.11}{\sqrt{4855}})$ and $\hat{b_1} = -0.9166(\frac{0.04}{\sqrt{4855}})$ using classical procedures. $e^{\hat{b}_1} = e^{-0.9166} \approx 0.40$ is the odds ratio of the i^{th} level versus the $(i+1)^{th}$, meaning that for every increase of the educational level by one level, an increase of $1/e^{\hat{b}_1} = 2.5$ on the changes of reporting a {good, very good} SRHS is reported. In figure 3.1 the contour plot of the log likelihood values of the above considered model is presented. We can see that parameter β_0 cover almost three times the interval of β_1, which points that correlation worthy to be mentioned between the two parameters should exist.

The MCMCpack (*see Appendix for details*) was used to implement the following examples. The function MCMCmnl which simulates from the posterior distribution of a multinomial logistic regression model - in our case a simple bivariate logistic regression - using either an Independent MH (IndMH), a random walk Metropolis algorithm (rwM) or a univariate slice sampler. The simulation part is done in compiled C++ code to maximize speed and efficiency.

Even though we mentioned earlier that the two parameters seems to be correlated, we adopt a multivariate Normal prior on β with independent components for convenience $\beta \backsim N(\beta_i, \tau_{\beta_i}^{-1})$

The Metropolis proposal distribution is centered at the current value of β, except stated otherwise and has variance-covariance matrix $V = T(\tau_{\beta_i} + \Sigma^{-1})^{-1}T$, where T is a matrix

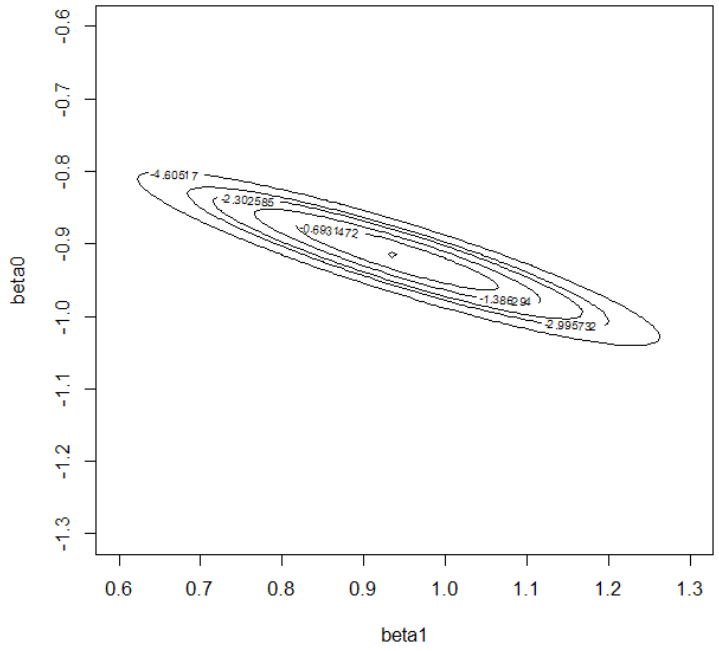


Figure 3.1: Logit Regression Contour plot of Log Likelihood values of β_0 Vs. β_1, with confidence intervals of {0, 0.5, 0.75, 0.9, 0.95, 0.99}

formed using the tune parameter, τ_{β_i} is the prior precision, and Σ is the large sample variance-covariance matrix of the MLEs.

In our example an independent bivariate Normal prior on $\beta = \begin{pmatrix} \beta_0' \\ \beta_1' \end{pmatrix}$ is assumed.

$$\begin{pmatrix} \beta_0' \\ \beta_1' \end{pmatrix} \sim Normal \left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} S_{\beta_0^2} & 0 \\ 0 & S_{\beta_1^2} \end{pmatrix} \right) \tag{3.4}$$

Therefore in each step, the proposed values of β' will be generated from $\beta_0' \sim N(\beta_0, S_{\beta_0^2})$ and from $\beta_1' \sim N(\beta_1, S_{\beta_1^2})$

The likelihood of the model will be

$$f(y|\beta_0, \beta_1) = \prod_{i=1}^{N} \left(\frac{\beta_0 + \beta_1 x_i}{1 + \beta_0 + \beta_1 x_i} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} \tag{3.5}$$

while the posterior distribution

$$\begin{aligned} f(\beta_0, \beta_1|y) &\propto f(y|\beta_0, \beta_1) f(\beta_0, \beta_1) \\ &\propto \prod_{i=1}^{N} \left(\frac{\beta_0 + \beta_1 x_i}{1 + \beta_0 + \beta_1 x_i} \right)^{y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{1-y_i} \\ &- exp\left(- \left(\frac{\beta_0 - \mu_{\beta_0}}{\sqrt{2}\sigma_{\beta_0}} \right)^2 - \left(\frac{\beta_1 - \mu_{\beta_1}}{\sqrt{2}\sigma_{\beta_1}} \right)^2 \right) \end{aligned} \tag{3.6}$$

In our examples $i = 1, 2$, while the unknown parameters are β_0 the constant and β_1 the effect of the model. In most cases, any prior distribution can be used to represent prior information. The choice of the prior may include informative priors, if previous knowledge of our data is available. In our case we take a "non - informative" prior indicating our possession of no clear prior information with two different starting points. In the last case though, a prior with small variance and mean far away from the MLE estimates to test how well the algorithm work in conditions where wrong a priori information is available. The information of the examples that follow are summarized in table 3.3. The calibration of the RWM acceptance rate should be even more careful, because high acceptance rates have the tendency to ruin the sampling procedure. The acceptance rates were stabilized each time around 0.30 using the procedures tuning parameter T, except for the Slice Sampler in which by definition the acceptance rate equals to 1.

Quantities of MCMC examples initial runs

Set	Initial Beta	Prior Mean	Prior Precision	Acceptance Rate	Burn-in	Iterations
1_{st} & 2_{nd} set	(MLE) / (10,10)	(0,0)	$\begin{pmatrix} 0.001 & 0 \\ 0 & 0.001 \end{pmatrix}$	≈ 0.30*	3000	5000
3_{rd} set	(10,10)	(10,10)	$\begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$	≈ 0.30*	3000	5000

Table 3.3: Summary of the changes in various quantities of examples 1.1 to 3.3 *Except for the Slice Sampler which acceptance rate is by definition equal to 1.

Diagnostic test are to be discussed and at the same time applied, while a function to summarize the posterior information into a graph, containing graphs mostly produced from the coda package was created and used in the following examples.

3.2.1 Diagnostic Plots

In order to check if the chain has reached convergence and other properties, a series of plots is available and will be plotted. The **traceplot** follows the iterations and plots sampled values for each variable, while **the running mean (ergodic mean) plot**, generates a time series plot of the "running" mean for each parameter. The "running" mean is the mean of all sampled values up to a given iteration, including that of the last iteration. **The density plot**, creates the probability density function from the simulated data, while the **autocorrelation plot** is a tool for checking if randomness exists in our simulated data. This property is ascertained by evaluating autocorrelations for data values at a variety of time lags.

From now on when noticing a particular sampling scheme with number $\in (1, 2, 3)$ we will be referring to table's 3.3 first column's numbers, describing each of our sampling schemes. To distinguish the use of sampling schemes among different MCMC algorithms, the names of each algorithm will be mentioned as IndMH (Independent Metropolis Hastings), RWM (random walk Metropolis) and Slice (Slice sampler).

3.2.1.1 Example 1 - Independent Metropolis Hastings

IndMH algorithms are of certain interest, but as we have discussed their practical implementation is often problematic. Both the construction and the choice of the proposal are crucial and often complicates the procedure. In the IndMH algorithm as we already have mentioned every next draw is independent from the previous. The step is based on a bivariate normal proposal with mean equal to the posterior mode.

In the following section the IndMH example's plot diagnostics are presented. The 2_{nd} sampling scheme converged without difficulties. A small autocorrelation issue was noticed on the last scheme. The IndMH reached convergence for the set of three schemes, with some difficulties noticed for the 3_{rd} sampling scheme. The lag in the first two schemes was limited to 5 and for the last scheme was raised to 10 for the autocorrelation plot. The diagnostic tests showed that the IndMH has issues when wrong information is known and as a fact a not so appropriate prior is selected, in comparison to selecting a non informative prior with prior mean either equal to zero or derived from the MLE estimators.

3.2.1.2 Example 2 - random walk Metropolis

While the IndMH algorithm applies mostly in specific problems, RWM can be applied to a wider set of cases. In the RWM algorithm every next draw is dependent to the previous. So the step is based on a bivariate normal proposal with mean equal to the previous step. Even though random walk has the ability to deal with a lot of different situations it is not always the most efficient solution. Difficulties such as regions with low probability between modal regions require many iterations to be explored and due to its symmetric features lot of time is spent revisiting regions that are have already been explored. Alternatives that are not that easy to implemented exist, that overcome difficulties using perfect symmetry (see Robert & Casella 2009, chapter 6).

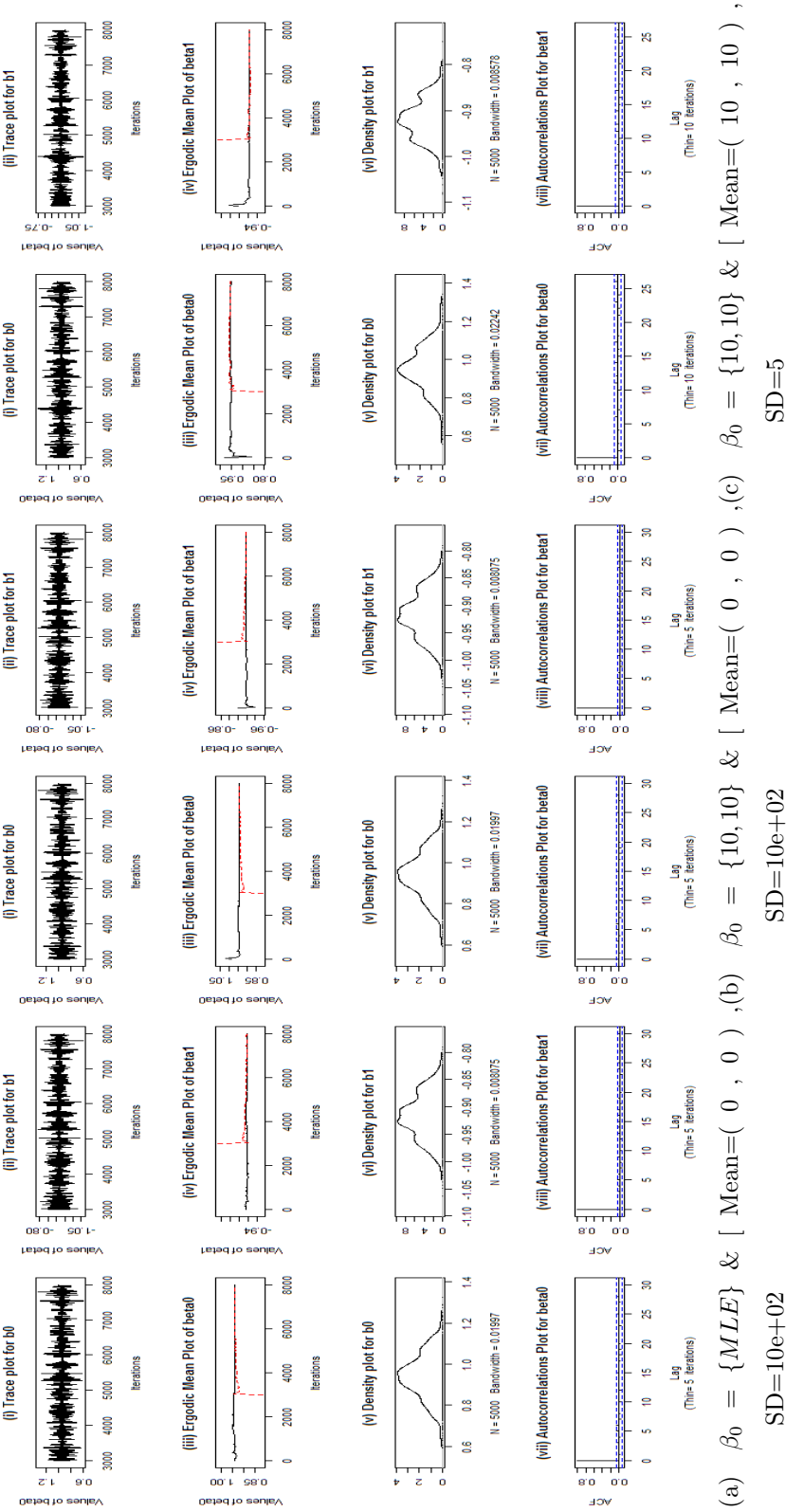
As it was anticipated the autocorrelations of the chains in comparison to the IndMH are higher in all sampling sets. More iterations seems to be needed for convergence to be reached in comparison to the IndMH. In particular, in the 2_{nd} & 3_{rd} set, the algorithm had not reached convergence even after 1500 iterations.

3.2.1.3 Example 3 - Slice sampler

The Slice sampler uses latent variables to sample from the target distribution. That feature is what makes the sampling schemes under the Slice sampler to need more time to execute given the number of iterations. In the slice sampler the proposals are generated by a truncated bivariate normal distribution. Details on how the slice sampler is implemented for Generalized Linear Models can be found in Damlen et al. (1999)

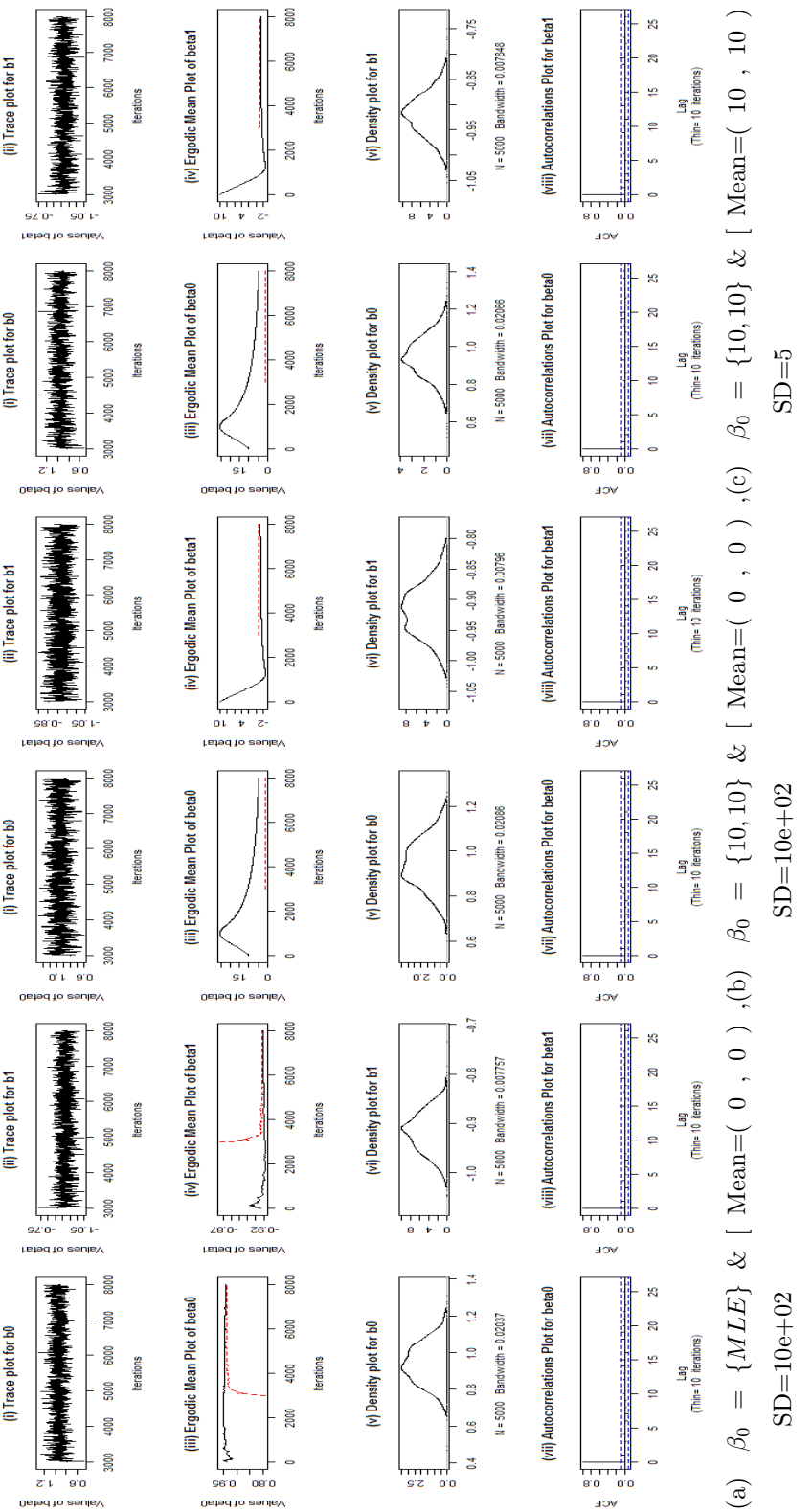
In our examples the Slice sampler performed really well and with only just a few iterations reached convergence in the sets of schemes, it seems to performs similar to the IndMH but not outperforming it.

As we have noticed in section 2.4.4 the sample would, by construction, be high correlated due to expansion. The slice sampler seems to have the biggest autocorrelations of all examples presented in this chapter. Comparisons between algorithms and between their sampling schemes will be presented in sections 3.2.2 to 3.2.3



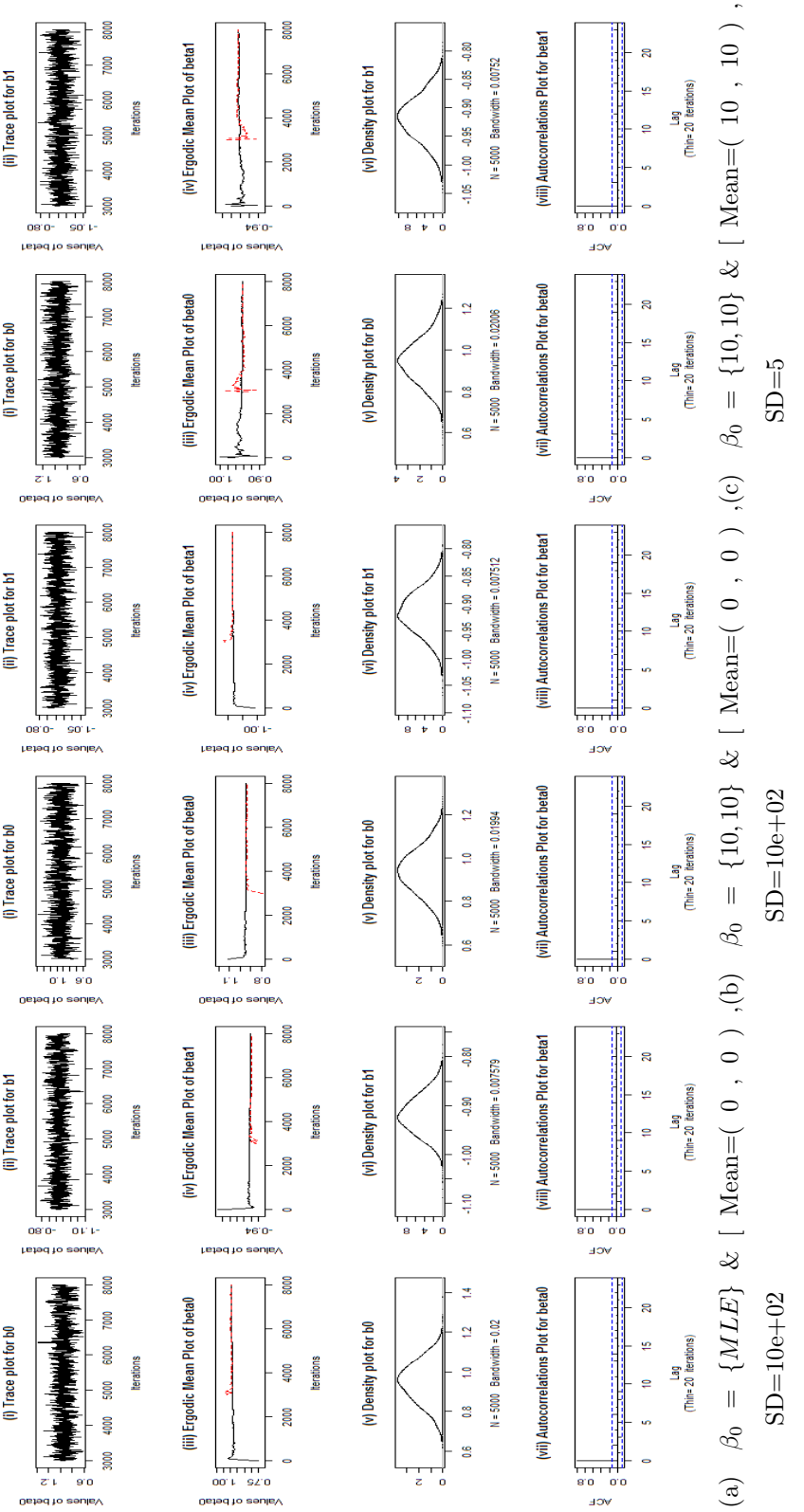
(a) $\beta_0 = \{MLE\}$ & [Mean=(0 , 0) ,(b) $\beta_0 = \{10, 10\}$ & [Mean=(0 , 0) ,(c) $\beta_0 = \{10, 10\}$ & [Mean=(10 , 10) , SD=10e+02 SD=10e+02 SD=5

Figure 3.2: Indepedent Metropolis Hastings - MCMC Diagnostic Plots



(a) $\beta_0 = \{MLE\}$ & [Mean=(0 , 0) ,(b) $\beta_0 = \{10, 10\}$ & [Mean=(0 , 0) ,(c) $\beta_0 = \{10, 10\}$ & [Mean=(10 , 10) , SD=10e+02 SD=10e+02 SD=5

Figure 3.3: Random Walk Metropolis - MCMC Diagnostic Plots



(a) $\beta_0 = \{MLE\}$ & [Mean=(0 , 0) ,(b) $\beta_0 = \{10, 10\}$ & [Mean=(0 , 0) ,(c) $\beta_0 = \{10, 10\}$ & [Mean=(10 , 10) , SD=10e+02 SD=10e+02 SD=5

Figure 3.4: Slice Sampler - MCMC Diagnostic Plots

3.2.2 Diagnostic Tests

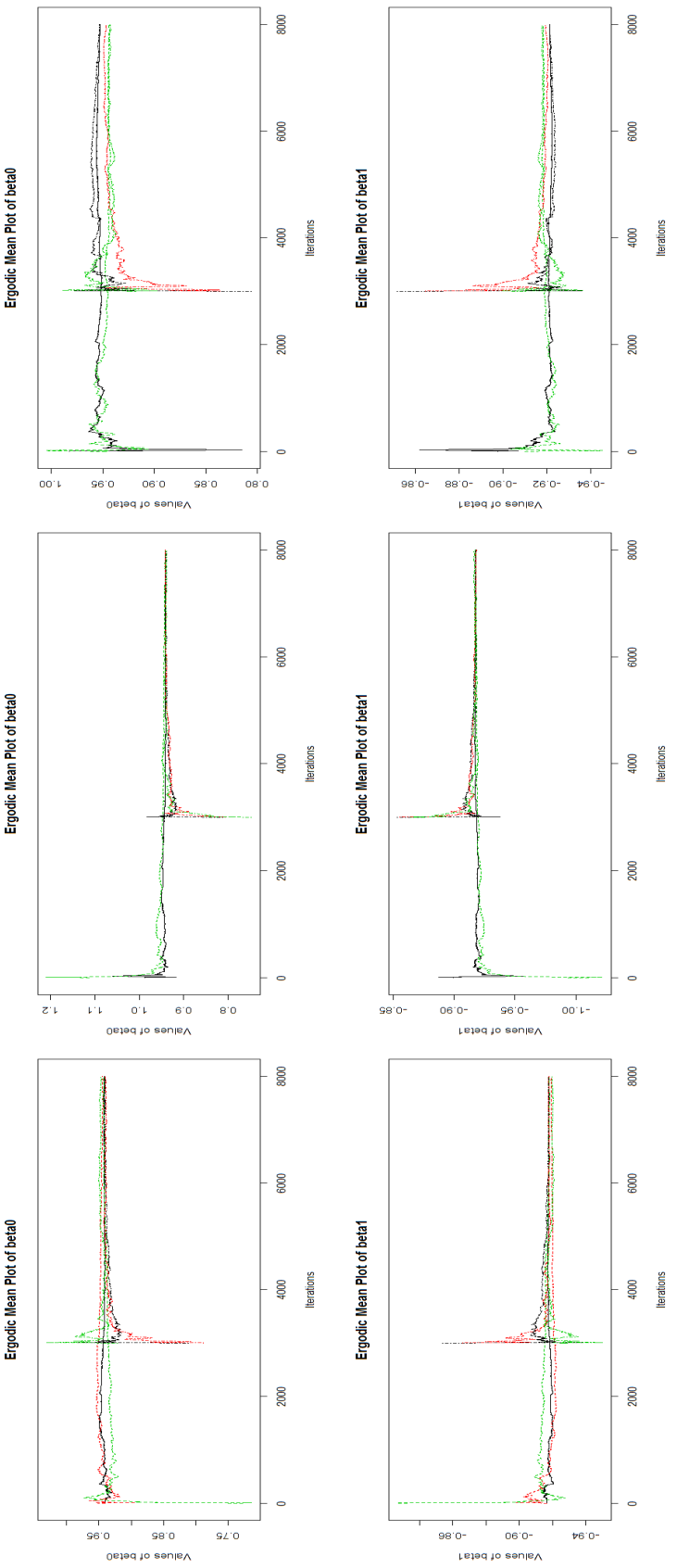
Diagnostic tests were carried out, with the majority of them pointing to similar results. The data used in the following diagnostics are the iterations of each sampling scheme with the initial burn-in sample included, therefore $N = 8000$.

The whole set of sampling schemes created chains that converged. The RWM seems to have the most difficulties in the 2^{nd} and the 3^{rd} sampling schemes, but eventually reaching convergence. As a whole, the slice sampler seems to be working faster (needing the least iterations) than RWM and slightly better than IndMH, even when taking into account the autocorrelation introduced by the slice sampler into the produced MCMC sample. In the 3^{rd} sampling scheme the IndMH seems not to have reached its equilibrium and more iterations are suggested by the Heidelberger-Welch diagnostic.

For more details on the application of the tests with the coda package refer to the Appendix's section 9.3.5 on the functions used, see coda's functions and parameters placed in the appendix. For a comparative study on diagnostic tests one can refer to Cowles & Carlin (1996).

3.2.3 Samplers Comparison for Simple Logistic Regression

Further on, plots presenting the ergodic mean of each algorithm given the sampling scheme (1 - 3) are provided. With starting points, both from the beginning of the chain and after discarding the sample of the burn-in period (T=3000). The two random walk's schemes 2^{nd} and 3^{rd}, for which we assumed distant initial betas, seem not to be working well. For that reason, those schemes were excluded from the plotting of the whole chain and only plotted after discarding the burn-in period in figures 3.5b and 3.5c. This is because the first 1.500 iterations of the RWM wandered away and only after a considered number of iterations the chain converged to its equilibrium. The last figure (3.5c) of this section includes the ergodic mean of the slow converging RWM algorithm for the 2^{nd} sampling scheme, we can clearly see that the algorithm's property of exploring local regions is not very helpful, leading the algorithm to slower reach its equilibrium.



(a) $\beta_0 = \{MLE\}$ & [Mean=(0,0), SD =(b) $\beta_0 = \{10, 10\}$ & [Mean=(0,0), SD = (c) $\beta_0 = \{10, 10\}$ & [Mean=(10,10), SD = 5 10e+02 10e+02

Figure 3.5: Comparison of Indepedent Metropolis Hastings, Random Walk Metropolis and the Slice Sampler. Lines : {Black : IndMH - Red : RWM - Green : Slice}

3.2.4 Other ways of diagnosing convergence

3.2.4.1 Graphical analysis of contour plots

A presentation of the 3^{rd} sampling scheme is attempted in figures 3.6, 3.7 and 3.8, where the prior mean=(10,10), the prior Standard Deviation = 5 and the $\beta_{initials} = \{10, 10\}$. The plots describe how each one of the three algorithms (IndMH, RWM, Slice sampler) move around the space discovering the joint distribution. Of the three sampling schemes considered only the 3^{rd} will be presented, as we noticed that all algorithms had difficulties to reach convergence given this scheme.

Here we can notice again that the IndMH explores the posterior distributions with much ease and achieves to produce values from the whole space much quicker that the other two algorithms, though having difficulty exploring the tails of the distribution. The RWM seems to be having the worst exploration, one can notice that after 50 iterations the mode has been visited very few of a times but after one thousand iterations seems to be sampling better that the IndMH. The Slice sampler behaves well throughout the sampling process, but needs the most running time for the same number of iterations to be attempted (see table 9.1 in the Appendix)

3.2.5 Posterior Summaries - Interpretation

After checking the diagnostic plots and the diagnostic tests, we concluded that the IndMH and the Slice sampler seem to converge faster in the whole set of sampling schemes while the RWM is slower for the 2^{nd} and the 3^{rd}. In that way we have to use a smaller burn-in period for those two, as we are in a way throwing away valuable sample that would lower the MCMC error in a given number of iterations. The burn-in period for all sets of posterior summaries is considered to be 3000.

We will present the posterior summaries of each MCMC algorithm sampling scheme in three separate tables, one for each of the algorithms. The set corresponds to the number of the sampling scheme, while the parameter to each of the two parameters estimated for my simple logistic model.

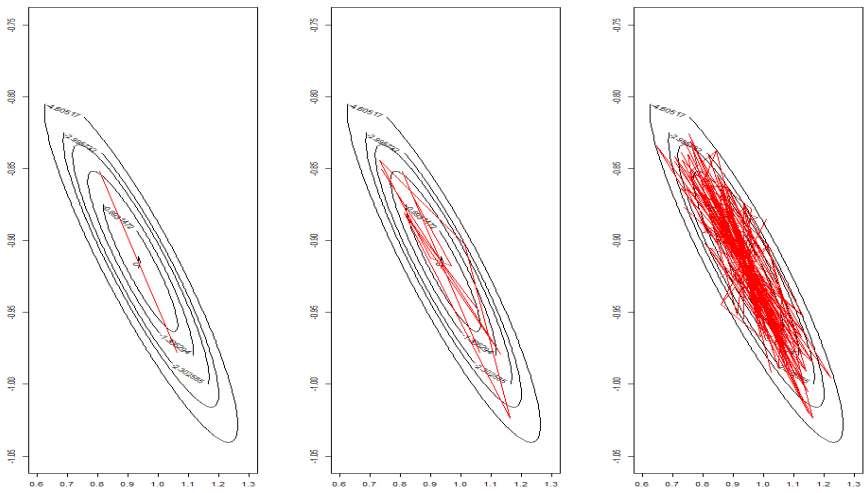


Figure 3.6: IndMH - Logit Regression Contour plot of the joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations

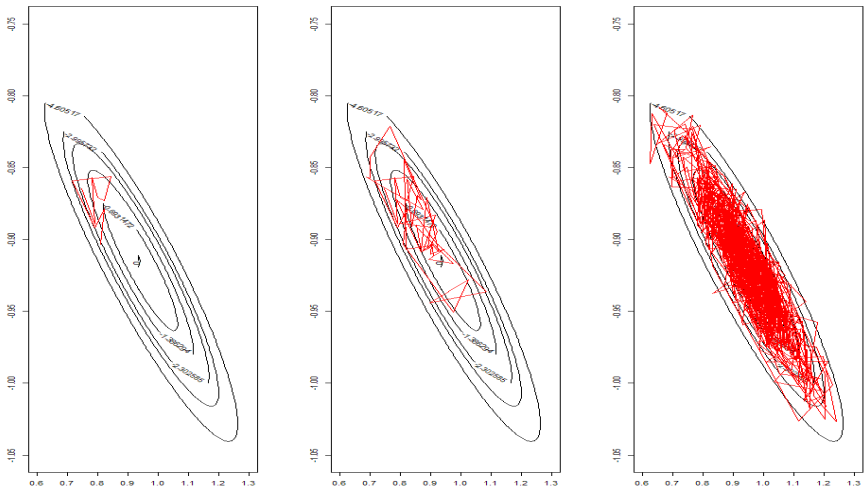


Figure 3.7: RWM - Logit Regression Contour plot of joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations

Using the standard method of MLE we computed $\beta_0 = 0.9402$ with SD equal to 0.0015 and $\beta_1 = -0.9166$ with SD equal to 0.0006. The estimates are really close to these values, this may be due to the size of our dataset's sample $N = (4855)$, that diminishes any prior knowledge

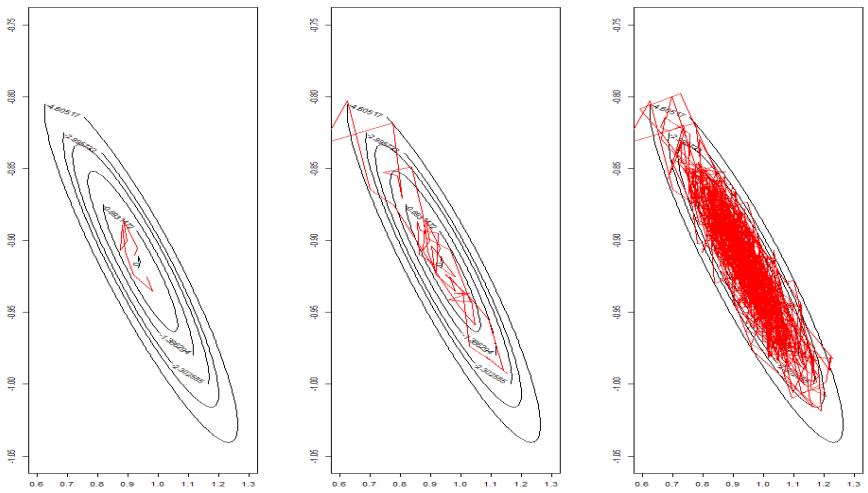


Figure 3.8: Slice sampler - Logit Regression Contour plot of joint Log Likelihood, $\{\beta_0, \beta_1\}$ for 10, 50, 1000 Iterations

provided. The 1^{st} and the 2^{nd} sampling schemes provide closer estimates to the MLE in comparison to the 3^{rd} only for the IndMH and the RWM algorithms. We should notice that for the Slice sampler the 2^{nd} sampling scheme performed better in the way that it returned means closer to the MLEs.

As for the MC error it is clear that given the separate algorithms and constant burn-in period equal to 3000, for the IndMH the MC error takes lower values under the 1^{st} and higher under the 2^{nd} scheme for both parameters, while for the RWM algorithm it is not clear which of the sampling schemes has lower/higher MC errors. As far as the Slice sampler is concerned even though the 2^{nd} scheme seems to have the most accurate mean estimates, it also has the higher MC error among the different sampling schemes of the algorithm. In all cases though the MC error is well below the threshold of **three significant digit accuracy** ($MC\ error < 0.01$).

For the running times of each sampling scheme under the three algorithms used in MCMCpack's function MCMCmnl see the Appendix (table 9.1).

Parameter	Set*	Mean	SD	Naive SE	Time-series SE	2.5%	97.5%
	1^{st}	0.9408	0.1120	0.0016	0.0024	0.7338	1.1413
β_0	2^{nd}	0.9405	0.1055	0.0015	0.0040	0.7434	1.1401
	3^{rd}	0.9535	0.1194	0.0017	0.0035	0.7331	1.1871
	1^{st}	-0.9175	0.0418	0.0006	0.0009	-0.9972	-0.8401
β_1	2^{nd}	-0.9180	0.0402	0.0006	0.0015	-0.9956	-0.8452
	3^{rd}	-0.9214	0.0444	0.0006	0.0012	-1.0052	-0.8402

Table 3.4: Posterior Summaries of the IndMH according to the scheme of iteration (*details in table 3.3). Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006).

Parameter	Set	Mean	SD	Naive SE	Time-series SE	2.5%	97.5%
	1^{st}	0.9405	0.1055	0.0015	0.0040	0.7434	1.1401
β_0	2^{nd}	0.9424	0.1081	0.0015	0.0038	0.7330	1.1567
	3^{rd}	0.9468	0.1071	0.0015	0.0042	0.7306	1.1548
	1^{st}	-0.9180	0.0402	0.0006	0.0015	-0.9956	-0.8452
β_1	2^{nd}	-0.9183	0.0412	0.0006	0.0015	-0.9963	-0.8341
	3^{rd}	-0.9194	0.0408	0.0006	0.0015	-0.9992	-0.8369

Table 3.5: Posterior Summaries of the random walk MH according to the scheme of iteration (details in table 3.3. Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006).

3.3 Model Checking

There are a number of techniques used when the chain fails to converge. The researcher can change the priors carefully but avoiding the chance of truncating. Moreover, an increase of the

Parameter	Set	Mean	SD	Naive SE	Time-series SE	2.5%	97.5%
	1st	0.9471	0.1037	0.0015	0.0049	0.7376	1.1378
β_0	2nd	0.9387	0.1033	0.0015	0.0058	0.7390	1.1427
	3rd	0.9431	0.1039	0.0015	0.0046	0.7480	1.1541
	1st	-0.9198	0.0393	0.0006	0.0019	-0.9933	-0.8436
β_1	2nd	-0.9165	0.0389	0.0006	0.0021	-0.9933	-0.8403
	3rd	-0.9178	0.0390	0.0006	0.0018	-0.9953	-0.8429

Table 3.6: Posterior Summaries of the Slice Sampler according to the scheme of iteration (details in table 3.3. Contains the mean, the standard deviation, a naive s.e. of the mean (ignoring autocorrelation of the chain) and a time-series s.e. based on an estimate of the spectral density at 0 (Plummer et al. 2006).

number of chains should be considered, the alteration of parameters are another way to tune the algorithm so as to reach convergence faster. More information can be found in Gilks et al. (1996), Cowles & Carlin (1996).

Caution

Even after checking histograms, traceplots, autocorrelation plots and the convergence diagnostics there is no guarantee that a chain is stationary.

A limitation of the techniques used in this chapter is that they are not capable of identifying multimodal distributions and draw sample. One way to make sure a bigger part of the space is explored is by using multiple chains with different starting points. A lot of proposals are presented for dealing with an unknown change point analysis in logistic regression see e.g. Gössl & Kuechenhoff (2001). Lastly, a menu-driven program in R obtaining convergence diagnostics and other functions for MCMC outputs is called BOA (Smith 2005) and is suggested as an alternative of coda package mostly used throughout this chapter.

3.4 Closing Remarks

Part of chapter's 2 mentioned algorithms were presented in practice using data from the EHIS 2009 study. Comparisons between the techniques and advantages/disadvantages of each one was presented in theory as well as in practice.

The next chapter (4) is devoted to an introduction in variable and model selection techniques when question over the model itself is considered. Details for Model comparison both for the Classical and the Bayesian approach are included. A comparison between three presented techniques using MCMC is performed in WinBUGS and thoroughly explained in chapter 5 with the use of an Empirical Bayes dependent prior. This brief review on Bayesian variable selection methods is based on O'Hara & Sillanpää (2009), Ntzoufras (1999, 2011), Congdon (2007) and other sources. In chapter 6 we discuss the Reversible Jump MCMC introduced by Green in 1995 and we present the Jump Interface created by Lunn et al for WinBUGS in 2006 (Lunn et al. 2006), for practical use of the reversible jump MCMC for variable selection using a hierarhical model with Gibbs sampling.

Chapter 4

Introduction to Model Selection

All models are wrong,
but some are useful.
- *George Box* -

When we have a number of possible predictors it can be difficult to find the "best" model. Questions of the type, which covariates should be included in the model as main effects and which interactions to take into account, arise. Model selection attempts to make this task easier for the researcher. For such a procedure to be created one needs, a criterion to be able to compare two models and to create a strategy for comparison.

4.1 Classical Model Comparison

In that way model selection attempts to find this subset of predictors (covariates) that explain part of the response variability in the "best" way. Any unnecessary predictors should be excluded, as they will only add noise in how quantities are being estimated. Classical ways to deal with variable selection have been developed both using significant tests (e.g. in GLMs using the F and X^2 distribution) and alternative methods mostly using model selection criteria.

4.1.1 Stepwise Procedures

Methods that became widely used after their initial development of Efroymson in the early 70's are the stepwise methods (Efroymson 1960). The initial idea concerned implementation of the procedure in multiple linear regression. It is a fact that stepwise methods contain a plethora of different strategies e.g. forward and backward selection, stepwise forward and backward selection.

Problems arise when considering the limitation that are introduced by the use of such tests during model selection. In large datasets, p-values tend to be pretty small, and the continuous calculation of sequential significant tests introduces Type I and Type II error, including not important covariates or dropping out of the model significant ones. However, the main issue with the usage of such techniques is the selection of just one model between a huge number of possible models, producing effects considering only the selected model, while at the same time ignoring model ascertainty. In that way, selection of one model may insert bias in our inference.

Another measure to assess the fit of a model, the C_p statistic, which can be also used as criterion to stop the above stepwise procedures, was introduced by Mallow's in 1973 (Mallows 1973). Other popular model selection criteria, choose the model with the minimum values of MSE or looking at types of R^2 coefficients of determination. Among the different types of R^2 the most widely used is the $R^2_{adj.}$ a measure that takes account the number of variables contained in each model, while also maintaining the ability to be used for comparison among models of unequal size.

Instead of using the simple formula for calculating R^2 we can compute it using cross validation or bootstrap techniques. There are many cross-validation type techniques. A summary of different cross validation schemes is described briefly in (Syed 2011). Austin (2008) claims that bootstrap model selection appears to have similar performance compared to backward selection.

Finally, the level of efficiency of such methods has been summarized by Copas (1984) in the next quote "Stepwise methods are frequently used, frequently abused and poorly understood procedure of applied statistics".

4.1.2 Information Criteria

Model selection criteria contains a family Information Criteria who are based on the idea of maximizing the likelihood. The IC family can be written in the most general form as $IC = -2logL(\hat{\theta}|y) + d(p,n)$. This equation contains the negative logarithm of the likelihood and a function of mostly (p) the parameters of the models and (n) the number of observations. The values of IC are used to compare model of different size. They try to balance between penalizing for extra parameters adding complexity in the model and goodness of fit. A great number of different IC were introduced taking into account relaxed initial assumptions or just trying to boost its initial properties.

In Appendix's table 9.3 some Information Criteria are summarized, with both the equation for their computation and the date they were published. Among the ones we have chosen to cite, all have the same elements in the first half (likelihood part) but differ on how they weight the information in the second part. In other IC that is not always the case. AIC tries to find the true model from where the data were generated asymptotically. TIC (Takeuchi 1976) can be considered a more general AIC, where it is often hard to estimate $J(\hat{\theta})$ and $I(\hat{\theta})$, therefore becomes more difficult to be applied. AIC and BIC (Schwarz 1978) have a lot in common, however the second one is more strict whenever the model becomes more complex with the introduction of extra covariates. Hanna-quinn (Hannan & Quinn 1979) is a special case, which while often cited, seems to have seen little appraisal in practice as stated by Burnham & Anderson (2002). AICc (Hurvich & Tsai 1989) is a corrected form of AIC, which tries to work out the problem of selecting overfitted models when the number of observations is small compared to the number of regressors.

4.2 Bayesian Model Comparison

4.2.1 Bayes Factor

The formal Bayesian model choice procedure rests on work by Jeffreys (1961). As there is no constrain in the number of simultaneous hypotheses testings nor do any need to be nested

within any of the others, from now on we change the notation of "hypotheses" H_o to "models" $M_i, i = 1, ..., m$ (Carlin & Louis 1997).

Comparison between models, involves comparison of marginal likelihoods. The marginal likelihood is the probability of the data y given a particular model m, and is obtained by taking the average over the priors assigned to the parameters of m. The comparison of two competing models is based on the ratio of marginal likelihoods, or Bayes factor of model M_1 against model M_2. The introduction of that ratio diminishes the normalization constant, as it is ruled out by the devision.

Let us assume that we only have two competing models in the world M_1 and M_2, θ_0 and θ_1 the parameters corresponding to each of one of the models. Each model has a prior $p(M_i)$ with i =1,2 and $p(M_1) + p(M_2) = 1$. Applying the Bayes theorem (section 1.2.2) the posterior probability of any of the models is given by

$$p(M_i|y) = \frac{p(y|M_1)p(M_1)}{p(y|M_1)p(M_1) + p(y|M_2)p(M_2)}, \tag{4.1}$$

for i =1,2.

The posterior odds PO_{12} of model M_1 versus model M_2 are given by

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_2)}{p(y|M_1)} \frac{p(M_2)}{p(M_1)}, \tag{4.2}$$

The fraction BF_{12} is called Bayes Factor of model M_1 versus model M_2. $\frac{p(M_1)}{p(M_2)}$ is the prior odds of model M_1 versus model M_2. In other words (4.2) can be rewritten as

$$Posterior\ Odds = Bayes\ Factor\ X\ Prior\ Odds \tag{4.3}$$

$p(y|M_i)$ is the marginal likelihood of the data given a specific model M_i and is given by

$$p(y|M_i) = \int_{\theta_1} p(y|\theta_i, M_i)p(\theta_i|M_i)d\theta_i \tag{4.4}$$

The above comparison of two models can be extended to more. Approximate values for interpreting the Bayes factors provided by Kass & Raftery (1995) and are given in Tables 4.1 and

4.2. For values of BF<1, evidence for supporting the other model exist. Taking twice the log of the Bayes factor returns the same scale as the conventionally used deviance and likelihood ratio statistics (see Congdon 2007, page 27). Even though those are consider to be the most appropriate interpretations, one should be very careful when applying them in practice. The logarithms of the marginal probability of the data can also be used as a predictive score.

$Log_{10}\ BF_{10}$	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 4.1: Bayes Factor and its logarithm interpretation (Kass & Raftery 1995)

$2Ln\ BF_{10}$	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	Decisive

Table 4.2: Bayes factor and twice its natural logarithm interpretation (Kass & Raftery 1995)

Lindley-Bartlett Paradox

In 1957 Lindley (1957) noticed a rather interesting behavior of the Bayes Factor and called it a paradox. When $n \to \infty$ then the posterior odds $\to \infty$ for any given significance level, leading to the support of the simpler hypothesis. In classical statistics, significance tests, when n is very large, tend to reject the null hypothesis. Therefore, according to what methodology is selected for drawing inference, the researcher will end up "correctly" supporting different hypotheses.

Bartlett then observed that for variance of great magnitude the posteriors odds support the null hypotheses. Given this finding, one should carefully select priors with large variance and

the use of improper priors should be avoided. More specifically this paradox had lead to many suggestions for priors on model selection, some of which, will be examined further on.

4.2.2 Marginal Likelihood

Bayesian Model assessment with the use of Bayes Factor, posterior model probabilities and odds can be implemented with the computation of the marginal likelihood

$$f(y|m) = \int f(y|\theta_m, m)\pi(\theta_m|m)d\theta_m \tag{4.5}$$

while m is a model under assessment, $\pi(\theta_m|m)$ is the the density of the parameter vector θ_m for model m. From now on, in this section, for convenience we will omit the term m from the presented equations, except where necessary.

It is a fact that for the computation of the marginal likelihood high dimensional integrals are getting involved making most of the times the analytical approximation unachievable.

The predictive density can be an expectation of the likelihood with respect to the prior

$$f(y) = E_p[f(y|\theta)] \tag{4.6}$$

In Bayesian thinking, marginal likelihoods have a way to impose a natural penalty on more complicated models. In comparison to the likelihood, which will continuously increase with the addition of extra parameters, the marginal likelihood, while reaching a top, then starts decreasing while the complexity of the models increase. Is therefore, immune to overfitting issues.

The methods described in secton 4.3 utilize MCMC outputs from separate models in order to acquire the estimates of their marginal likelihoods and as a result estimate Bayes Factors. Most of the times we are not capable to derive this integration analytically, therefore we have to estimate of the marginal likelihood making use of other approaches such as asymptotic based methods (e.g. Laplace method) or with the use of simulation via MCMC. Only a brief mention of these methods which use iterative MCMC schemes to estimate the marginal likelihood will be attempted as they are not this MSc thesis main theme. More information can be found in Kass & Raftery (1995) and Gamerman & Lopes (see 2006, chapter 7).

4.3 Bayesian variable selection with direct methods

A direct estimate of the marginal likelihood, using Monte Carlo is given by $\hat{f}_0(y) = \frac{1}{K}\sum_{j=1}^{N} f(y|\theta^{(j)})$ where $\theta^{(1)}, \ldots, \theta^{(N)}$ is a prior distribution's $f(\theta)$ sample. This approach is rather simple and has a lot of drawbacks, including the fact that the estimator does not work well when disagreement between the likelihood and the prior exists (McCulloch & Rossi 1991). Even if N is large enough, this simple direct estimate will be influenced by a few sampled values, making it rather unstable.

Laplace approximation (Tierney & Kadane 1986) is based on the Normal distribution and results in $\hat{f}_1(y) \approx (2\pi)^{d/2}|\tilde{\Sigma}|^{1/2}p(y|\tilde{\theta})p(\tilde{\theta})$ where $\tilde{\theta}$ is the posterior mode of the parameters of model, $\tilde{\Sigma}$ equals to the minus of the second derivative matrix of the $log(\theta|y)$ evaluated at $\tilde{\theta}$ One can avoid to compute analytically $\tilde{\Sigma}$ and $\tilde{\theta}$ and estimate them with use of the output of a MCMC algorithm using the point that maximizes $p(y|\theta)p(\theta)$ as an estimate of $\tilde{\theta}$ (posterior mean) and the variance-covariance matrix of the generated values (**Metropolized Laplace Estimator** (Lewis & Raftery 1997).

Another way is to use importance sampling, described in section 2.3, aiming to boost values being sampled from regions where the integrand is large. These approaches are based on generating sample from the importance density $g(\theta)$. The predictive density given by 4.5 can be rewritten as $f(y) = E_g[\frac{f(y|\theta)p(\theta)}{g(\theta)}]$. Therefore the **importance sampler estimator** (Newton & Raftery 1994) can be given by $\hat{f}_2(y) = \frac{1}{K}\sum_{j=1}^{N}\frac{f(y|\theta^{(j)})p(\theta^{(j)})}{g(\theta^{(j)})}$ where now $\theta^{(1)}, \ldots, \theta^{(N)}$ are sampled from $g(\theta)$.

When $g(\theta)$ is the posterior $\pi(\theta)$ the importance sampling estimator is called **Harmonic mean (HM) estimator**. The HM estimator, even though is rather simple to use, a problem that cannot be ignored is the fact that it is affected by small likelihood values (Lopes & West 2004).

The Newton - Raftery (NR) estimator is a way to combine both $\hat{f}_1$ and the HM estimator in such a way that $g(\theta) = \delta p(\theta) + (1-\delta)\pi(\theta)$. The NR estimator uses a mixture of the prior and the posterior, with weight equal to δ where $0 < \delta < 1$.

Bridge sampling estimator Meng & Wong (1996) claimed that for any bridge function $\alpha(\theta)$ with support deriving from both the posterior density π and the proposal density g, ratios of normalizing constants can be estimated by$f(y) = \frac{E_g\{\alpha(\theta))p(\theta)p(y|\theta)\}}{E_\pi\{\alpha(\theta)g(\theta)\}}$ It is interesting to see that if $\alpha(\theta) = \frac{1}{g(\theta)}$ reduces to the simple MC estimator $\hat{f}_0(y)$ and in the same way if $\alpha(\theta) = \{p(\theta)p(y|\theta)g(\theta)\}^{-1}$ the Bridge sampler becomes just a variation of the HM estimator . Details of the iterative scheme to estimate the predictive density can be found in Meng & Wong (1996).

Another common method to calculate the marginal likelihood was introduced by Chib (1995) called the **Chib's estimator** . For the estimation of the marginal likelihood $\hat{\pi}(y|m)$ presented as $\hat{\pi}(y)$, an estimate of the predictive density $\hat{\pi}(\theta^{'}|y)$ at $\theta^{'}$ is required. When knowledge of the full conditional distributions is available, Chib's proposed a Gibbs iterative scheme to be applied in order to calculate $\pi(\theta^{'}|y)$. The posterior distribution can be written as

4.3.1 Conclusion - Further reading

Techniques of classical statistics were mentioned at the beginning, while model comparison with analytical computation of the marginal likelihood or with the use of MCMC algorithms (Chib & Jeliazkov 2001) were briefly discussed.

A lot of extensions of the above methods have been proposed e.g. Gelman & Meng (1998) extended the bridge sampling estimator with a proposal named path sampling estimator. Other approximating methods among others include the Jeliazkov estimator (Chib & Jeliazkov 2001). Chib and Jeliazkov in 2001, estimated the marginal likelihood using a single run metropolis algorithm using the same idea. The power posterior estimator (Friel & Pettit 2008). A detailed review and comparison of some marginal likelihood estimators can be found in Gamerman & Lopes (2006), in Perrakis (2008) one can find comparisons of some marginal likelihood estimators with code for implementing them in R, Ntzoufras (see 2011, pg.392-397) briefly presents them and provides many references for further reading and Congdon (2007) makes a brief theoretical presentation of Laplace and BIC approximations.

Chapter 5

Bayesian Variable/Model Selection Using MCMC

Statistical problems where
"the things you don't know
is one of the thing you don't know"
are abiquitous in statistical modelling.
*G*reen P. 1995

The Bayesian approaches, for directly estimating the marginal likelihood, presented in section 4.3 are closely related to model-selection criteria : during our search for the best model we choose the one maximizing the marginal likelihood. However, the estimation of the marginal likelihoods, even when the covariates being considered for inclusion are of normal number, is rather impossible. The total number of models under consideration is equal to 2^p, when p covariates are included in our analysis.

In order to find the most probable model, the evaluation of the most promising candidates, is required. For that reason algorithms that search large parts of the model and parameter space were introduced firstly by George & McCulloch (1993). George and McCulloch presented the first Bayesian algorithm that searches the model space and a posteriori choose the most probable models named "Stochastic Search Variable Selection". Other algorithms, considerate

variable selection, the Kuo & Mallick sampler (Kuo & Mallick 1998), the Gibbs Variable Selection of Dellaportas et al. (Dellaportas et al. 2002). Another set of algorithms extending by generalizing the Gibbs-based algorithms, search the model space are the Carlin & Chib method (Carlin & Chib 1995), Metropolized Carlin and Chib by Dellaportas et al. (2002) and the reversible jump MCMC (Green 1995). Latest methods derived from those will be just mentioned with appropriate references for further reading.

In this section such variable and model selection algorithms will be described. For the conditional posterior distributions of the algorithms presented in the next section, in general we follow Ntzoufras notation (see Ntzoufras 2011, pg 409-412, 436-438). These techniques seems to be more appropriate when dealing with problems of high dimensions in comparison even to Bayesian Model Averaging, a technique which is discussed, analyzed in the next section and implemented again in chapter 7 using the BMA, BMS and BAS R packages created by Hoeting et al. (1999), Zeugner & Feldkircher (2009) and Clyde et al. (2011) respectively.

5.1 Bayesian Model Averaging

Following Hoeting et al. (1999) who provide us with a tutorial for executing BMA, let assume that Δ is the quantity in which we are interested for. Δ might be an effect size, then given the data, the average posterior distribution under each of the considered models, can be expressed as

$$\pi(\Delta|D) = \sum_{k=1}^{K} \pi(\Delta|M_k, \Delta)\pi(M_k|\Delta) \tag{5.1}$$

while the posterior distribution for every model from

$$\pi(M_k|\Delta) = \frac{\pi(D|M_k)\pi(M_k)}{\sum_{l=1}^{K} \pi(D|M_l)\pi(M_l)} \tag{5.2}$$

where,$\pi(M_k)$ is the prior probability that M_k is the true model and the likelihood ($\pi(D|M_k)$) of each model is given by $\pi(D|M_k) = \int \pi(D|\theta_k, M_k)\pi(\theta_k|M_k)d\theta_k$, θ_k is the vector containing the parameters of a particular model, $\pi(D|\theta_k, M_k)$ the likelihood.

Hoeting et al. (1999) provide equations for the posterior mean and the posterior variance of Δ and points out difficulties of Bayesian averaging over all models. One of the most important

being the fact that equation's 5.1 might contain a huge number of terms and therefore be impossible the summation procedure to be implemented.

Bayesia Model Averaging - Pros / Cons

(+) The technique avoids the problem of having to defend the choice of model, while considering model uncertainty.

(+) Model averaging is more naturally correct than considering only one model for estimating parameters.

(–) Is rather complicated for a simple presentation for large audiencies.

(–) Place much focus on the posterior effect probabilities (PEP).

(–) Higher estimates of variance than choosing only a single model.

(– / +) If no close form exists, approximations can be implemented e.g. for generalized linear models where no close form of the marginal likelihood exists a Laplace approximation can be applied.

5.2 Variable Selection Initial notions

It is common that an analysis in Bayesian terms begins by assigning to the unknown model parameters prior distributions. As we have noticed, there are also cases where uncertainty lies not only on the model covariates but on the whole model.

For variable selection methods, following George & McCulloch (1993, 1997), we assume a $\gamma \in \{0,1\}^m$ where γ contains inclusion indicators of a set of available covariates m. Considering the above we can now write the linear predictor of a generalized linear model as

$$\eta = \sum_{j=0}^{m} \gamma_j X_j \beta_j \tag{5.3}$$

where X is the design matrix and β the vector of the parameters of the full model, p the covariate included in the linear predictor. X_0 being the constant β_0 and equal to the first

column of X being filled with 1s. We then, following Ntzoufras (2011), create the partitions of γ, $(\beta_\gamma, \beta_{/\gamma})$, separating the β into those variables that are included ($\gamma = 1$) and those that are not included in the model ($\gamma = 0$). The above expanded regression formula is used for the Kuo-Mallick sampler and for Gibbs Variable selection.

Except for that way of constructing the model by substituting the parameter vector θ_j by $\beta_\gamma = \gamma_j \beta_j$, another way is to do not insert the γs in the above formula leading to the Stochastic Search Variable selection. In that way the parameter vector only contains the covariate effects and the indicator is involved in the model through a hierarchical structure of type $f(\gamma, \beta) = f(\gamma) f(\beta|\gamma)$, a structure that changes according to the sampler utilized. More details provided in sections 5.4.1, 5.4.2 and 5.4.3.

For the approaches that keep the dimension of model subspace constant, the variable selection algorithm can be seen as a decision problem of which β_js ,where $j = 1, ..., p$, p being the number of covariates, are equal to zero (O'Hara & Sillanpää 2009).

5.3 Zellner's g-prior and extensions

Before diving in model and variable selection techniques, we present the Zellner's g-prior and the more general scheme of Hyper g-priors introduced by Liang et al (Liang et al. 2008) as a generalization of the first g-prior. In 1986 Zellner introduced the g-prior, a prior for the special case of normal linear regression,

$$\beta_\gamma | \sigma^2, g \sim N_{q_\gamma}(0_{q_\gamma}, g\sigma^2 (X_\gamma' X_\gamma)^{-1}) \tag{5.4}$$

where $X_\gamma = (x_{\gamma 1}, x_{\gamma 2} \ldots, x_{\gamma n})$ is the design matrix, σ^2 usually considered unknown and assigned a Jeffreys prior.

The hyperparameter $g > 0$ is very influential on how the prior changes the result. Very large values of the hyperparameter lead paradoxically the process to prefer models which are less complex, this behavior is known as the Lindley - Jeffrey's Paradox (see section 4.2.1), while very small values of g make the posterior model probabilities to spread more equally, without considering differences in model size nor any other additional term. As a result the specification

of this hyperparameter is of great importance and many approaches for automatic specification were presented during the last years, e.g. (George & Foster 2000, Cui & George 2008).

The following can be considered as specific cases of g-priors, where g is the hyperparameter for the **linear case**.

- Unit Information Prior : $g = N$, where N is the number of observations, corresponds to the unit information prior (Kass & Wasserman 1995), an empirical version of which we are going to use in our examples and comparisons.
- Benchmark Prior $g = max(N, K^2)$, where K is the number of covariates corresponds to the benchmark prior proposed by Fernandez et al. (2001).
- Risk Inflation Criterion $g = K^2$ corresponds to the risk inflation criterion introduced by Foster & George (1994).
- $g = log(N)^3$, asymptotically resembles the Information Criteria proposed by Hanna-Quinn (see 4.1.2).
- Empirical Bayes procedures, containing Local Empirical Bayes and Global Empirical Bayes, for details see in Liang et al. (2008)
- Mixture of g-priors by Liang et al. (2008), Ley & Steel (2012) considering that g is not a constant and can be calibrated.

Liang et al. (2008) proposed the hyper g-prior, making g random, which retain a closed form expression for $f(y|\gamma)$ which is crucial for proper and efficient model inference (Sabanés Bové & Held 2011).

Sabanés Bové & Held (2011) in their paper proposed an extension of the classical g-prior for implementation in Generalized Linear Models, in which the hyperprior on g is handled in such a way that any continuous proper hyperprior $f(g)$ can be of use. As $n \to$ inf the prior on β_γ converges to the normal distribution and has the following form

$$\beta_\gamma | g, \gamma \sim N_{q_\gamma}(0_{q_\gamma}, gc\sigma^2(X_\gamma' W X_\gamma)^{-1}) \tag{5.5}$$

where c is a constant taking various values according to the generalized cases, while W is the weight matrix.

In the binomial regression, both parameters introduced in equation 5.5, are important. More specifically under the logit link, $c = 4$, under the probit $c = \pi/2$ and under the cloglog link, $c = e - 1$. For cases of not binomial regressions the most common value of c is 1.

Priors on the Model Space / Variable Indicators

The most common prior setting for the model space is the uniform prior, which in terms of variable selection using indicators results to

$$\gamma_i \sim \text{Bernoulli}(\theta), \text{ where } \theta = \frac{1}{2}$$

Other prior setting on model space include the suggestion of Ley & Steel (2009) for a binomial beta hyperprior, $\theta \sim Be(\alpha, \beta)$, on the a priori probability of including a variable. Lastly, if prior knowledge for a regressor exists then a fixed inclusion prior on the particular regressor can be applied.

5.4 Indicator variable selection algorithms

5.4.1 Stochastic search variable selection (SSVS)

The Stochastic search variable selection (SSVS) was one of the first Gibbs-based family algorithms proposed by George & McCulloch (1993, 1997). In this approach the linear predictor has the form

$$\eta = \sum_{j=0}^{m} X_j \beta_j \tag{5.6}$$

the indicators are not part of the linear predictor and the number of parameters under consideration each time is constant.

George and McCulloch introduced the binary indicator γ that shows if the i_{th} covariate will be included in the model ($\gamma_i = 1$) or not ($\gamma_i = 0$),

Let M be the set of models under consideration, this set can be represented by a vector of γ over all the possible sets of covariates and explain which of them are to be included in the model and which are not. Each covariate is modeled, as a mixture of two normal distributions. The ones with density concentrated around zero (but not equal to zero) and the others with density spread out over large regions.

This can be achieved by placing a spikeslab prior (O'Hara & Sillanpää 2009) on each parameter β_j. George & McCulloch (1993) proposed a mixture of two normal distributions as priors on β_j as shown in equation 5.7. In that way the indicators γ_j are involved in the model as a latent variables and have some convenient properties, noted further on.

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2) + \gamma_j N(0, g_j^2\tau_j^2) \tag{5.7}$$

This prior does not set non important regressors exactly equal to zero but rather places them in a region close to zero. By setting a very small value for $\tau_j > 0$ we achieve in successfully estimating the effect of the candidates β_j, if $\gamma_j = 0$, close to zero. By the same time setting, a rather large value for $g_j > 1$, we achieve when $\gamma_j = 1$, β_j to be non zero. So if $\gamma_j = 0$, β_j is not candidate for inclusion, while when $\gamma_j = 1$, β_j is important and available as candidate for a model covariate. Keep in mind that the model dimensions do not change while the β_j are evaluated. George & McCulloch (1993) give insight on how to choose g_j^2 and τ_j^2.

The full conditional posterior distribution of the regressors, considering the mixture of Normal densities in 5.7, can be written from

$$f(\beta_j|y,\gamma,\beta_{/j}) \propto f(y|\beta,\gamma)f(\beta_j|\gamma_j) \tag{5.8}$$

to

$$\begin{aligned} \gamma_j = 1 \quad &: \quad f(\beta_j|y,\gamma,\beta_{/j}) \propto f(y|\beta,\gamma)f(\beta_j|\gamma_j) \\ \gamma_j = 0 \quad &: \quad f(\beta_j|y,\gamma,\beta_{/j}) \propto f(y|\beta,\gamma) \end{aligned} \tag{5.9}$$

Due to the form of the model structure, independence among γ and y is assumed. Therefore, the full conditional posterior distribution of the indicators γ_j is given by

$$\gamma_j|\beta,\gamma_{/j},y \;\propto\; Bernoulli\left(\frac{O_j}{1+O_j}\right)$$

where

$$O_j = \frac{f(\gamma_j=1|y,\gamma_{/j},\beta)}{f(\gamma_j=0|y,\gamma_{/j},\beta)} = \underbrace{\frac{f(\beta|\gamma_j=1,\gamma_{/j})}{f(\beta|\gamma_j=0,\gamma_{/j})}}_{\text{Betas Prior}}\underbrace{\frac{f(\gamma_j=1|\gamma_{/j})}{f(\gamma_j=0|\gamma_{/j})}}_{\text{Prior Odds}} \tag{5.10}$$

5.4.2 Unconditional priors Gibbs sampler (KM)

The most straight forward approach for dealing with variable selection is to use the spikeslab prior by setting slab $\theta_j|(\gamma_j=1)$ equal to the parameters included and the spike, $\theta_j|(\gamma_j=0)$, equal to zero.

We can always factorize the joint distribution as

$$f(\gamma,\beta) = f(\gamma)f(\beta|\gamma) \tag{5.11}$$

Kuo & Mallick (1998) sampler (KM) made the convenient assumption that the indicators are a priori independent of the effects $f(\gamma,\beta) = f(\gamma)f(\beta)$. The above assumption makes the priors of β and γ indicators independent. By now partitioning β to $(\beta_\gamma,\beta_{/\gamma})$ then $f(\gamma,\beta) = f(\gamma)f(\beta_\gamma|\beta_{/\gamma})$.

As a result of the independence the full conditional posterior distribution is

$$\begin{aligned}\gamma_j = 1 \;&:\; f(\beta_j|y,\gamma,\beta_{/j}) \propto f(y|\beta,\gamma)f(\beta_j|\beta_{/j})\\ \gamma_j = 0 \;&:\; f(\beta_j|y,\gamma,\beta_{/j}) \propto f(\beta_j|\beta_{/j})\end{aligned} \tag{5.12}$$

One might notice that when the effect is estimated by zero, $\gamma_j = 0$, the proposed values only depends on the conditional likelihood $f(\beta_j|\beta_{/j})$.

The full conditional posterior distribution of the indicators γ_j is given by

$$\gamma_j|\beta,\gamma_{/j},y \;\propto\; Bernoulli\left(\frac{O_j}{1+O_j}\right)$$

where

$$O_j = \frac{f(\gamma_j = 1|y, \gamma_{/j}, \beta)}{f(\gamma_j = 0|y, \gamma_{/j}, \beta)} = \underbrace{\frac{f(y|\gamma_j = 1, \gamma_{/j}, \beta)}{f(y|\gamma_j = 0, \gamma_{/j}, \beta)}}_{\text{Likelihood}} \underbrace{\frac{f(\gamma_j = 1, \gamma_{/j})}{f(\gamma_j = 0, \gamma_{/j})}}_{\text{Prior Odds}} \tag{5.13}$$

In the above equation we can easily notice that the prior on βs are not included in the calculation of the posterior probability, due to the assumption of independence and only depends on the likelihood. KM is easy to be implemented but attention should be given because the algorithm considers one prior $f(\beta)$ over all models independent of the model structure. Such a hypothesis is restrictive and even though the algorithm is pretty simple to apply, there is no way for the researcher to improve its efficiency.

5.4.3 Gibbs Variable Selection (GVS)

Dellaportas et al. (2000) proposed the Gibbs Variable Selection, a similar to the Kuo-Mallick sampler technique, which main difference summarizes in the fact that the parameters proposal had no independency assumption and had also an hierarchical structure of $f(\gamma, \beta) = f(\gamma)f(\beta|\gamma)$. Considering that β can be separated into $(\beta_\gamma, \beta_{/\gamma})$, $f(\gamma, \beta) = f(\gamma)f(\beta|\gamma) = f(\gamma)f(\beta_\gamma|\gamma)f(\beta_{/\gamma}|\beta_\gamma, \gamma)$.

the full conditional posterior distribution can be written as

$$\begin{aligned} \gamma_j = 1 \quad &: \quad f(\beta_j|y, \gamma, \beta_{/j}) \propto f(y|\beta, \gamma)f(\beta_\gamma|\gamma)f(\beta_{/\gamma}|\beta_\gamma, \gamma) \\ \gamma_j = 0 \quad &: \quad f(\beta_j|y, \gamma, \beta_{/j}) \propto f(\beta_{/\gamma}|\beta_\gamma, \gamma) \end{aligned} \tag{5.14}$$

and the full conditional posterior distribution of the indicators γ_j is given by

$$\gamma_j|\beta, \gamma_{/j}, y \ \propto \ Bernoulli\left(\frac{O_j}{1 + O_j}\right)$$

where

$$O_j = \frac{f(\gamma_j = 1|y, \gamma_{/j}, \beta)}{f(\gamma_j = 0|y, \gamma_{/j}, \beta)} = \underbrace{\frac{f(y|\beta, \gamma_j = 1, \gamma_{/j})}{f(y|\beta, \gamma_j = 0, \gamma_{/j})}}_{\text{Likelihood}} \underbrace{\frac{f(\beta|\gamma_j = 1, \gamma_{/j})}{f(\beta|\gamma_j = 0, \gamma_{/j})}}_{\text{Betas Prior}} \underbrace{\frac{f(\gamma_j = 1|\gamma_{/j})}{f(\gamma_j = 0|\gamma_{/j})}}_{\text{Prior Odds}} \tag{5.15}$$

We should notice that the pseudopriors of $f(\beta_{/\gamma}|\beta_\gamma, \gamma)$, solo role is for improving the efficiency of the algorithm. In order for optimal convergence to be achieved they have to be tuned, most of the times just a pilot run of the full model is required to create the proposal for the pseudoprior (Dellaportas et al. 2002).

Synopsis

The main differences between the above tools lies in how the prior and the linking densities, function in each algorithm. In the variable selection step of GVS both likelihood and prior appear, while in SSVS the likelihood is absent and in KM sampler the prior on β is omitted as β and γ are independent by definition. Further theoretical comparisons of those variable selection techniques and an explanatory table of their association can be found in Ntzoufras (see 2011, pg 412 - Table 11.7), along with application of GVS using different priors on β.

5.5 Model space search algorithms

Let us assume once again, as we already did in section (4.3) describing the direct methods , that we have M competing models and that the dependent data are the product of a particular model $m \in M$. We will extend the posterior model distribution from two competing model, which was given in equation (4.1), to m competing models. Let $f(m)$ be the prior distribution for model m, then $f(m|y) = \dfrac{f(m)f(y|m)}{\sum f(m)f(y|m)}$ from which $f(y|m) = \int f(y|m, \beta_m)f(\beta_m|m)d\beta_m$ where $f(\beta_m)$ is the conditional distribution of β_m, β_m are the model parameters for the model under consideration $\beta_m \in B_m$ and B_m are set of all possible values for the models regressors.

It is a fact that most model settings require the model and the parameter space to be jointly searched by the MCMCs. The joint sampling space is $M\ x\ \prod_{m\in M} \beta_m \subset M\ x\ \prod_{m\in M} R^{N_m}$. The marginal posterior probabilities for models can be acquired by $f(m|y)$, while the posterior estimation of the regressors of each model $f(\beta_m|m, y)$ can be acquired just by conditioning on the samples produced when the chain is in state m (Carlin 2001).

During the last years, many techniques were proposed that generate values from the posterior distribution $f(m, \beta_m)$ and then estimate the posterior probability and the conditional distribution of β_m.

The most straight forward and easy to implement way of generating values from the joint posterior of (m, β_m) is the **Indepedence Sampler** . A proposal of $(m^{'}, \beta_m^{'})$ for a given state of (m, β_m) is generated from a proposal distribution and the proposal is accepted with acceptance probability of MH $\alpha = min\left(1, \frac{f(y|m^{'}, \beta_{m^{'}}^{'}) f(\beta_{m^{'}}^{'}) f(m^{'}) q(m, b_m|m^{'}, b_{m^{'}}^{'})}{f(y|m, \beta_m) f(\beta_m) f(m) q(m^{'}, b_{m^{'}}^{'}|m, b_m)}\right)$. We should notice that the Independence sampler works most efficiently if the q is somehow approximate to the target distribution. In practise one should construct an approximate estimate of both $f(m|y)$ and $f(\beta_m|m, y)$ for every model m. When the number of possible models is large the computations become cumbersome and a different way for estimating the posterior should be considered.

In the **method of Carlin and Chib** (CC) (Carlin & Chib 1995), the likelihood of the indicator model m is given by $f(y|\beta_m, m)$, while the prior for the same model by $f(\beta_m|m)$. We then denote as $m^{'}$ all the models under consideration. The marginal likelihood for the indicator model, which by assuming independence between β_m of each model for convenience, is given by $f(y|m) = \int f(y|\beta, m) f(\beta|m) d\beta = \int f(y|\beta_m, m) f(\beta_m|m) d\beta_m$. The Carlin and Chib's method then uses a Gibbs sampler over the full conditional distribution of the regressors.

All the required full conditional densities are defined and the algorithm (Gibbs sampler) will produce samples from the right joint posterior model distribution (Han & Carlin 2000). In comparison to the methods described in the previous section 5.4.3 and 5.4.2 where only one prior at each step is required to be computed, Carlin and chib's method main drawback is the need to specificate all possible priors on the parameters under each model, $f(\beta_{m^{'}=m}|m)$, in order to compute the full conditional distributions, making the method computationally expensive when M is pretty large (Han & Carlin 2000). Since then, the method has been extended leading to the even more general composite model space framework of Godsill (2001)

Dellaportas, Forster and Ntzoufras 2002 proposed a Metropolis within Gibbs strategy, containing a model selection step which is based on proposal for making a move from model m to $m^{'}$. That method is called " **Metropolized Carlin and Chib**". Therefore, by introducing MCMC in the model selection step, therefore "Metropolizing" the step, the above method requires only to sample from the pseudo–prior for the model under consideration $m^{'}$. In Han & Carlin (2000) a comparative study indicates that the MCC performs two times faster than the CC method under all examples considered, while Dellaportas et al. (2002) indicate that Model

Composition MCMC (MC^3) (Madigan et al. 1995) is a special case of MCC algorithm.

Bayes factor and posterior model probabilities can be estimated as presented in section 5.7

5.6 Latest Variable / Model selection algorithms

Hans et al. (2007) created an algorithm (**Shotgun Stochastic Search (SSS)**) to explore more effectively the model space, especially for higher dimension problems. Their proposal searches region of models with high posterior probability. To achieve that they run parallel chains, making possible the simultaneous evaluation of more models. In particular, let assume that the current state of the algorithm is in model $\gamma_{current}$, where 1 $\leq$ current $\leq$ p, and p the number of candidate models. In each iterative scene three possible proposals are made. $\gamma^{+}_{current}$, when proposing to add a variable in the model, $\gamma^{o}_{current}$ when proposing to change a variable of the model, while $\gamma^{-}_{current}$ when proposing to exclude a variable from the model. After evaluating each proposed move, only one is selected and the steps are repeated, $\gamma_{current'} \in \{\gamma^{+}_{current}, \gamma^{o}_{current}, \gamma^{-}_{current}\}$. By that way the model space is explored in a faster way.

Another rather recent proposed algorithm is the one called **Subspace Carlin and Chib** (SCC) algorithm (Petralias & Dellaportas 2012). Petralias and Dellaportas created a way to combine the Metropolised Carlin and Chib with the SSS algorithm. SCC algorithm was inspired by the SSS initial idea of creating neighborhoods and sampling simultaneously from them.

5.7 Posterior model/variable selection inference

Ntzoufras (2011) presents alternative ways for model comparison according to what family of algorithms one chooses to use. We present the most simple, while for details and alternatives the reader is prompted to the aforementioned citation. For those using an indicator γ, it is most common to estimate the maximum a posteriori model (MAP) by estimating

$$f(\gamma_j = 1) = \frac{1}{T-B} \sum_{t=B+1}^{T} I(\gamma_j^{(t)} = 1) \tag{5.16}$$

or the posterior model probabilities

$$f(m|y) = \frac{1}{T-B} \sum_{t=B+1}^{T} I(m^{(t)} = m) \tag{5.17}$$

where I() is the indicator function, having value equal to 1 when $\gamma_j^{(t)} = 1$, for 5.16 or when $m^{(t)} = m$, for 5.16, and 0 otherwise. m is a model indicator transforming the indicators to a unique decimal number through $m(\gamma) = 1 + \sum_{j=1}^{p} \gamma_j^{2^{j-1}}$ when the constant is included in all models compared.

The posterior Bayes factor is considered an appropriate way of comparing two models, based on the posterior predictive densities of the data being observed. The posterior Bayes factor comparing two models $M1, M2$ equals to

$$PBD = \frac{f(y|m_1)}{f(y|m_2)} = \frac{\int f(y|\beta_{m_1}, m_1) f(\beta_{m_1}|m_1) d\beta_{m_1}}{\int f(y|\beta_{m_2}, m_2) f(\beta_{m_2}|m_2) d\beta_{m_2}} \tag{5.18}$$

which can be estimated easily from the MCMC output by the posterior mean of the likelihood over all sampled parameter values obtained.

$$f(y|m) = \frac{1}{T} \sum_{t=1}^{T} f(y|\beta_m^{(t)}, m) \tag{5.19}$$

5.8 GVS, SSVS and KM implementation in BUGS

The three techniques for variable selection using γ indicator for GLMs introduced in section 5.4 are : SSVS by George & McCulloch (1993), KM by Kuo & Mallick (1998) and GVS by Dellaportas et al. (2000), while further extensions for GLM can be found, for log-linear by Ntzoufras et al. (2000) and for multivariate regression by Brown et al. (1998). The main aspect of the aforementioned techniques is that the most probable covariates can be pointed out with the use of their posterior probability. The "optimal" subset of covariates is the one which appears most frequently during the MCMC implementation. This "optimal" subset can be identified through the analysis of the marginal posterior distribution of γ, $f(y|\gamma)f(\gamma)$, containing information both from our data and the prior of γ.

The results will the be compared to then Bayesian Model Averaging package (BMA) which assumes a Laplace (BIC) approximation for Generalized Linear Models for parameter priors and a uniform distribution over the model priors. BMA as we will point out in chapter chapter:rBasVarSel has a very restrictive toolkit in comparison to BMS package and BAS packages also performing Bayesian model averaging in R, though being the only one currently able to perform Bayesian model averaging on Generalized Linear Models.

The difference as already briefly noticed lies on where and how γ is used in the model. SSVS place γ in the prior of the regression coefficients, in KM they become a part of the regression formula, while GVS bridge SSVS and KM. We will present those three algorithms and their difference by implementation in WinBUGS for an extended logistic regression model of EHIS 2009 and for simulated data from a binomial regression generated using an inverse logit function.

As we will not take into account interactions of the covariates included in the models of the rest of this thesis, when the "Full Model" is mentioned, we will be referring to the model with all main effects.

In the next section we will implement model and variable comparisons on both simulated data and our EHIS 2009 example, using three of the aforementioned algorithms (GVS, SVSS, KM) and comparing the posterior inclusion probabilities to those derived from BMA, keeping the prior on β constant to compare the efficiency and speed of each procedure. Lastly, we will argue on whether or not one model is more probable than the others by reporting the Bayes Factors comparing the probabilities of each model to the most probable.

5.8.1 Simulated data

Following among others Ntzoufras (2002, 1999, 2011) and Dellaportas et al. (2002), the code in Appendix's section (9.4.1) for implementing the Kuo - Mullick Sampler, the Stochastic Search variable selection and the Gibbs variable selection to our data was modified appropriately, to fit logistic regression models. The differences noted in theory can also be noticed in the presented code. Attention should be given on the way the priors for each method changes, while the

general set of variable selection remains the same, always under the convenient hypothesis of independence between the covariates.

The analysis was first performed on simulated data to check the consistency of the algorithms and the modified codes and then in the next section on our EHIS 2009 data. The simulated data were drawn according to $X_i \sim N(0,1)$, where $i = 1, \ldots, 15$ of total size of 400. The model was formed from X_1 & X_2 using a normal distribution on the linear predictor $Y_i = log\left(\frac{\pi_i}{1-\pi_i}\right)$. The covariates were standardized and by that achieving minimaziation the influence of the constant. and the prior precision of each parameter was acquired by a pilot run of the full model's algorithm, $\beta_j \sim N(0, nS^2_{\beta_j})$. While taking into account the number of the sample, we mimic an UIP Empirical Bayes independent prior distribution accounting for one data point. The indicators γ will be distributed by $\gamma \sim bin(0.5)$ for all covariates except for the constant which is included in all models under consideration.

Some comments on **BMA** package which will be further reviewed and compared in chapter 7. BMA for variable less than 30 variables enumerates the model space with the use of a Laplace approximation on the likelihoods of each model. It also uses a technique called Occam's window (Madigan & Raftery 1994) that acts like a way for the summation of the denominator to be reduced, be excluding models that are not so probable and adjusting the PIP according to the new reduced model space. It reports a variety of quantities among which are the posterior probability of the effect not being equal to zero, which is computed after summation of the number of the models in which the effect is present and the posterior model probabilities of the reduced model space.

The following tables (5.1, 5.2) summarize the variable / model selection results of the different sampling schemes implemented.

We shall notice that all MCMC methods conclude to Model {x1, x2} as the one with the highest posterior model probability (PMP). The maximum a posterior model (MAP) is, in each application, around 7 times more likely than the 2_{nd} in rank. Predictors {x1, x2} are present in all five most probable models, leading to more clear evidence that are important for the logistic model under consideration. The fact that no great differences are noticed among the reported PIPs may be due to the fact that the number of the dataset is large (n=400), as

in bibliography is noticed that SSVS, having the likelihood not taking part in the computation of the PIPs usually reports different results.

BMA fully enumerated the model space and chose using the Occam's window a small amount of models. When deciding to create a bigger window setting the Odds Ration equal to 500000, considering all models which BF(MAP) against all other models is equal or smaller than 500000. Considering this approach a total number of 482 models were included, resulting to similar results to the WinBugs methods.

Like BMA's Occam's window, the researcher can reduce the model space and compute the model selection probabilities, for the MCMC methods implemented in WinBUGS, of the most probable regressors, over the chosen reduced model space.

	WinBugs*			R**
	SSVS	KM	GVS	BMA#
Model	Posterior Model Probabilities			
X1+X2	39,65	41,95	41,45	39,65
X1+X2+X9	5,43	6,19	6,15	6,03
X1+X2+X8	4,96	5,45	5,81	6,21
X1+X2+X5	3,78	4,53	4,51	4,52
X1+X2+X10	4,13	5,33	5,23	5,02

Table 5.1: Model selection results of simulated data for SSVS, KM, GVS and BMA.# Occam's Window OR=500000, * UIP Empirical Bayes independent prior, **Laplace Approximation. Iterations=15000, Burnin Period = 5000

5.8.2 EHIS 2009 data

Having the three algorithmic variable selection schemes in terms of simulated data presented in the previous section, we will now apply them to the EHIS 2009 data which were presented in Chapter 2, identifying relation between "Self Reported Health Status" and the "Level of Education" and revisited by the introduction of other possible covariates in section (5.8).

	SSVS		KM		GVS		BMA#	
#	PIP	MCerror	PIP	MCerror	PIP	MCerror	PIP	P.Mean
X1	99,2	<0,0001	100,0	<0,0001	100,0	<0,0001	100,0	0,0908
X2	98,8	0,001208	100,0	<0,0001	100,0	<0,0001	100,0	0,0838
X3	6,7	0,001951	4,1	0,002452	4,1	0,001569	4,8	-0,0001
X4	7,0	0,002124	4,5	0,002787	4,7	0,001635	5,1	0,0003
X5	8,8	0,002593	10,0	0,004153	9,7	0,003093	10,0	-0,0013
X6	6,8	0,002194	6,2	0,003038	5,8	0,002188	6,2	-0,0002
X7	7,3	0,002201	6,6	0,003567	6,1	0,001957	6,3	0,0003
X8	10,9	0,003099	12,8	0,0052	13,3	0,003721	13,9	0,0057
X9	11,9	0,003151	12,6	0,005363	13,6	0,003631	13,9	-0,0014
X10	9,5	0,002579	11,2	0,004649	10,7	0,00336	11,0	-0,0009
X11	7,5	0,002234	6,4	0,003464	6,7	0,002251	6,8	0,0003
X12	9,5	0,002507	9,4	0,00419	9,7	0,002517	9,8	0,0016

Table 5.2: SSVS - KM - GVS - BMA variable selection indicators for binomial simulated data. #Occam's Window OR=500000. Iterations=15000, Burnin Period = 5000

We will again use the same algorithmic scheme as the one described in the section of the simulated data. The prior precision of the parameters will be acquired from a pilot run of the full model, then divided by the length of our dataset to mimic an UIP Empirical Bayes independent prior distribution accounting for one data point, while the prior mean will be equal to zero. The indicators γ will be distributed by $\gamma \sim bin(0.5)$ for all covariates except for the constant which is included in all models under consideration.

The results provided, clearly state that "Age", "Education" and "Long - Illness" should be included in the model. They are all included in the MAPs models, while retain a PIP equal to one, independently of the method applied. Furthermore, "Sex" is indicated by the majority of the methods as of medium importance predictor. SSVS reports a lower PIP for both "Sex" and "Urban", which was anticipated due to major difference in calculating the PIPs compared to GVS and KM approach. BMA reports results once again close to the other methods, except for SSVS that seems to prefers a more parsimonious model.

	Posterior Model Probabilities			
	WinBugs*			R**
Model	SSVS	KM	GVS	BMA#
Age+Educat+Longill	81,54	45,98	49,84	47,23
Age+Sex+Educat+Longill	11,08	35,82	35,55	34,31
Age+Urban+Educat+Longill	6,54	11,34	9,42	11,57
Age+Sex+Urban+Educat+Longill	0,84	6,86	5,18	6,89

Table 5.3: SSVS - KM - GVS - BMA model selection results of EHIS 2009 dataset. Age = Age(in Years), Sex = Sex(Male/Female), Education=Please refer to table 9.4, Long-illness = Long Illness(Yes/No), Urban=Urban(Yes/No). #Occam's Window OR=500000, * Empirical Bayes independent prior, **Laplace Approximation. Iterations=15000, Burnin Period = 5000.

	SSVS		KM		GVS		BMA#
Predictors	PIP	MCerror	PIP	MCerror	PIP	MCerror	PIP
Age	100	<0.0001	100	<0.0001	100	<0.0001	100
Educat	100	<0.0001	100	<0.0001	100	<0.0001	100
Longill	100	<0.0001	100	<0.0001	100	<0.0001	100
Sex	11,92	0,004	42,68	0.0188	40,74	0.0106	41,20
Urban	7,38	0,003	18,2	0.01268	14,61	0.00497	18,50

Table 5.4: SSVS - KM - GVS - BMA variable selection indicators, Standard Deviations and MC errors for EHIS 2009 Data. #Occam's Window OR=500000. Iterations=15000, Burnin Period = 5000

5.9 Closing remarks

This chapter tried briefly to touch many aspects of how model comparison is dealt with from the Bayesian perspective. The Bayesian point of view, on model choice, seems more natural. As at the end we are not given the "optimal" solution picked by an algorithm but the chances that our data are produced by the models under consideration. As a result the way to compare models is more straightforward. Moreover, Bayesian Model Averaging provides a more natural and accurate estimation of the effects, while incorporating model uncertainty and is claimed to be better than most of the classical methods for model parameters estimation.

For example, SSVS to be implemented the close form of the full conditionals should be available, otherwise the most common way to deal with cases were no close form is available is through adaptive MCMCs. In problems though, when the parameter do not have the same role from iteration to iteration the recent history cannot and should not be used.

We should notice a package named "R2WinBUGS" (Sturtz et al. 2005) that automates the sampling procedure through R working environment, without the researcher having to work in WinBUGS/OpenBUGS. In that way, the researcher should only create a small file ".bug" containing the Gibbs sampling statements for BUGS and then by programming just in R can call that function, saving all quantities in R for easier further handling.

More information about Model Choice and comparison can be found in a plethora of resources (e.g.Congdon (2007), chapter 2)

The transdimendial algorithm "reversible jump" introduced by Green (1995), used also when both the number of the parameters under consideration and the parameters themselves are unknown will be presented in chapter 6 for linear regression model and compared to GVS method under the same prior considerations.

Chapter 6

Reversible Jump MCMC

6.1 Introductory notions

Reversible jump MCMC (rjMCMC) can be claimed that was initially based on Birth and Death processes. The general idea is to create in continuous time a Markov chain, e.g. a Markov jump process, which has the ability to move between models, by births (increasing of dimension), deaths (decreasing of dimension), and other moves in general. Details can be found in Ripley (1977) and Stephens (2000) among others.

The method introduced by Green (1995), is another variable selection strategy that help us sample over the model space, avoiding the full product space search of CC method, while introducing a less straight forward algorithm.

The reversible jump has the ability to jump between models with different number of parameters, while at the same time achieve to maintain the properties of a MCMC chain needed for convergence – Aperiodicity,irreducibility and positive recurrence – (see section 2.2). The iterative schemes for updating the model parameters given that we are in a particular model is just a standard MCMC algorithm procedure. While the phase where a proposal to move from model m to a new model $m^{'}$ is the new part of the rjMCMC and will be discussed in detail below.

Let assume that we currently are in a state of $\{m, \beta_m\}$, β_m the m model parameter vector, for each reverisible jump step of the algorithm a new $\{m^{'}, (\beta_m)^{'}\}$ is proposed and the dimensions from model $m \to m^{'}$ now differ $dim(\beta_m) \to dim((\beta_m)^{'})$. For the reversible jump algorithm to retain the properties for convergence, both the condition of reversibility (a condition that by definition is satisfied in a simple MCMC) and that of dimension matching (Green 1995) should be satisfied. For achieving a $m \to m^{'}$ move where the dimesions of the two models differ, a random parameter vector u_m is introduced, so as $\{\beta_m, u_m\} = \{\beta^{'}_{m^{'}}, u^{'}_{m^{'}}\}$ to hold.

If the transformation $\{\beta_m, u_m\}$ to $\{\beta^{'}_{m^{'}}, u^{'}_{m^{'}}\}$ is a diffeomorphism, in other words, the transformation and its inverse are differentiable (Green 2003), a valid choice for α is $\alpha_{m \to m^{'}} = min\left(1, \frac{f(\beta^{'}_{m^{'}})q^{'}(u^{'}_{m^{'}})}{f(\beta_m)q(u_m)} \left| \frac{d(\beta^{'}_{m^{'}}, u^{'}_{m^{'}})}{d(\beta_m, u_m)} \right| \right)$

Therefore, one way of representing the reversible jump algorithm is through the following steps

- Let assume that the present state of the Markov chain is (m, β_m), where β_m is of dimension N_m.
- Similarly to MCC propose a new model $m^{'}$ with probability equal to $j(m, m^{'})$.
- Then, generate u from a proposal distribution $q(u_m|\beta_m, m, m^{'})$.
- Specify the "dimension – matching" function by setting $(\beta^{'}_{m^{'}}, u^{'}) = g_{m \to m^{'}}(\beta_m, u)$, where $g_{m,m^{'}}$ is a deterministic invertible 1-1 function.
- Accept the proposed mode from model m to model $m^{'}$ with acceptance probability equal to
- With probability $1 - \sum_{m:k} j(m \to m^{'})$ no move is attempted.

$$\begin{aligned} \alpha_{m \to m^{'}} &= min\left(1, A\left[(\beta_m, m) \to (\beta_{m^{'}}, m^{'})\right]\right) \\ &= min\left(1, \frac{f(y|m^{'}, \beta^{'}_{m^{'}})f(\beta^{'}_{m^{'}})f(m^{'})j(m^{'} \to m)q(u^{'}_m|m, m^{'}, b_m)}{f(y|m, \beta_m)f(\beta_m)f(m)j(m \to m^{'})q(u_m|m^{'}, m, b_{m^{'}})} X \left| \frac{dg_{m \to m^{'}}(\beta_m, u_m)}{d(\beta_m, u_m)} \right| \right) \end{aligned} \tag{6.1}$$

where $\beta_m = g_{m\to m'}(\beta_m, u_m)$ for a random parameter vector $u_m \sim q_m(u)$, while $q(m \to m^{'}$ is the transition probability from model m to $m^{'}$. The second part of the above acceptance probability contains a Jacobian matrix, which in fact is the basic difference between a simple metropolis hastings algorithm and the reversible jump.

The move in rjMCMC can be either a standard MH step or a Gibb step. Bayes factor and posterior model probabilities can be estimated as presented in section 5.7.

The reversible jump algorithm is not limited to countable set of models M, although is frequently presented in such way (Sisson 2005). The researcher can perform implementation of the sampler without previous knowledge on the size of the model space at all. An example of a problem's setting with unknown number of models is Bayesian nonparametrics, e.g. Fractional polynomial regression (Royston & Altman 1994).

When the full (normalized) conditional probabilities of each model m are known in closed form, $f(\beta_m|x)$, which is the case as far as regression is concerned. If the random parameter vector $u_m \sim q(u) = f(\beta^{'}_{m'}|x)$ is a direct draw from the posterior distribution and the proposal state $\beta^{'}_{m'} = u_m$ the second part of the acceptance rate of equation 6.1 simplifies to

$$A\left[(\beta_m, m) \to (\beta_{m'}, m^{'})\right] = \frac{f(m^{'})j(m^{'} \to m)l_{m'}(x)}{f(m)j(m \to m^{'})l_m(x)} \tag{6.2}$$

which is now independent of the current and the proposed states of the parameter, resulting to a fixed dimensional sampler over the space of models (Clyde 1999). This simplified simulation as described above it is possible to be implemented in WinBUGS using the "jump Interface" presented by Lunn et al. (2006, 2009) as an add-on to the initial WinBUGS suite.

The rjMCMC was widely used in the literature from the time it was introduced. It is a fact that one in every five citations of Green's algorithm can be classified as genetic-related research, the vast majority of implications lies on the generic problem of model selection. Sisson (2005) summarizes some programs created for implementation of transdimensional samplers providing enough detail for the reader.

To conclude this section, we should notice that even though the algorithm has dominated between the available methodologies from Bayesian model selection, there are situations where

an different algorithm will provide a simpler implementation. Ntzoufras (2002) claims that when the total number of models under consideration is small, the adoption of a product-space approach may be more useful.

6.1.1 Comparison notes Reversible Jump Vs. Carlin and Chib

All the methods that use transdimensional sampling for comparing models, at least until today, should retain the dimensionality of all compared models matched. One of the key differences between Carlin and Chib method and the reversible jump was the fact that in reversible jump the dimension of the transdimensional model becomes equal to the dimension of the highest-dimensional of the models under consideration, while in Carlin and Chib the dimension of the transdimensional model is the product space of all compared models. When the two methods were firstly introduced the reversible jump made use of the Metropolis Hastings algorithm, while Carlin and Chib used the Gibbs sampler.

However, further research bridged the first key difference, as the whole product space of Carlin and Chib can be lowered by considering that some parameters contained the same kind of information and be considered as part of the same information, making the number of pseuroparameters used smaller. In that way, when some parameters are shared between models, the product space method does not always apply a purely product space, diminishing the boundaries between the two methods (Lodewyckx et al. 2011). Dellaportas et al. (2002) managed to insert the more general metropolis hastings sampling to the Carlin and Chib method while Lunn et al. (2009) proposed a way that reversible jump can be implemented with the use of Gibbs sampling as well.

6.1.2 Population-Based Reversible-Jump MCMC

When dealing with multimodals distributions, is quite impossible to explore adequately the space in and between the distribution's mode. To overcome such situations Jasra et al. (2007) proposed the Population based Reversible Jump, having the ability to sample from more than one chains in simultaneously. Details can be found, together with a suggestion of a practical

way of moving between the distribution modes, initiating different chains, in Fouskakis et al. (2009).

6.2 DAG for probabilistic models

Noticing a simple way to treat DAGs, which were discussed in section 1.5 to show dependencies distributions in the same way as probabilities. Expressing conditional independence graphically one can easily understand them and later implement computational tricks (e.g. vectorization - parallel computing).

Let us assume that we have the following simple model

$$\begin{aligned} k &\sim Bernouli(\pi) \\ B_2 &\sim Uniform(k, k+2) \\ B_1 &\sim Uniform(k, k+3) \\ \Psi &\sim Normal(B_1 + B_2, 1) \end{aligned} \tag{6.3}$$

In our example distributions of B_1 & B_2 only depends on k, while the distribution of Ψ only depends on the values of B_1 & B_2. Then, the above model can be formalized using a DAG, expressing the way of causality as in figure 6.1.

Having the basic idea on how a probabilistic model is depicted as a Directed acyclic graph, we will in the next section illustrate the way the jump interface works and how is represented as a DAG.

The basic hierarchical model for Bayesian model selection derived from the joint distribution of $\{\beta_m, \gamma\}$ is presented in figure 6.2, where γ represents $p(\gamma)$, θ represents $p(\beta_m, \gamma)$, referring to Jump's Interface for WinBUGS. More information is provided in the next section.

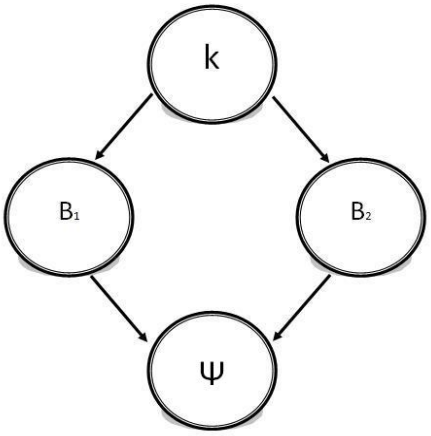


Figure 6.1: Directed Acyclic Graph of model presented in equation 6.3

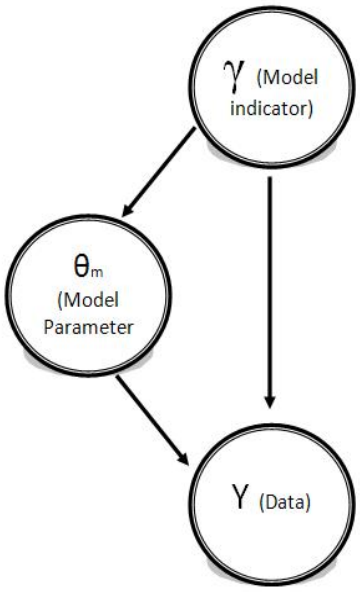


Figure 6.2: Basic hierarchical model for Jump's Interface reversible jump as a DAG. (Lunn et al. 2006)

6.3 Jump Interface rjMCMC

WinBUGS from 2006 has been extended and now can handle variable selection among models with the use of Reversible Jump MCMC. In this section we will first describe the new add-on of WinBUGS, then illustrate it using the simulated data from a linear regression model, following Clyde et al. (2011) simulated linear regressors, the results obtained will be presented and then compared in the next chapter with the Bayesian model averaging programs of R and code for SSVS, KM and GVS for variable selection of linear models.

With the use of reversible jump we are no longer restricted in the number of models to be evaluated. The algorithm uses the observed data and search over different models, as it considers the model as an extra discrete parameter. The posterior distribution is now consisted of both the parameter and the model space, and the reversible jump has the ability to search both spaces simultaneously.

We will now explain how the Jump extension of WinBUGS (Lunn et al. 2006) can be of use on variable selection of several types of regression and more particular the logistic regression. In the code presented in the Appedix 9.4.2, one can notice the presence of an index k and an indicator variable *id*. k is the number of βs that are currently in the model when a certain draw is performed. The parameter β_0 is alway part of the model. The indicator *id* is of certain interest and it is used to produce the variable selection plot and the probabilities of inclusion of each variable, it indicates in which model is the sampler located during a MCMC draw.

In our example of simulated data the indicator showing that the parameters β_1 & and β_{15} are in the model will be equal to id=100000000000001. The researcher has the ability to directly change only the prior distribution of k and not the joint distribution of $\{\theta, k\}$, which can be indirectly been configured when changing the prior on k. The initial choice of the Jump extension for $f(k)$ is $k \sim Binomial(0.5, N)$, where N is the total number of parameters under consideration. This leads to $p(\theta, k) = p(\theta|k)p(k) = \dfrac{1}{2^Q}$, which yield equal probability of selection for all models under consideration.

Another way to model the prior on k is to assume that the parameter p is not equal to 0.5 but is generated by a beta distribution with parameter $p \sim Beta(\alpha, \beta)$. A structure that tends to "shrunk" the probabilities towards zero as indicated by Fridley (2009).

We should notice that the priors specified for the parameters are able to greatly influence the posterior model probabilities. If not prior sensitivity analysis is performed, one should be really careful when specifying them, so as to represent as closely as possible any a priori information. In our example considering the previous issue and also for comparison to the aforementioned in Chapter 5 search algorithms (SSVS,KM,GVS) and the next chapter's (BMA, BMS, BAS) on the same simulated data from linear regression model, we will retain a constant prior precision equal to the mean precision of the standardized variables acquired form a pilot run of the full model divided by the length of the dataset, in one way creating a Unit Information Prior, stating that we have information for just one point. The prior even though takes a very small part of the data into consideration is considered empirical. The prior mean of the parameters is by default assumed to be equal to 0.

Tests on jump interface have been conducted by Lunn et al. (2009, 2006) and Gimenez et al. (2009)

6.3.1 Clyde's Simulated Data Scheme

Following (Clyde et al. 2011) let us consider the following regression parameters for data generation. β being the effects of the generated data. So we assume that β_0=2, β=(-0.48,8.72,-1.76,-1.87,0,0,0,0,0*,0,0,0,0,0,0) * The 9^{th} covariate is assumed to be correlated with the 2^{nd} with $\rho_{2,9} = 0.99$, in difference with Clyde's approach we do not include x9 in the linear predictor, while the dispersion parameter is assumed to be equal to ϕ=1.

We will provide a presentation of the Jump Interface on the simulated data. Two (2) chains were initiated and the posterior quantities were plotted to check if the chain mix sufficiently (figure 6.3), in a way indicating that convergence has been reached, while each separate parameter used by the jump interface add-on was graphically reported to be mixing sufficiently (chain crossing).

The black regions of figure 6.3 corresponds to the coincidence between variables and models between the two chains initiated. We can notice that for the predictors that are most highly probable the chains mixed sufficiently well, while for the less probable ones the mixing was not of the same magnitude.

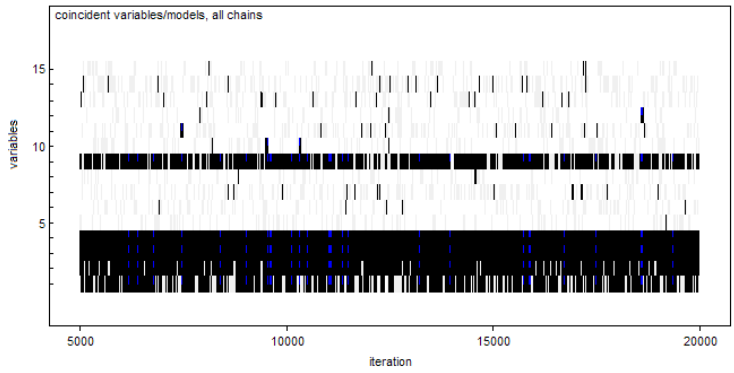


Figure 6.3: Jump Interface model mixing representation, for the two chains initiated for the linear simulated regression predictors using Jump Interface in WinBUGS.

The Jump Interface had similar characteristics to the previous chapter as far as the prior on parameters and on models are concerned, the only differences being that by default Jump Interface assumes a multivariate normal prior with prior mean equal zero and prior variance equal to a constant variance of all the parameters. This variance we chose to be acquired be a pilot run of the full model. The program did well identifying the predictors that were designed to be important but also the ones { x2 , x9 } that were assumed having a correlation of magnitude $\rho = 0.99$. Covariates { x1 , x2 , x3 , x4 , x9 } were included in the MAP model, while being part of the majority of the most probable models, indicating that the Jump Interface does well in both identifying the PIPs of the predictors while not performing so conveniently as far as correlation is concerned, under our initial assumptions for parameter / model priors, being itself kind of restrictive by default. For an analytical report and a comparison to GVS, refer to tables (6.1 & 6.2). Both techniques conclude to similar results,

Predictor	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
rjMCMC	0.85	0.97	1.00	1.00	0.10	0.10	0.16	0.07	0.91	0.13	0.15	0.08	0.15	0.18	0.08
GVS	0.93	0.97	1.00	1.00	0.11	0.08	0.09	0.07	0.88	0.08	0.08	0.12	0.11	0.17	0.09

Table 6.1: Posterior inclusion probabilities of the linear simulated dataset predictors using the Jump Interface and GVS in WinBUGS . Iterations=15000, Burnin Period = 5000

Reversible Jump		GVS	
Model	PMP	Model	PMP
X1+X2+X3+X4+X9	0.210	X1+X2+X3+X4+X9	0.275
X1+X2+X3+X4+X9+X14	0.053	X1+X2+X3+X4	0.039
X1+X2+X3+X4+X9+X13	0.040	X1+X2+X3+X4+X12	0.036
X1+X2+X3+X4+X9+X11	0.039	X1+X2+X3+X4+X5+X9	0.034
X1+X2+X3+X4+X7+X9	0.036	X1+X2+X3+X4+X9+X10	0.025
X1+X2+X3+X4+X9+X10	0.028	X1+X2+X3+X4+X7+X9	0.024
X1+X2+X3+X4+X5+X9	0.026	X1+X2+X3+X4+X7+X8+X11	0.023
X1+X2+X3+X4+X9+X15	0.024	X1+X2+X3+X4+X8+X9	0.023
X2+X3+X4+X9	0.024	X2+X3+X4+X9	0.023

Table 6.2: Posterior model probabilities and corresponding Bayes Factors for the models with high probability versus the most probable model for the simulated data using Jump Interface and GVS in WinBUGS. Iterations=15000, Burnin Period = 5000

6.4 Conclusion

But why choosing rjMCMC over the others algorithms? Sisson (2005) claims that over all the approaches to trans-dimensional modeling proposed in recent years, the "reversible jump" method, appears to be the most effective. Is surely is the one that had received the most popularity in recent years. As we have already described, reversible jump has the ability to make jumps both on the parameter space and the model space, is closely related to other algorithms like MCC and MC^3 and it tends to perform better than the variable selection techniques of constant model space (SSVS, KM, GVS) when higher dimensions are considered.

In the last chapter of this thesis, we will briefly mention same characteristics of the 3 available packages in R for Bayesian Model Averaging and we will present a comparison of all the programs described until then, on Clyde's linear regression simulated data. We will also report which of the programs have trouble identifying the effect magnitude and possible correlation.

Chapter 7

Bayesian Variable Selection in R

7.1 Raftery's et al. "Bayesian Model Averaging"

For the purposes of this thesis just an implementation for comparison reasons will be attempted and described in section 7.4. The package created in R, can be used to apply model averaging both in linear models and generalized linear models, making use of a Laplace approximation on the model likelihoods and then use a leap and bound algorithm to select the most probable regressors. Therefore finding the most probable subsets without the need of examination of all possible subsets. It also applies the Occam's Window, a technique for reducing the summation in the denominator be lowering the mode state space.

It also contains a function to perform variable selection with simultaneous recognition of outliers by the use of MC^3 introduced by Madigan et al. (1995), MC^3 being a special case of the reversible jump MCMC described in chapter 6.

BMA is rather restrictive as for the priors that can be placed on parameters and on the models. We get only one choice to perform a type of BIC approximation on the parameter priors, resulting to a prior close to the Unit Information Prior, and is also not flexible regarding model priors, leaving us with the choice of equal a priori chances for each model. No possible options for application of Zellner's g-prior is available, which is considered one of the main drawbacks of this package in comparison to the other two.

It also provide the user with a graph containing each effect posterior distribution.

7.2 Feldkircher's and Zeugner's "Bayesian Model Averaging Library"

In comparison to the newest package BMS created by Amini & Parmeter (2011), performing Bayesian model averaging only in Linear Models but with a wide alternative ways of stating the parameters and model priors. BMS has a extented toolkit in comparison to BMA. It has the ability to apply a variety of different types of Zellner's g-prior on covariates : (1) Unit Information Prior, (2) Risk Inflation Criterion, (3) An asymptotic use of Hanna - Quinn Criterion, (4) BRIC, which converges to UIP or RIC according to the dataset, (5) Local Empirical Bayes and the (6) Hyper g-prior. Furthermore, has the ability to choose between several model prior options : (1) Uniform prior, (2) Binomial prior, (3)Beta-binomial prior and also allows to place custom prior on models.

When the number of the models under consideration is of high dimensions and make the algorithm unable to perform asymptotic computations, model sampling with the use of MCMC methods is applied. For over than 15 covariates the full enumeration of the model space is replaced with the application of the MC^3 sampler.

For graphical reporting, it returns a graph containing the model size distribution (both prior and posterior) and the posterior model probabilities comparing their exact computation to an initiated MCMC.

7.3 Clyde's Bayesian "Model Averaging using Bayesian Adaptive Sampling"

Clyde et al. (2011) in their paper introduced a different way of sampling the model space before applying model averaging. They created an innovative way to update the initial sampling probabilities of each regressor using the marginal sampling probabilities. By introducing binary

trees, attempts to sample models with higher posterior probabilities from the model space. An improved way for sampling without replacement is introduced with the use of binary trees. They claim that BAS succeeds in comparison to Simple random sampling without replacement (SRSWOR) to leave only 20% of the mass unsampled while on the same number of iterations SRSWOR leaves a huge 95% of mass unsampled. Which is easy to be understood, due to the completely random choice of models through the model space. Comparison with other methods is also presented. It should be pointed out that BAS uses the Median probability model in comparison to the Highest probability model as it is reported to be more preferable Barbieri & Berger (2004).

When the predictors placed in BAS overcome the number of 30, the package does not perform enumeration of the whole model space, but introduce algorithms for model searching instead: (1) "BAS", described just above or (2) an adaptive MCMC algorithm. The computation of the initial sampling probabilities can be of equal difficulty or computationally intensive as the enumeration of the model space. For that reason Clyde et al. (2011), proposed 3 alternative ways for computing the initial sampling probabilities, which will be then used be the BAS algorithm. It considers (1) uniform probabilities, (2) a p-value calibration or (3) a MCMC calibration.

BAS also contains an extended toolkit, with many choices on parameter and mode priors. The following choices is are available for parameter priors : (1) Akaike Information Criterion, (2) Bayesian Information Criterion, (3) the g-prior, (4) Zellner's Siow prior, (5) the hyper g-prior, (6) the hyper g-prior with a Laplace approximation on the prior of g, (7) Local empirical Bayes and (8) Global Empirical Bayes. While for the model priors it allows us to choose between : (1) Uniform, (2) Binomial and (3) Beta-binomial distributions.

Lastly, it provides plots for checking how the model fitted the data and an parameter inclusion probability plot.

7.4 General Comparisons

In this final section of this Thesis, we present the results of the aforementioned and analyzed in different degrees programs and raw codes both in R and in WinBUGS.

Firstly, we present a summary of their main characteristics in table (7.1), referring to the prior setup that was kept approximate similar for comparison efficiency.

The Bayesian Model Averaging packages return an estimate of the effect of each predictor which is reported alongside with the real value of the generated linear regression parameters and a classical estimate derived from the application of the "glm()" command on our simulated data in table (7.2). The posterior quantities each packages / code slightly differs, while all of them report Posterior Inclusion Probabilities (PIPs) which are presented and compared in this thesis final table (7.3).

Category	WinBUGS				R		
Program / Code	SSVS	KM	GVS	Jump	BMA	BMS	BAS
Type	HM	HM	HM	HM	BMA	BMA	BMA
Search Alg.	Gibbs	Gibbs	Gibbs	RJGibbs	LP/MC^3	BD / RJ	BAS/AMCMC
Model/γ Priors	Ber($\frac{1}{2}$)	Ber($\frac{1}{2}$)	Ber($\frac{1}{2}$)	Bin($\frac{1}{2}$)	Bin($\frac{1}{2}$)	(3) Bin($\frac{1}{2}$)	(4) Bin($\frac{1}{2}$)
Param. Priors	EUIP	EUIP	EUIP	EUIP*	BIC≈UIP	(6) UIP	(9) UIP
Summaries	PIPs - Posterior Means - Posterior Standard Deviations (PMPs)						
Plots	PIP linechart / density plot or PMP plot						

Table 7.1: Characteristics of the packages and code to be used. BD= Birth - Death, RJ=Reversible Jump,HM=Hierarhical Mixture Model, LP= Leaps and Bounds Algorithm, UIP=Unit Information Prior, EUIP=UIP Empirical Bayes Independent Prior, BIC=Bayesian Information Criterion, AMCMC= Adaptive MCMC, EUIP*= considering constant prior variance between covariates. The number in parenthesis refer to the number of the programs alternative choices available.

After controlling the prior distributions in such way, so as to be able to compare the different techniques, placing an Empirical Bayes parameter prior in Winbugs and using a Unit informa-

	Sim.	C. Estimate	BMA		BMS		BAS	
Predictor	Value	Full Model	PIP	P.Mean	PIP	P.Mean	PIP	P.Mean
X1*	-0.480	-0.7004	87.80	-0.6832	71.70	-0.5499	74.13	-0.5696
X2*	8.720	45.6643	77.10	12.8945	67.07	11.0107	68.92	11.0661
X3*	-1.760	-1.6365	100.00	-1.6349	100	-1.6241	100	-1.6245
X4*	-1.870	-1.8847	100.00	-1.8426	100	-1.8275	100	-1.8239
X5	0.000	-0.1727	3.50	-0.0038	7.73	-0.0069	9.55	-0.0091
X6	0.000	-0.0983	0.80	-0.0007	13.90	-0.0081	9.32	-0.0062
X7	0.000	0.2519	13.30	0.0390	18.63	0.0571	16.81	0.0511
X8	0.000	0.0061	0.00	0.0000	5.87	0.0033	9.32	0.0047
X9”	0.000	-37.3293	41.00	-4.3347	48.17	-2.5176	46.01	-2.5729
X10	0.000	0.1379	4.90	0.0090	14.63	0.0294	11.95	0.0242
X11	0.000	-0.3247	12.60	-0.0383	14.37	-0.0435	15.65	-0.0467
X12	0.000	-0.0632	0.00	0.0000	11.10	0.0023	9.20	0.0022
X13	0.000	0.2200	5.10	0.0104	8.00	0.0165	11.73	0.0242
X14	0.000	-0.1674	11.20	-0.0327	14.40	-0.0429	14.68	-0.0430
X15	0.000	0.0583	0.00	0.0000	10.80	-0.0018	9.16	-0.0014

Table 7.2: BMA - BMS - BAS Posterior Inclusion Probabilities for linear simulated data, Param. Prior = Empirical or UIP (g-prior, with g=n), Model Prior = Uniform, considering full enumeration of the model space. PIP=Posterior inclusion Probabilities, P.Mean=Posterior Mean. *Predictors chosen from Stepwise method with AIC step. ” x9 is correlated with x2 $\rho = 0.99$.

tion prior in R packages, while in all case and a Binomial / Bernoulli model prior with $p = 0.5$, stating our prior ignorance on which is the best model, we should comment on the findings. The first four programs, mentioned in table 7.1, performs Gibbs steps to estimate the quantities, while the last three implemented in R enumerate or approximate the space for a small number of covariates, which number, if surpassed a different MCMC scheme is initiated.

We noticed that all MCMC based techniques supported the model with parameters {x1, x2, x3, x4, x9 }, while all enumeration programs supported the model with parameter {x1, x2, x3, x4}. Excluding from their highest probability model predictor 9 which was created to be high correlated with predictor 2. Even with the default initial parameter and model priors of BAS

& BMS model {x1, x2, x3, x4, x9 }, was not supported as the highest probability model.

After applying a classical stepwise procedure using AIC criterion from the "Rcmdr" package in R to test the steps, one would conclude in a model with the following parameters : {x1, x2, x3, x4} having effects of really close magnitude to the initials. All three BMA methods succeed in giving 100% inclusion probability for regressors {x3,x4}, while giving a very high inclusion probability for the 1^{st} regressor. They also supported with a probability around 80% the 2^{nd} regressor, while giving a possibility of less than 50% that the 9^{th} regressor has to be part of the model. BMA package seems to be more supportive for the inclusion of x2 and the exclusion of x9, in comparison to the other packages.

Moreover, considering table 7.3, in which posterior summaries derived from the MCMC techniques implemented in WinBUGS on the linear simulated dataset are presented, one can notice that all four algorithms with success give a posterior inclusion probability of 1 to parameters {x3, x4}. We should also notice that GVS and rjMCMC give a higher PIP to the 1^{st} regressors compared to the packages that perform full enumeration in R. Finally in table 7.3 one can find the synopsis of the PIPs of all programs summarized in this thesis. It is clear that the MCMC algorithms fail to understand the ongoing correlation between regressor x2 and x9. All reporting high PIPs, except for KM which had a posterior change in its assumption, so as to be able to even report a PIP. KM under our initial assumptions failed to report a PIP and we changed the prior precision for the two correlated regressors so as to much their piloted precision, while for the other regressors we assumed a precision equal to the mean precision of all, while excluding {x2, x9} which had reported a very high precision in the pilot run. That is why KM seems to be working better even compared to the full enumeration programs.

GVS, KM and rjMCMC succeed in reporting a higher PIP for the first regressor. Covariates {x3,x4} are supported equally from all different implementations. Moreover, the BMA programs seems to be succeeding to understand the correlation between {x2, x9} placing a very low PIP on the ninth covariate. The zeros reported on BMA, maybe due to the aftermatch of the Occam's Window, excluding any model in which X13 and X19 where included and as a matter of fact placing zero PIP for those two covariates.

Running time of each procedure can be found in the Appendix's table 9.2

		WinBugs				R		
Predictors	Pr(>|t|)**	SSVS	KM	GVS	rjMCMC	BMA	BMS	BAS
X1*	0,0183	64,91	93,85	93,78	85,35	87,80	71,70	74,13
X2*	0,0722	63,59	92,15	80,95	97,05	77,10	67,07	68,92
X3*	<0.0001	100.00	100,00	100,00	100,00	100,00	100,00	100,00
X4*	<0.0001	99,92	100,00	100,00	100,00	100,00	100,00	100,00
X5	0,5558	12,06	9,32	9,24	9,99	3,50	7,73	9,55
X6	0,7248	10,66	7,20	6,72	9,81	0,80	13,90	9,32
X7	0,3365	11,55	8,32	8,39	15,53	13,30	18,63	16,81
X8	0,9783	11,45	8,42	8,24	7,29	0,00	5,87	9,32
X9”	0,1442	51,3	42,32	72,58	91,01	41,00	48,17	46,01
X10	0,5932	11,49	8,17	7,87	12,80	4,90	14,63	11,95
X11	0,2228	12,01	7,82	7,94	15,40	12,60	14,37	15,65
X12	0,7935	13,84	11,64	11,71	7,46	0,00	11,10	9,20
X13	0,4220	11,95	9,51	9,71	15,01	5,10	8,00	11,73
X14	0,5520	13,95	15,68	15,85	18,27	11,20	14,40	14,68
X15	0,8076	12,73	9,56	9,52	8,29	0,00	10,80	9,16

Table 7.3: Posterior Inclusion Probabilities of the linear simulated dataset regressors under the set of summarized programs in R / WinBUGS. C. p-value corresponds to the p-value returned from the classical regression of the full model. *Predictors chosen from Stepwise method with AIC step in R, ** Classical p-value, ” X9 is correlated with x2 $\rho = 0.99$. BMA -BMS - BAS consider full enumeration of the model space. Iterations = 20000, Burnin Period=20000. For more information see table 7.1

Chapter 8

Conclusion - Further Research

In this thesis we attempted an introduction to basic notions required for implementing Bayesian variable selection, presented a summary of the available methods for variable selection in both Classical and Bayesian statistics while analyzing more extensively the Bayesian variable selection methods. Programs and codes in both R and WinBugs were used, created or altered to make comparisons between existing ways of implementing Bayesian variable selection.

Most of the theoretical disadvantages of each method were noticed in the results, while R seems to contain the most compact and easy way to implement Bayesian variable selection for linear models using one of the three available packages. As far as Bayesian variable selection for Generalized linear models is concerned, only the BMA package can implement it in R, while by using code in WinBugs even though the running times of the procedures may increase, the way to program the needed code is much easier.

All of the compared methods had problems while dealing with correlated variables which is one topic that more research should be attempted. Moreover, the use of different parameter priors which have different properties while performing Bayesian variable selection is a rather hot topic in the scientific community. Different algorithmic scenes are introduced often, for example the EMVS algorithm that was introduced by Rockova and George in 2013 while this thesis was written.

Generalizations and close forms of the integrals used for Bayesian variable selection will surely make those techniques more appealing for usage by a broader audience.

Chapter 9

Appendix

9.1 Appendix A

9.1.1 Notations

$P(A_w)$	: The probability that A_w the wooden door is selected.
$P(A_p\|W)$	: The probability that A_p the plastic door is selected given that the wooden is already opened.
$f(y\|\theta)$	: or $l(\theta)$ the likelihood as a function of parameter θ.
$\pi(\theta)$	: Prior distribution of parameter θ.
$f(y,\theta)$	: Joint distribution of y and parameter θ can be written as $\pi(\theta,y)f(y)$.
$\pi(\theta\|y)$	: Posterior distribution of parameter θ given the data.
$f(y\|x)$	: Posterior predictive distribution of Y given observed x.
$f(y)$	: Marginal distribution that can be obtained by integrating out parameter θ from the joint distribution.
$\pi(\theta\|x) \propto \pi(x)f(x\|\theta)$	: Symbol of proportionality, the left part is proportional to the right up to constant.

$Y \sim Binomial(p)$	: Y is distributed as of a binomial distribution with probability equal to p
$I_{ij}(\theta)$	: Information matrix with columns equal i and rows equal j.
$J(\theta)$	: Jeffrey's prior on θ.
$Pr(C\|y)$	: Probability of C.
$pa(E) = \{B, C\}$	: Node E is a parent of B and C.
$ch(E) = \{F, G\}$	: Node E is a children of F and G.
$\alpha(\theta^{'}\|y)$	: Acceptance probability of proposed $\theta^{'}$.
$\hat{\mu}$	: Estimation of mean.
$\Re^p$	: Whole set of real numbers.
$E(Y)$	: Expectation of Y.
$V(Y)$	: Variance of Y.
$logit(p)$	: p is modeled using the logit link in generalized linear models.
$Cov(\theta^t, \theta^{t+1})$	: Correlation between simulated θs of distance equal to 1.
$Y \sim N(\beta, \tau)$	: Y is distributed as Normal with mean equal to β and precision equal to τ.
$\prod_{i=1}^{N}$	: The product of the part included each time.
$exp(X)$	: e=2.74 raised to a value X that can be a number or a result of a distribution.
$logL(\theta, y)$	: The loglikelihood of θ.
PO_{12}	: Posterior Odds of Model 1 compared to model 2.
$p(M_i\|y)$	: The posterior distribution of model M_i given the data.
$p(y\|M_i)$	: The likelihood given a particular model.

$\hat{f}_i(y)$: Direct estimators of the marginal likelihood $f(y)$.

η : The linear predictor of a particular modeling scheme.

O_j : The Odds of a particular model.

γ_j : Indicator variable, used for indicating whether a variable should be excluded from our model or not.

$\alpha_{m \to m'}$: Transition probability of reversible jump and any other algorithm that changes the model in every step of the algorithm.

9.2 Appendix B

9.2.1 Abbreviations

Statistical Abbreviations

AIC :	Akaike Information Criteria
AICc :	Generalized Akaike Information Criteria
AIS :	Adaptive Importance Sampling
AR :	AutoRegressive
BF :	Bayes Factor
BIC :	Bayesian Information Criteria
BOA :	Bayesian optimization algorithm
BUGS :	Bayes using Gibbs sampling
CC :	Carlin and Chib Method
CDF :	Cumulative distribution function
CLT :	Central Limit Theory
CODA :	Convergence Diagnosis and Output Analysis
DAG :	Directed Acyclic Graph
ESS :	Effective Sample Size
GLM :	Generalized Linear Models
GVS :	Gibbs Variable Selection
HM :	Harmonic Mean estimator
HPD :	Highest Posterior Density
IC :	Information Criteria
IID :	Independent Identically distributed
IndMH :	Independent Metropolis Hastings

KM :	Kuo and Mallick sampler
MAP :	Maximum a posteriori Probability Model
MC :	Monte Carlo
MCC :	Metropolized Carlin and Chib
MCE :	Monte Carlo Error
MCMC :	Markov Chain Monte Carlo
M-H :	Metropolis-Hastings
MLE :	Maximum Likelihood Estimation
MSE :	Mean Square Errors
MWG :	Metropolis within Gibb
NR :	Newton and Raphson estimator
PBD :	Posterior Bayes Density
PDF :	Probability distribution functions
PIP :	Posterior Inclusion Probabilities
PMP :	Posterior Model Probabilities
PoprjMCMC :	Population based reversible jump Markov Chain Monte Carlo
rjMCMC :	reversible jump Markov Chain Monte Carlo
RWM :	Random walk Metropolis
SAS :	Statistical Analysis System
SCC :	Subspace Carlin and Chib
SD :	Standard Deviation
SIR :	Sequential Importance Resampling
SLICE :	Slice sampler
SLLN :	Strong Law of Large Numbers
SSS :	Shotgun Stochastic Search
SSVS :	Stochastic Search Variable Selection
TIC :	Information Criteria

Other Abbreviations

ENIAC :	Electronic Numerical Integrator and Computer
MANIAC :	Mathematical Analyzer, Numerator, Integrator, and Computer
EHIS :	European Health Interview Survey
EU :	European Union
GNU :	Graphical N User
ISCED :	International Standard Classification of Education
SRHS :	Self-reported Health Status

9.3 Appendix C

9.3.1 R Packages

9.3.1.1 MCMCpack package

MCMCpack (Martin et al. 2011) is a R statistical package that perform Bayesian Inference with the use of posterior simulation for a variety of statistical problems. The majority of the simulations' code is compiled in C++ using the Scythe Statistical Library (Pemstein et al. 2007). The result of each function is a mcmc coda object, providing in that way summarize them easily using the "coda" package.

Website : http://mcmcpack.wustl.edu/

9.3.1.2 MCMCmnl

Markov Chain Monte Carlo for Multinomial Logistic Regression

Description

This function generates a sample from the posterior distribution of a multinomial logistic regression model using either a random walk Metropolis algorithm or a slice sampler. The user supplies data and priors, and a sample from the posterior distribution is returned as an mcmc object, which can be subsequently analyzed with functions provided in the coda package. Details of the function MCMCmnl used from package MCMCpack as presented by Martin et al. (2011) in the package's instruction manual.

Initial Values of MCMCmnl function

```
MCMCmnl(formula, baseline=NULL, data=NULL, burnin = 1000, mcmc = 10000,
       thin = 1, mcmc.method = c("IndMH", "RWM", "slice"), tune = 1, tdf=6,
       verbose = 0, beta.start = NA, b0 = 0, B0 = 0, ...)
```

Arguments

formula : Model formula. Individual specific covariates can be entered into the formula normally.

baseline : The baseline category of the response variable. In our case the response variable had two categories, so this parameter was not used.

data : The data frame used for the analysis. Each row of the data frame corresponds to an individual who answered a question

burnin : The number of burn-in iterations for the sampler.

mcmc : The number of iterations to run the sampler past burn-in.

thin : The thinning interval used in the simulation. The number of mcmc iterations must be divisible by this value.

mcmc.method : Can be set to either "IndMH" (default), "RWM", or "slice" to perform independent Metropolis-Hastings sampling, random walk Metropolis sampling or slice sampling respectively.

tune : Metropolis tuning parameter. Can be either a positive scalar or a k-vector, where k is the length of. The acceptance rate should be tune around 0.2 to 0.4 before making inference from the posterior sample.

...

beta.start : The starting value for the vector. This can either be a scalar or a column vector with dimension equal to the number of betas. If this takes a scalar value, then that value will serve as the starting value for all of the betas. The default value of NA uses the MLE of β as the starting value.

b0 : The prior mean of β.

B0 : The prior precision of β. Default value of 0 is equivalent to an improper uniform prior for beta.

This function returns a mcmc.object that then can be analyzed using the coda package which is described in brief in the next section.

9.3.2 CODA package

Details of the functions used from package CODA as presented by Plummer et al. (2006) in the package's instruction manual.

9.3.2.1 CODA diagnostics

geweke.diag

```
geweke.diag(x, frac1, frac2)
```

x : An object of type mcmc.
frac1 : fraction to use from beginning of chain, default value $A = \frac{1}{10}$.
frac2 : fraction to use from end of chain, default value $B = \frac{1}{2}$.

heidel.diag

```
heidel.diag(x, eps , pvalue )
```

x : An object of type mcmc.
eps1 : Target value for ratio of halfwidth to sample mean, default value $eps = \frac{1}{10}$.
pvalue : significance level for the test, default value $\alpha = 0.05$.

raftery.diag

```
raftery.diag(x, , r , s , converge.eps )
```

x : An object of type mcmc.

q : the quantile to be estimated, default value $q = 0.025$

r : margin of error of the estimate desired, default value $r = 0.005$

s : the chance to get an estimate in $(q - r, q + r)$, default value $s = 0.95$

convergence.eps : Desired precision for the time of convergence to be estimated, default value $Precision = 0.001$

effective.size

```
effectiveSize(x)
```

x : An object of type mcmc.

autocorr.diag

```
autocorr.diag(x)
```

x : An object of type mcmc.

9.3.2.2 CODA plots

traceplot

```
traceplot(x, smooth=F, ...)
```

x : An object of type mcmc.

smooth : Draw smooth line through the plot

... : For further graphical parameters one should refer to (par and plot)

densplot

```
densplot(x,  ...)
```

x : An object of type mcmc.

... : For further graphical parameters one should refer to (par and plot)

autocorr.plot

```
autocorr.plot(x, lag.max, ...)
```

x : An object of type mcmc.

lag.max :

... : For further graphical parameters one should refer to (par and plot)

geweke.plot

```
geweke.plot(x, frac1 = 0.1, frac2 = 0.5, nbins = 20,
         pvalue = 0.05, auto.layout = TRUE, ask, ...)
```

x : An object of type mcmc.

nbins : Number of segments.

pvalue : p-values to plot H_0 confidence limits.

frac1 : fraction to use from beginning of chain, default value $A = \frac{1}{10}$.

frac2 : fraction to use from end of chain, default value $B = \frac{1}{2}$.

... : For further graphical parameters one should refer to (par and plot)

9.3.2.3 summary.mcmc

```
summary(object, quantiles = c(0.025, 0.25, 0.5, 0.75, 0.975))
```

x : An object of type mcmc.

quantilies : an evaluated vector of quantiles for every mcmc chain.

9.3.2.4 plot.diagnostics

(User created function) The function was created for simultaneous plotting of graphs required for graphical diagnoses of convergence.
Depends : R(≥2.15.3), coda

```
plot.diagnostics(x=NULL, burnin=1000, iter=10000, thin=1,
        step.dens=20, step.acf=250, ind=c("tra","erg","den","acf"))
```

x : An object of type mcmc.
burnin : The number of burn-in iterations of the mcmc chain.
iter : The number of iterations of the mcmc chain.
thin : The thinning interval of the mcmc chain.
step.dens : The step of the density plot.
step.acf : The lag of the autocorrelation plot (can be adjusted to find the best thinning interval value).
ind : An indicator that can take each one of the next values and all of them together ("tra","erg","den","acf"), corresponding to traceplots, ergodic.mean plots, density plots and autocorrelation plots, respectively.

A function creating a summary of plot diagnostics for objects in coda mcmc format. The function requires the burnin period to be known. Can plot only part of the available graphs. Three of the graphs plotted for each of the parameters are from the CODA package described above. Among the diagnostics plot, a density plot is created for the researcher to check more efficiently the posterior information through graphs.

See code in attached folder (R Code//plot.diagnostics)

9.3.3 xtable package

Details of the functions used from package CODA as presented by Dahl (n.d.*b*) in the package's instruction manual.

```
xtable(x, caption = NULL, label = NULL, align = NULL, digits = NULL,
display = NULL, ...)
```

x : An R object of class found among methods, it usually works for every matrix / data frame objects.
label : Number of segments.
align : A character vector equal to the number of columns of the matrix specified. Choices can be between "l", "r", and "c" corresponding to left, right, and center alignment, respectively.
digit : A vector of length equal to one, or to the number of columns of the matrix specified. It determines the number of digits to be displayed.
frac2 : fraction to use from end of chain, default value $B = \frac{1}{2}$.

The result can be then be transfered without further intervention to LaTeXto be printed.

9.3.3.1 xtableMCMCsummaries

See code in attached folder (R Code//xtableMCMCsummaries)

9.3.3.2 xtableMCMCdiagnostics

See code in attached folder (R Code//xtableMCMCdiagnostics)

9.3.4 Other R functions

9.3.4.1 ContourPlots

(User created function) A function that creates the contour plot for the joint distribution of two variables e.g β_0 vs. β_1.

```
contourplot(x, )
```

x : A matrix of dim with two columns accounting for the first and the second variable plotted.
xlim :A vector of size two, responding to the limits of x-axis to be plotted.
ylim :A vector of size two, responding to the limits of y-axis to be plotted.
sepx :A scalar that corresponds to the detail of the plotted x-axis of the contour plot.
sepy :A scalar that corresponds to the detail of the plotted y-axis of the contour plot.
conf.int : The confidence intervals to be plotted, default value is (0, 0.5, 0.75, 0.9, 0.95, 0.99).

9.3.4.2 writeDatafileR

Details of the function as presented by (Dahl n.d.*a*) in his revised script.

```
writeDatafileR(x, towhere, fill )
```

x : Either a data frame or else a list consisting of scalars, vectors, arrays or data frames. The values should all be numeric for the function to not return an error.
towhere : Directory and file to receive the output, default value is "toWinBUGS.txt"
fill : If numeric, it represents the number of columns for output, while if FALSE the output will be produced on one line. Default value, fill=80.

See code along with comments in the attached folder (R Code//writeDatafileR)

9.3.4.3 erg.mean

(User created function) See code along with comments in the attached folder (R Code//erg.mean)

9.3.4.4 Extra tables/graphs

S. Scheme	User	System	Elapsed
IndMH (1)	14.86	0.02	16.37
IndMH (2)	14.68	0.00	14.98
IndMH (3)	14.53	0.02	14.79
RWM (1)	14.65	0.00	15.02
RWM (2)	14.81	0.05	15.24
RWM (3)	15.11	0.08	15.47
Slice (1)	230.81	0.31	236.82
Slice (2)	272.00	0.31	296.27
Slice (3)	249.36	0.24	251.65

Table 9.1: Running (System, user, elapsed) times of MCMCmnl of MCMCpack under all sampling schemes for 8000 iterations in seconds

Program (Changes)	time
SSVS	43
KM	55
GVS	298
rjMCMC	337
BMA	0.17
BMS	0.14
BAS	0.71

Table 9.2: Running (System, user, elapsed) times for WinBUGS (SSVS - KM - GVS - rjMCMC) and R (BMA - BMS - BAS) linear regression variable selection for Clyde's simulated dataset in seconds. R programs performed full enumeration, WinBUGS programs were measured for 20000 iterations.

Information Criteria		
Akaike (AIC) :	$c_n(k) = (3.1.1) + 2p$	1974
Takeuchi :	$c_n(k) = (3.1.1) + 2[\hat{tf}\{J(\hat{\theta})[I(\hat{\theta}_k)]^{-1}\}]]$	1976
Schwarz (BIC) :	$c_n(k) = (3.1.1) + plog(n)$	1978
Hanna-quinn :	$c_n(k) = (3.1.1) + c\frac{log(log(n))}{n},\ c \geq 2$	1979
AICc :	$c_n(k) = (3.1.1) + plog(n) + \frac{2p(p+1)}{n-p-1}$	1989

Table 9.3: Information on five frequently used Information Criteria placed according to date of development. (3.1.1) part corresponds to $-2logL(\hat{\theta}|y)$

What is your highest education degree?

ISCED Level	ISCED 0	ISCED 1	ISCED 2	ISCED 3	ISCED 4	ISCED 5	ISCED 6
Description	no formal education or below	primary education	lower secondary education	upper secondary education	post-secondary but non-tertiary education	first stage of tertiary education	second stage of tertiary education

Table 9.4: Number of answers in each category of the variable " How is your Health in general? "" according to sex of the responder (Male/Female) and the marginal distribution of each of the two variables.

9.3.5 Convergence Tests

Here you should find a section for the application of different diagnostic tests on MCMC convergence and autocorrelation. The techniques are briefly explained most of the times using a graphical image and applied in chapter's 3 example, using CODA (see 9.3.2.1).

9.3.5.1 Geweke Diagnostic

Geweke diagnostic compares two different time intervals of the created chain where the sampled parameter is located. So, when the mean values of the parameter in two different time intervals are close, one can assume that those two samples come from the same distribution. It is most frequent one to compare the last part of the chain which is supposed to have reach convergence, over one narrower interval coming usually from the beginning of the chain.

In coda's implementation of the Geweke diagnostic the windows sizes are $A = \frac{n}{10}$ and $B = \frac{n}{2}$ for window A and B (see figure 9.1), which hold the 10% of the first part of the chain and the last half. These sizes are also suggested by Geweke et al. (1991). First, the statistic is applied to the whole chain, if the Z-statistic is outside the 95% confidence interval, we continue to apply the diagnostic after discarding 10%, 20%, 30% and 40%. If during the last test the Z-statistic is still not inside the 95% confidence interval, the chain is reported as "failed to reach convergence". One of the disadvantages of Geweke's method is that it is sensitive to the specification of the windows.

In table 9.5 it seems that Geweke's diagnostic is not showing an unusual scheme. The values of the test are within the ranges of the N(0,1). Except for those betas, there is no evidence provided by this test that the chains have not converge. Caution, one should also take into account that Geweke's diagnostic performs many sequential hypothesis testing, it should be used with attention due to possible increasing of type II error probability. Both Geweke's and Heidelberger's diagnostic require sequential testing, so for the tests to be realistic, a correction should be applied.

In figure 9.2 one of the corresponding plots is provided making even clearer what we already have mentioned. The 3^{rd} sampling scheme of RWM algorithms seems to need some time to

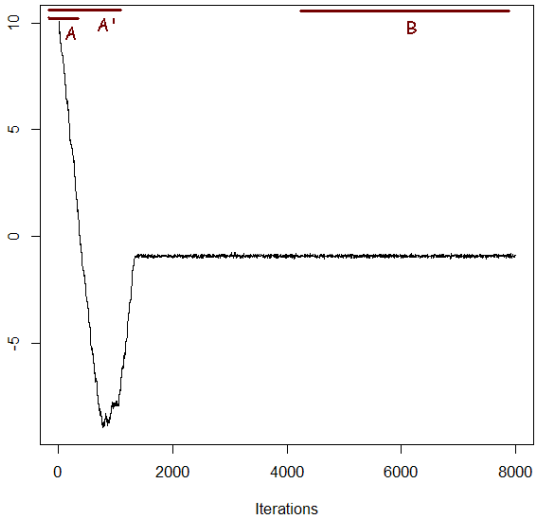


Figure 9.1: Geweke-Brooks plot for the 3^{rd} sampling scheme of the RWM algorithm, showing how the Geweke diagnostic chooses and testes window A with window B and then widening window $A \rightarrow A^{'}$ to repeat the test.

Set	β_0	β_1
IndMH (1)	1.1578	-0.7828
IndMH (2)	1.2210	-0.9582
IndMH (3)	0.4823	-0.1781
RWM (1)	1.0750	-0.8734
RWM (2)	1.3445	-0.3525
RWM (3)	1.3396	-0.3476
Slice (1)	-1.8227	1.9277
Slice (2)	1.4853	-1.3074
Slice (3)	1.2305	-1.4272

Table 9.5: Comparative table of Geweke Diagnostic for three algorithms used given a particular sampling scheme, see table 3.3, for windows sizes of $A = \frac{n}{5}, B = \frac{n}{2}$

reach convergence and then stabilizes within the limits of Geweke's Z-score.

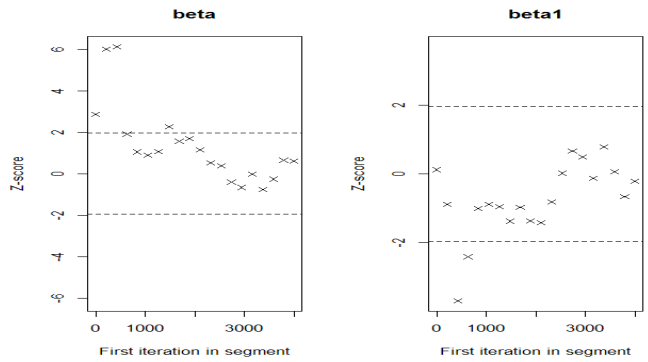


Figure 9.2: Geweke-Brooks plot for the 3^{rd} sampling scheme of the RWM algorithm, showing what happens to Geweke's Z-score when repeatedly bigger number of iterations are being discarded from the beginning of the chain, the plot never discards more than half of the chain.

9.3.5.2 Heidelberger-Welch Diagnostic

The diagnostic of Heidelberger & Welch (1983) is actually based on the assumption that a weakly stationary process is considered when the chain has already converge. Such a stationary process has the property that, if X_i is defined as the i^{th} iteration in the sequence, the mean function $E[X_i]$ is constant in time and $Cov(\theta^i, \theta^{i+s})$ does not depend on i but only on how large s is. This is a logical assumption since our sequence is generated by a Markov chain and therefore should satisfies full stationarity.

In coda's implementation of Heidelberger diagnostic, if the null hypothesis of stationarity is rejected, the test is repeated with 10%, 20% ... until stationarity is reached or at least 50% of the chain is discarded. If stationarity has not been reached and half of the chain is discarded the outcome results in a message of type "the chain failed to reach stationarity", indicating that a longer run may be needed. The Cramer-von-Mises (Anderson 1962) statistic is used to test the H_0 that the sampled values come from a distribution that have reached stationarity.

As we notice in tables 9.6 and 9.7, the majority of the sampling schemes were not rejected by this test. The algorithm that did not perform well in comparison to the others is the

Set	Stationarity test	Start iteration	Cramer-von -Mises statistic	Halfwidth test	Mean	Halfwidth
IndMH (1)	Pass	1.0	0.1587	Pass	0.9408	0.0047
IndMH (2)	Pass	1.0	0.1587	Pass	0.9408	0.0047
IndMH (3)	Pass	1501.0	0.0581	Pass	0.9502	0.0067
RWM (1)	Pass	1.0	0.8458	Pass	0.9405	0.0078
RWM (2)	Pass	1.0	0.3007	Pass	0.9424	0.0075
RWM (3)	Pass	1.0	0.5073	Pass	0.9468	0.0083
Slice (1)	Pass	1.0	0.8644	Pass	0.9471	0.0088
Slice (2)	Pass	1.0	0.7888	Pass	0.9387	0.0089
Slice (3)	Pass	1.0	0.8610	Pass	0.9431	0.0090

Table 9.6: Comparative table of Heidelberger Diagnostic ($\beta(0)$)for the three algorithms used given a particular sampling scheme (see table 3.3)

Set	Stationarity test	Start iteration	Cramer-von -Mises statistic	Halfwidth test	Mean	Halfwidth
IndMH (1)	Pass	1.0	0.0942	Pass	-0.9175	0.0018
IndMH (2)	Pass	1.0	0.0942	Pass	-0.9175	0.0018
IndMH (3)	Fail	–	0.0369	Fail	–	–
RWM (1)	Pass	1.0	0.9013	Pass	-0.9180	0.0030
RWM (2)	Pass	1.0	0.2157	Pass	-0.9183	0.0030
RWM (3)	Pass	1.0	0.4423	Pass	-0.9194	0.0030
Slice (1)	Pass	1.0	0.8719	Pass	-0.9198	0.0034
Slice (2)	Pass	1.0	0.8726	Pass	-0.9165	0.0034
Slice (3)	Pass	1.0	0.9046	Pass	-0.9178	0.0030

Table 9.7: Comparative table of Heidelberger Diagnostic ($\beta(1)$)for the three algorithms used given a particular sampling scheme (see table 3.3)

IndMH from which the two sampling schemes were on the verge of rejection and the third was rejected by Heidelberger diagnostic. This indicates that by following Heidelberger test, even after discarding 50% of the chain, convergence has not been reached for IndMH(3), and more iterations should be considered for the distribution to reach stationarity.

9.3.5.3 Raftery-Lewis Diagnostic

Suppose that we monitor the parameter x and we are interested in estimating the value of u such that $P(\theta(x) \leq u) = q$ for some quantile value q. We then choose which precision a and probability p we want to have on $\hat{u}$. In that way, we want $\hat{u}$ to belong in a interval $[u-a, u+a]$ for a probability p. Raftery & Lewis (1992*b*) presented a way to test convergence of the chain, a method that actually calculates the total length of the run and the estimated burn-in period of the chain, to estimate the aforementioned probability with accuracy a and probability p.

In coda's implementation of Raftery-Lewis diagnostic, the burn-in period, the length of the run and the dependence factor is provided. Values of the dependence factor larger than 5 indicate high autocorrelation which may appear ,among other reasons because of poor choice of starting values, high posterior correlations or the fact that the MCMC algorithm sticks in particular parts of the space. The output presents the number of iterations for convergence to be reached (N), the minimum number is kept constant and equal to 3476, the dependence factor indicating the distance of each independent iteration (I) and the needed number of burn-in period (M).

Set	M	N	N(Minimum)	I=(M+N)/Nmin (DF)
indMH (1)	18	20562	3746	5.5
indMH (2)	18	20562	3746	5.5
indMH (3)	30	27560	3746	7.4
rwMH (1)	17	18012	3746	4.8
rwMH (2)	18	19428	3746	5.2
rwMH (3)	17	18012	3746	4.8
Slice (1)	12	11842	3746	3.2
Slice (2)	15	20049	3746	5.3
Slice (3)	16	18916	3746	5.0

Table 9.8: Comparative table of Raftery - Lewis Diagnostic ($\beta(0)$) for the three algorithms used given a particular sampling scheme (see table 3.3)

Set	M	N	N(Minimum)	I=(M+N)/Nmin (DF)
indMH (1)	35	42210	3746	11.3
indMH (2)	35	42210	3746	11.3
indMH (3)	40	44500	3746	11.9
rwMH (1)	16	17023	3746	4.5
rwMH (2)	14	15560	3746	4.2
rwMH (3)	17	18115	3746	4.8
Slice (1)	11	11484	3746	3.1
Slice (2)	10	11784	3746	3.1
Slice (3)	14	14510	3746	3.9

Table 9.9: Comparative table of Raftery - Lewis Diagnostic ($\beta(1)$) for the three algorithms used given a particular sampling scheme (see table 3.3)

High autocorrelation can be seen in table 9.8 of parameter β_0, on all different sampling schemes. The dependence factor was around 5 and most of the times surpassing this upper limit. For both parameters β_0 and β_1 the slice sampler seems to be have the best performance. Accordingly to Raftery-Lewis diagnostic the chain produced under IndMH for parameter β_1 under any sampling scheme achieve an dependence factor of value around 11, the worst performance among all. Lastly, in general the 1^{st} and the 2^{nd} sampling scheme in given each algorithm

seem to be performing better. By using the Raftery – Lewis diagnostic we can now consider a thinning interval for every chain. This thinning interval could be equal to the dependence factor provided by the test. Lastly, it should be noted that Raftery-Lewis tends to underestimate the total burn-in period.

Having applied three diagnostic, until now, after having discarded the first 3000 iterations, only the 3^{rd} sampling scheme of the IndMH seems to not have reached stationarity.

9.3.5.4 The Effective Sample Size Diagnostic

Let suppose that we run our chain long enough, to obtain a particular set of samples (x). It is more than certain that one can ask himself. "After obtaining a set of samples (x), how much information do we really know about a parameter?". If correlation exists between successive samples from (x), then we can easily consider that our sample has not revealed as much information as one might expect if the samples were independent. This quantity of information is often estimated from the MCMC outputs and are most often seen in literature as the Effective Sample Size (ESS) (Kass et al. 1998). Therefore, ESS is a quantity that tries to estimate how many really independent sample are obtain from a set of samples (x). The ESS is a quantity that estimates the number of independent samples obtained from x.

The effective sample size is defined as $T\dfrac{1-p}{1+p}$, where T is the size of the MCMC sample and p the correlation between the IID sample. We consider 1^{st} degree correlations, assuming an AR(1) process. In our example T=5000.

Set	β_0	β_1
IndMH (1)	2166.1	2157.4
IndMH (2)	2166.1	2157.4
IndMH (3)	1176.2	1279.7
RWM (1)	695.9	687.0
RWM (2)	792.0	708.9
RWM (3)	636.7	724.8
Slice (1)	531.5	505.4
Slice (2)	522.2	516.7
Slice (3)	510.7	658.0

Table 9.10: Comparative table of Effective Size Diagnostic for the three algorithms used given a particular sampling scheme (see table 3.3)

The IndMH seems to produced more independent samples from the other two techniques used, while the Slice sampler produces the most autocorrelated. For the same number of set of samples (N=5000), the IndMH produces n={1176-2166}, the RWM produces n={792-636} and the Slice sampler n={510-658}. When having wrong prior information the IndMH performs even poorer in comparison to sampling schemes (1) and (2).

9.3.5.5 Autocorrelation Diagnostic

Set	Lag 0	Lag 1	Lag 5	Lag 10	Lag 50
IndMH (1)	1.0000	0.3772	[0.0030	-0.0202	-0.0391]
IndMH (2)	1.0000	0.3772	[0.0030	-0.0202	-0.0391]
IndMH (3)	1.0000	0.4978	[0.0907	0.0090	0.0042]
RWM (1)	1.0000	0.7556	0.2629	[0.0808	-0.0033]
RWM (2)	1.0000	0.7669	0.2781	[0.0689	-0.0117]
RWM (3)	1.0000	0.7691	0.2493	[0.0675	-0.0040]
Slice (1)	1.0000	0.8166	0.3281	0.1159	[-0.0344]
Slice (2)	1.0000	0.8108	0.3648	0.1481	[-0.0389]
Slice (3)	1.0000	0.8146	0.3312	[0.0465	0.0296]

Table 9.11: Comparative table of Autocorrelation Diagnostic ($\beta(0)$) for the three algorithms used given a particular sampling scheme (see table 3.3)

Set	Lag 0	Lag 1	Lag 5	Lag 10	Lag 50
IndMH (1)	1.0000	0.4025	[0.0130	-0.0262	-0.0351]
IndMH (2)	1.0000	0.4025	[0.0130	-0.0262	-0.0351]
IndMH (3)	1.0000	0.4998	[0.0814	0.0186	-0.0007]
RWM (1)	1.0000	0.7583	0.2591	[0.0776	-0.0005]
RWM (2)	1.0000	0.7649	0.2433	[0.0405	0.0026]
RWM (3)	1.0000	0.7664	0.2343	[0.0620	-0.0107]
Slice (1)	1.0000	0.8164	0.3365	0.1138	-0.0352]
Slice (2)	1.0000	0.8127	0.3538	0.1553	-0.0477]
Slice (3)	1.0000	0.8166	0.3362	[0.0475	0.0231]

Table 9.12: Comparative table of Autocorrelation Diagnostic ($\beta(1)$) for the three algorithms used given a particular sampling scheme (see table 3.3) , the brackets show the interval in which the autocorrelation has decreased in non significant values.

The autocorrelation is diminished in the IndMH scheme around lag 5, while for the rwM around lag 10, while for the Slice sampler one needs the biggest lag among those approach to create a sample with no autocorrelation. These results are in accordance to the results of the ESS as both count the autocorrelation of successive samples of x.

9.3.5.6 Diagnostics synopsis

We will briefly summarize the advantages and disadvantages of the above diagnostics and comment on their usage.

Geweke diagnostic seems not to reveal more information than a simple trace plot for the burn-in period, this is explainable as it is a simple transformation of the values plotted on a trace plot to a comparison to the mean location. However, it quantifies the optical examination of the traceplot into a z-score comparing means of two different parts of the chain, therefore being easy to be interpreted. How the appropriate window is chosen requires a minimal examination of the plot. (in our case $A = \frac{1}{5}$ was chosen.)

The Heidelberger-Welch diagnostic is a rather more complete way investigating properties of the chain's trace than the Geweke diagnostic. We compare the means and the variation between two different part of the chain, providing a more comprehensive, but more difficult to be interpreted diagnostic than the previous.

While evaluating the diagnostic of Raftery and Lewis we noticed that the estimations for the burn-in given were often too small compared to the burn-ins suggested by examining the traceplots of the whole chain. Another fact to take into consideration when applying diagnostics upon chains is the following. Do one need to see how precise are his estimates on the true B, B being the burn-in iterations, or which B gives the most precise estimate of the parameter he is interested for.

9.4 Variable Selection Functions

9.4.1 SSVS, KM & GVS

The algorithm presented below is a readjusted version, for the EHIS2009 data, of the Logistic Regression Model Selection in WinBUGS, written by J. Ntzoufras for BUGS software in 1997.

```
#####################################################################
# Description of the variables used in OpenBUGS SSVS, KM, GVS code.
 N = 4850,  # number of binomial responses
 Q= 62,  # Number of models under consideration for 5 covariates 2^5
 include,    # conditional prior probability for g
 pmodel,    # model indicator vector
 model,        # code of model
 beta,       # model coefficients
 gamma;     # term indicator vector
 beta.prior.mean,    # proposal mean used in pseudoprior
 beta.prior.sd,      # proposal stand.deviation used in pseudoprior
 beta.prior.mean, # prior mean for b depending on g
 tau,     # model coefficients precision
#####################################################################
# Calculation of the likelihood and model configuration,
# in the case of SSVS the gammas are not included in the likelihood.
for (i in 1:n){
HS01[i]~dbern(p[i])
logit(p[i])<-gamma0*beta0+
gamma[1]*beta[1]*age[i]+
gamma[2]*beta[2]*sex[i]+
gamma[3]*beta[3]*educat[i]+
gamma[4]*beta[4]*longill[i]+
gamma[5]*beta[5]*urban[i]}

# Prior for b model coefficients
# Mixture normal depending on current status of gamma.

for (i in 1:Q) {
#  GIBBS VARIABLE SELECTION proposal
# Proposals parameters equal to
```

```
#  the posterior mean and variance of the full model's pilot run.
#  ------------------------------------------------------------------
#  if(gamma[i]=1) then precision[i] = 1/SD^2*n
#      elseif(gamma[i]=0) then precision[i]=1/SD^2
%#  if(gamma[i]=1) mb[i]=0
%#      elseif(gamma[i]=0) then mb[i]=Mean from Pilot
# mb[i]=0
beta0 ~ dnorm( mb0, taub0)
mb0 <- 0
taub0 <- (gamma0/(pow(prop.sd.beta0 ,2)*n)
                          + (1-gamma0)/pow(prop.sd.beta0 ,2))
for (j in 1:5){
  beta[j] ~ dnorm( mb[j], taub[j])
  mb[j] <- 0
  taub[j] <- (gamma[j]/ (pow(prop.sd.beta[j] ,2)*n)
                          + (1-gamma[j])/ pow(prop.sd.beta[j] ,2))
}
#  STOCHASTIC SEARCH VARIABLE SELECTION proposal
# Automatic proposal using g=1000 as in SSVS
#  ------------------------------------------------------------------
# if(gamma[i]=1) then precision[i] = 1/SD^2*n
#     elseif(gamma[i]=0) then precision[i]=1/SD^2
# mb[i]=0
#
beta0 ~ dnorm( mb0, taub0)
mb0 <- 0
taub0 <- (gamma0/(pow(prop.sd.beta0 ,2)*n)
                     + (1-gamma0) / pow(prop.sd.beta0 ,2))
for (j in 1:5){
  beta[j] ~ dnorm( mb[j], taub[j])
  mb[j] <-  0
```

```
  taub[j] <- (gamma[j]/(pow(prop.sd.beta[j] ,2)*n)
                    + (1-gamma[j]) / pow(prop.sd.beta[j] ,2))
}
#  KUO \& MALLICK SAMPLER  proposal
#  P(beta,gamma)=p(beta)*p(gamma),
#  assuming independence between the variable indicators and the parameters.
#  precision[i]=1/SD^2*n   and mb[i]=0
#
  tau00<-1/(pow(prop.sd.beta0 ,2)*n)
  beta0 ~ dnorm( 0, tau00)
for (j in 1:5){
tau01[j]<-1/(pow(prop.sd.beta[j] ,2)*n)
  beta[j] ~ dnorm( 0, tau01[j])
}

# ------------------------------------------------------------
# Defining prior information for gamma[i].
# Allow models to have the same initiative probability of inclusion.
#
gamma0 ~ dbern(1)
for (j in 1:Q){ gamma[j]~dbern(0.5) }
# Or place a prior on p[1:5]~beta(1,1)
# gamma0<- dbern(1)
# for (j in 1:Q){ gamma[j]~dbern(p) }
# for (j in 1:Q){ p[j]~dbeta(1,1) }
}
# Construction of model Indicator
   for (j in 1:5){ mindex[j] <- pow(2,j)}
   model <- inprod( gamma[], mindex[] )

# gamma0 is omitted as it is always included in the model.
```

```
#
# Posterior Model probabilities for all
# possible Model combinations Nmodel = Sigma 2^mi
   for( m in 1:Nmodel){pmodel[m]<-equals(m,model)}
```

9.4.2 reversible jump via OpenBUGS jump add-on

The OpenBUGS code for performing reversible jump in linear simulated data is presented in this section. For the code applied in digital BUGS file, please refer to the attached folder (OpenBUGS code//Jump).

```
# Description of the variables used in OpenBUGS jump extension code.
n=4850 # the number of responses.
Q=5 # number of covariates.
Y #  response variable of the model.
psi # a deterministic function of fixed dimension
   that contains each time an unknown number of covariates.
k # number of betas currently in the model.
X # the vector of x-values.
id # variable indicating a particular model.
pred # prediction with the use of specific set of linear predictors.
tau # prior precision
X.pred #
effect #

model {
for (i in 1:n) {
y[i]~ dnorm(psi[i],tau)
X[i, 1] <- x1[i]
X[i, 2] <- x2[i]
          ...
```

```
X[i, 15] <- x15[i]
}
tau~dgamma(0.01,0.01)
#
psi[1:n] <- jump.lin.pred(X[1:n, 1:Q], k, beta.prec)
id[1] <- jump.model.id(psi[1:n])
beta.prec <- 1 / (SD^2*n)
k ~ dbin(0.5, Q)
# Or place a prior on p[]~beta(1,1)
# k ~ dbin(p, Q)
# p~dbeta(1,1)
#
pred[1:(Q + 1)] <- jump.lin.pred.pred(psi[1:n], X.pred[1:(Q + 1), 1:Q])
#
for (i in 1:Q) {
X.pred[i, i] <- 1
for (j in 1:(i - 1)) {X.pred[i, j] <- 0}
for (j in (i + 1):Q) {X.pred[i, j] <- 0}
X.pred[(Q + 1), i] <- 0
effect[i] <- pred[i] - pred[Q + 1]
}}
```

9.4.3 BMA - BMS - BAS

Sample code for those three to be included in the digital version of this thesis.

Bibliography

Amini, S. M. & Parmeter, C. F. (2011), 'Bayesian model averaging in r', *Journal of Economic and Social Measurement* **36**(4), 253–287.

Anderson, T. (1962), 'On the distribution of the two-sample cramer-von mises criterion', *The Annals of Mathematical Statistics* pp. 1148–1159.

Austin, P. C. (2008), 'Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backwards variable elimination: a simulation study', *Journal of Clinical Epidemiology.* **61**, 1009–1017.

Barber, D. (2012), *Bayesian reasoning and machine learning*, Cambridge University Press.

Barbieri, M. M. & Berger, J. O. (2004), 'Optimal predictive model selection', *The Annals of Statistics* **32**(3), 870–897.

Bartlett, M. (n.d.), 'S.(1957). a comment on dv lindley's statistical paradox', *Biometrika* **44**, 533–534.

Bayes, M. & Price, M. (1763), 'An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs', *Philosophical Transactions (1683-1775)* pp. 370–418.

Berger, J. O. (2000), 'Bayesian analysis: A look at today and thoughts of tomorrow', *Journal of the American Statistical Association* **95**(452), 1269–1276.

Bernoulli, J. (1713), *Ars conjectandi*, Impensis Thurnisiorum, fratrum.

Bolstad, W. M. (2007), *Introduction to Bayesian statistics*, Wiley-Interscience.

Box, G. (n.d.), 'Ep and tiao, g. c. 1973', *Bayesian inference in statistical analysis* pp. 112–139.

Brown, P. J., Vannucci, M. & Fearn, T. (1998), 'Multivariate bayesian variable selection and prediction', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**(3), 627–641.

Carlin, B. P. & Chib, S. (1995), 'Bayesian model choice via markov chain monte carlo methods', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 473–484.

Carlin, B. P. & Louis, T. A. (1997), 'Bayes and empirical bayes methods for data analysis', *Statistics and Computing* **7**(2), 153–154.

Carlin, B. P. & Louis, T. A. (2011), *Bayesian methods for data analysis*, Vol. 78, Chapman and Hall/CRC.

Cheng, J. & Druzdzel, M. J. (2000), 'Ais-bn: An adaptive importance sampling algorithm for evidential reasoning in large bayesian networks', *Journal of Artificial Intelligence Research* **13**(1), 155–188.

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

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

Clyde, M. (1999), '[bayesian model averaging: A tutorial]: Comment', *Statistical Science* **14**(4), 401–404.

Clyde, M. A., Ghosh, J. & Littman, M. L. (2011), 'Bayesian adaptive sampling for variable selection and model averaging', *Journal of Computational and Graphical Statistics* **20**(1).

Congdon, P. (2007), *Bayesian statistical modelling*, Vol. 704, Wiley.

Copas, J. (1984), 'Discussion of dr miller's paper', *Journal of the Royal Statistical Society A* pp. 410–412.

Cowles, M. K. & Carlin, B. P. (1996), 'Markov chain monte carlo convergence diagnostics: a comparative review', *Journal of the American Statistical Association* **91**(434), 883–904.

Cui, W. & George, E. I. (2008), 'Empirical bayes vs. fully bayes variable selection', *Journal of Statistical Planning and Inference* **138**(4), 888–900.

Dahl, D. B. (n.d.*a*), 'writedatafiler : To write a r file in a form compatible with winbugs.'.
URL: *http://www.public.iastate.edu/ alicia/stat544/writeDatafileR.txt*

Dahl, D. B. (n.d.*b*), 'xtable : Export tables to latex or html'.
URL: *http://xtable.r-forge.r-project.org/*

Damlen, P., Wakefield, J. & Walker, S. (1999), 'Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(2), 331–344.

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2000), 'Bayesian variable selection using the gibbs sampler', *BIOSTATISTICS-BASEL-* **5**, 273–286.

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2002), 'On bayesian model and variable selection using mcmc', *Statistics and Computing* **12**(1), 27–36.

Devroye, L. (1986), *Non-uniform random variate generation*, Springer-Verlag.
URL: *http://books.google.gr/books?id=mEw_AQAAIAAJ*

Downs, O. B., MacKay, D. J., Lee, D. D. et al. (2000), 'The nonnegative boltzmann machine', *Advances in Neural Information Processing Systems* **12**, 428–434.

Edwards, A. (1986), 'Is the reference in hartley (1749) to bayesian inference?', *The American Statistician* **40**(2), 109–110.

Efron, B. (1986), 'Why isn't everyone a bayesian?', *The American Statistician* **40**(1), 1–5.

Efroymson, M. (1960), 'Multiple regression analysis, mathematical methods for digital computers', *Statistics and Computing* **1**, 191–203.

Euler, L. (1736), 'Solutio probelmatis ad geometriam situs pertinentis', *Commentarii Academiae Scientiarum Imperialis Petropolitanae* **8**, 128–140.

Fernandez, C., Ley, E. & Steel, M. F. (2001), 'Model uncertainty in cross-country growth regressions', *Journal of applied Econometrics* **16**(5), 563–576.

Fink, D. (1997), 'A compendium of conjugate priors'.

Foster, D. P. & George, E. I. (1994), 'The risk inflation criterion for multiple regression', *The Annals of Statistics* pp. 1947–1975.

Fouskakis, D., Ntzoufras, I. & Draper, D. (2009), 'Population-based reversible jump markov chain monte carlo methods for bayesian variable selection and evaluation under cost limit restrictions', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58**(3), 383–403.

Fridley, B. L. (2009), 'Bayesian variable and model selection methods for genetic association studies', *Genetic epidemiology* **33**(1), 27–37.

Friel, N. & Pettit, A. (2008), 'Marginal likelihood estimation via power posteriors', *Journal of Royal Statistical Society* (770), 589–607.

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

Gamerman, D. & Lopes, H. F. (2006), *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*, Vol. 68, Chapman & Hall/CRC.

Gelfand, A. E. & Smith, A. F. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American statistical association* **85**(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003), *Bayesian Data Analysis, (Chapman & Hall/CRC Texts in Statistical Science)*, Chapman and Hall/CRC.

Gelman, A. & Meng, X.-L. (1998), 'Simulating normalizing constants: From importance sampling to bridge sampling to path sampling', *Statistical Science* pp. 163–185.

Gelman, A., Meng, X.-L. & Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica Sinica* **6**, 733–759.

Gelman, A., Roberts, G. & Gilks, W. (1996), 'Efficient metropolis jumping hules', *Bayesian statistics* **5**, 599–608.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

George, E. & Foster, D. P. (2000), 'Calibration and empirical bayes variable selection', *Biometrika* **87**(4), 731–747.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

George, E. I. & McCulloch, R. E. (1997), 'Approaches for bayesian variable selection', *Statistica sinica* **7**, 339–374.

Geweke, J. et al. (1991), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Federal Reserve Bank of Minneapolis, Research Department.

Gilks, W. R., Best, N. & Tan, K. (1995), 'Adaptive rejection metropolis sampling within gibbs sampling', *Applied Statistics* pp. 455–472.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996), *Markov chain Monte Carlo in practice*, Vol. 2, Chapman & Hall/CRC.

Gilks, W. R. & Wild, P. (1992), 'Adaptive rejection sampling for gibbs sampling', *Applied Statistics* pp. 337–348.

Gimenez, O., Bonner, S. J., King, R., Parker, R. A., Brooks, S. P., Jamieson, L. E., Grosbois, V., Morgan, B. J. & Thomas, L. (2009), Winbugs for population ecologists: Bayesian modeling using markov chain monte carlo methods, *in* 'Modeling demographic processes in marked populations', Springer, pp. 883–915.

Godsill, S. J. (2001), 'On the relationship between markov chain monte carlo methods for model uncertainty', *Journal of Computational and Graphical Statistics* **10**(2), 230–248.

Gordon, N. J., Salmond, D. J. & Smith, A. F. (1993), Novel approach to nonlinear/non-gaussian bayesian state estimation, *in* 'IEE Proceedings F (Radar and Signal Processing)', Vol. 140, IET, pp. 107–113.

Gordon, N., S. D. & Smith, A. F. M. (1993), 'Non-linear / non-gaussian bayesian state estimation. iee proceedings of radar and signal processing', *Proceedings of Radar and Signal Processing* **140**(2), 107–113.

Gössl, C. & Küchenhoff, H. (1999), 'Bayesian analysis of logistic regression with an unknown change point'.

Gössl, C. & Kuechenhoff, H. (2001), 'Bayesian analysis of logistic regression with an unknown change point and covariate measurement error', *Statistics in medicine* **20**(20), 3109–3121.

Green, P. J. (1995), 'Reversible jump markov chain monte carlo computation and bayesian model determination', *Biometrika* **82**(4), 711–732.

Green, P. J. (2003), 'Trans-dimensional markov chain monte carlo'.

Guindani, M. (Spring 2008), 'Statistical computing'.

Hammersley, J. M. & Clifford, P. (1968), 'Markov fields on finite graphs and lattices'.

Hammersley, J. M.; Clifford, P. (1971), 'Markov fields on finite graphs and lattices', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* .

Han, C. & Carlin, B. P. (2000), 'Mcmc methods for computing bayes factors: A comparative review', *Biometrika* **82**(4), 711–732.

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

Hans, C., Dobra, A. & West, M. (2007), 'Shotgun stochastic search for "large p" regression', *Journal of the American Statistical Association* **102**(478), 507–516.

Harrell, F., Lee, K. L. & Mark, D. B. (1996), 'Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in medicine* **15**, 361–387.

Hastings, W. K. (1970), 'Monte carlo sampling methods using markov chains and their applications', *Biometrika* **57**(1), 97–109.

Heidelberger, P. & Welch, P. D. (1983), 'Simulation run length control in the presence of an initial transient', *Operations Research* **31**(6), 1109–1144.

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999), 'Bayesian model averaging: A tutorial', *Statistical science* pp. 382–401.

Hoffmann-Jørgensen, J. (1994), *Probability with a view towards statistics*, Vol. 2, CRC Press.

Hurvich, C. M. & Tsai, C.-L. (1989), 'Regression and time series model selection in small samples', *Biometrika* **76**(2), 297–307.

Jasra, A., Stephens, D. A. & Holmes, C. C. (2007), 'On population-based simulation for static inference', *Statistics and Computing* **17**(3), 263–279.

Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation problems', *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186**(1007), 453–461.

Jeffreys, H. (1961), *Theory of Probability (3rd edt)*, Oxford University Press.

Kass, R. E., Carlin, B. P., Gelman, A. & Neal, R. M. (1998), 'Markov chain monte carlo in practice: A roundtable discussion', *The American Statistician* **52**(2), 93–100.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

Kass, R. E. & Wasserman, L. (1995), 'A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion', *Journal of the American Statistical Association* **90**(431), 928–934.

Kuo, L. & Mallick, B. (1998), 'Variable selection for regression models', *Sankhyā: The Indian Journal of Statistics, Series B* pp. 65–81.

Lamport, L. (1994), 'Latex. a document preparation...', *System* .

Lauritzen, S. L. & Spiegelhalter, D. J. (1988), 'Local computations with probabilities on graphical structures and their application to expert systems', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 157–224.

L'Ecuyer, P. (2012), *Random number generation*, Springer.

Lewis, S. M. & Raftery, A. E. (1997), 'Estimating bayes factors via posterior simulation with the laplace–metropolis estimator', *Journal of the American Statistical Association* **92**(438), 648–655.

Ley, E. & Steel, M. F. (2009), 'On the effect of prior assumptions in bayesian model averaging with applications to growth regression', *Journal of Applied Econometrics* **24**(4), 651–674.

Ley, E. & Steel, M. F. (2012), 'Mixtures of g-priors for bayesian model averaging with economic applications', *Journal of Econometrics* .

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008), 'Mixtures of g priors for bayesian variable selection', *Journal of the American Statistical Association* **103**(481).

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

Liu, J. S. & Chen, R. (1998), 'Sequential monte carlo methods for dynamic systems', *Journal of the American statistical association* **93**(443), 1032–1044.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P. & Wagenmakers, E.-J. (2011), 'A tutorial on bayes factor estimation with the product space method', *Journal of Mathematical Psychology* **55**(5), 331–347.

Lopes, H. F. & West, M. (2004), 'Bayesian model assessment in factor analysis', *Statistica Sinica* **14**(1), 41–68.

Lunn, D. J., Best, N. & Whittaker, J. C. (2009), 'Generic reversible jump mcmc using graphical models', *Statistics and Computing* **19**(4), 395–408.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000), 'Winbugs – a bayesian modelling framework: Concepts, structure, and extensibility', *Statistics and Computing* **10**(4), 325–337.
URL: *http://dx.doi.org/10.1023/A:1008929526011*

Lunn, D. J., Whittaker, J. C. & Best, N. (2006), 'A bayesian toolkit for genetic association studies', *Genetic epidemiology* **30**(3), 231–247.

Madigan, D. & Raftery, A. E. (1994), 'Model selection and accounting for model uncertainty in graphical models using occam's window', *Journal of the American Statistical Association* **89**(428), 1535–1546.

Madigan, D., York, J. & Allard, D. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review/Revue Internationale de Statistique* pp. 215–232.

Mallows, C. L. (1973), 'Some comments on c p', *Technometrics* **15**(4), 661–675.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011), 'Mcmcpack: Markov chain monte carlo in r', *Journal of Statistical Software* **42**(9), 1–21.

McCullagh, P. & Nelder, J. A. (1989), 'Generalized linear models (monographs on statistics and applied probability 37)', *Chapman Hall, London* .

McCulloch, R. & Rossi, P. E. (1991), 'A bayesian approach to testing the arbitrage pricing theory', *Journal of Econometrics* **49**(1), 141–168.

Meng, X.-L. & Wong, W. H. (1996), 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration', *Statistica Sinica* **6**, 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**, 1087.

Meyn, S. P., Tweedie, R. L. & Glynn, P. W. (2009), *Markov chains and stochastic stability*, Vol. 2, Cambridge University Press Cambridge.

Miller, A. J. (2002), *Subset selection in regression*, Vol. 95, Chapman & Hall.

Morris, C. N. (1983), 'Parametric empirical bayes inference: theory and applications', *Journal of the American Statistical Association* **78**(381), 47–55.

Murray, I., Adams, R. P. & MacKay, D. J. (2009), 'Elliptical slice sampling', *arXiv preprint arXiv:1001.0175* .

Neal, R. M. (2003), 'Slice sampling', *Annals of statistics* pp. 705–741.

Neumann, V. (1963), 'Various techniques used in connection with random digits'.

Newton, M. A. & Raftery, A. E. (1994), 'Approximate bayesian inference with the weighted likelihood bootstrap', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 3–48.

Ntzoufras, I. (1999), 'Aspects of bayesian model and variable selection using mcmc', *Athens: University of economics and business* .

Ntzoufras, I. (2002), 'Gibbs variable selection using bugs', *Journal of statistical software* **7**(7), 1–19.

Ntzoufras, I. (2011), *Bayesian modeling using WinBUGS*, Vol. 698, Wiley.

Ntzoufras, I., Forster, J. J. & Dellaportas, P. (2000), 'Stochastic search variable selection for log-linear models', *Journal of Statistical Computation and Simulation* **68**(1), 23–37.

Nummelin, E. (1984), *General irreducible Markov chains and non-negative operators*, Vol. 83, Cambridge University Press.

Oetiker, T., Partl, H., Hyna, I. & Schlegl, E. (2010), 'The not so short introduction to latex 2'.

O'Hara, R. B. & Sillanpää, M. J. (2009), 'A review of bayesian variable selection methods: what, how and which', *Bayesian analysis* **4**(1), 85–117.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference*, Morgan Kaufmann Pub.

Pemstein, D., Quinn, K. M. & Martin, A. D. (2007), 'The scythe statistical library: An open source c++ library for statistical computation.', *Journal of Statistical Software* **1**, 29.

Perrakis, K. (2008), Comparison of MCMC Methods for the Estimation of the Marginal Likelihood for Bayesian Model Evaluation, PhD thesis, Athens University of Economic and Business.

Petralias, A. & Dellaportas, P. (2012), 'An mcmc model search algorithm for regression problems', *Journal of Statistical Computation and Simulation* (ahead-of-print), 1–19.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'Coda: Convergence diagnosis and output analysis for mcmc', *R News* **6**(1), 7–11.
URL: *http://CRAN.R-project.org/doc/Rnews/*

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Raftery, A. E. & Lewis, S. (1992*a*), 'How many iterations in the gibbs sampler', *Bayesian statistics* **4**(2), 763–773.

Raftery, A. E. & Lewis, S. M. (1992*b*), '[practical markov chain monte carlo]: Comment: One long run with diagnostics: Implementation strategies for markov chain monte carlo', *Statistical Science* **7**(4), 493–497.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**(437), 179–191.

Ripley, B. D. (1977), 'Modelling spatial patterns', *Journal of the Royal Statistical Society. Series B (methodological)* pp. 172–212.

Robert, C. & Casella, G. (2009), *Introducing Monte Carlo Methods with R*, Springer.

Robert, C. & Casella, G. (2011), 'A short history of markov chain monte carlo: subjective recollections from incomplete data', *Statistical Science* **26**(1), 102–115.

Robert, C. P. & Casella, G. (2004), *Monte Carlo statistical methods*, Vol. 319, Citeseer.

Roberts, G. O. (n.d.), 'Markov chain concepts related to sampling algorithms', *Markov chain Monte Carlo in practice* **57**, 45–58.

Roberts, G. O., Gelman, A. & Gilks, W. R. (1997), 'Weak convergence and optimal scaling of random walk metropolis algorithms', *The Annals of Applied Probability* **7**(1), 110–120.

Royston, P. & Altman, D. G. (1994), 'Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling', *Applied Statistics* pp. 429–467.

Rubin, D. B. (1984), 'Bayesianly justifiable and relevant frequency calculations for the applies statistician', *The Annals of Statistics* pp. 1151–1172.

Rubin, D. B. (1987), 'The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm', *Journal of the American Statistical Association* **82**(398), 543–546.

Rubin, D. B. et al. (1988), 'Using the sir algorithm to simulate posterior distributions', *Bayesian statistics* **3**, 395–402.

Sabanés Bové, D. & Held, L. (2011), 'Hyper-g priors for generalized linear models', *Bayesian Analysis* **6**(3), 387–410.

Sahlin, K. (2011), Estimating convergence of Markov chain Monte Carlo simulations, PhD thesis, MS Thesis, Stockholm University.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Selvin, S. (1975a), 'On the monty hall problem', *American Statistician* **29**(3), 134.

Selvin, S. (1975b), 'A problem in probability', *American Statistician* **29**(1), 67.

Selvin, S. (1975c), 'A problem in probability (letter to the editor)', *The American Statistician* **29**(3), 134.

Selvin, S. (n.d.), 'On the monty halls problem (letter to the editor)'.

Sisson, S. A. (2005), 'Transdimensional markov chains: A decade of progress and future perspectives', *Journal of the American Statistical Association* **100**(471), 1077–1089.

Smith, B. J. (2005), 'Bayesian output analysis program (boa) for mcmc', *R package version* **1**(5).

Smith, B. J. (2007), 'boa: an r package for mcmc output convergence assessment and posterior inference', *Journal of Statistical Software* **21**(11), 1–37.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003), 'Winbugs user manual', *Cambridge: MRC Biostatistics Unit* .

Stephens, M. (2000), 'Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods', *Annals of Statistics* pp. 40–74.

Stigler, S. M. (1983), 'Who discovered bayes's theorem?', *The American Statistician* **37**(4a), 290–296.

Stigler, S. M. (1986*a*), *The history of statistics.*, Harvard University Press.

Stigler, S. M. (1986*b*), *The Story of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press.

Stigler, S. M. (2002), *Statistics on the table: The history of statistical concepts and methods*, Harvard University Press.

Sturtz, S., Ligges, U. & Gelman, A. (2005), 'R2winbugs: A package for running winbugs from r', *Journal of Statistical Software* **12**(3), 1–16.
URL: *http://www.jstatsoft.org*

Syed, A. R. (2011), 'A review of cross validation and adaptive model selection'.

Takeuchi, K. (1976), 'Distribution of information statistics and criteria for adequacy of models', *Math. Sci* **153**, 12–18.

Tanner, M. A. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American statistical Association* **82**(398), 528–540.

Thomas, A., Spiegelhalter, D. J. & Gilks, W. (1992), 'Bugs: A program to perform bayesian inference using gibbs sampling', *Bayesian statistics* **4**(9), 837–842.

Thomas, N. (2010), '"overview". openbugs website'.
URL: *http://en.wikipedia.org/wiki/OpenBUGS*

Thompson, M. B. (2011), Slice Sampling with Multivariate Steps, PhD thesis, University of Toronto.

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

Wang, X. & Wood, D. (1996), 'Xtable: A tabular editor and formatter'.

Wild, P. & Gilks, W. (1993), 'Algorithm as 287: Adaptive rejection sampling from log-concave density functions', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **42**(4), 701–709.

Wilson, D. B. (1999), 'How to couple from the past using a read-once source of randomness', *arXiv preprint math/9910050* .

Zellner, A. (1986), 'On assesing prior distributions and bayesian regression analysis with g-prior distributions', *Studies in Bayesian Econometrics and Statistics* (ch. 5), 233–243.

Zeugner, S. & Feldkircher, M. (2009), 'Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in bayesian model averaging', *IMF Working Papers* pp. 1–39.